TOWARD THE DEVELOPMENT OF A MARKING AND REPORTING SYSTEM ACCEPTABLE TO BOTH CRITICS AND PROPONENTS OF LETTER GRADING

by

Lillian Rosenberg Hurwitz

A DISSERTATION

Submitted to the Office for Graduate Studies,
Graduate Division of Wayne State University,
Detroit, Michigan
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
1979

MAJOR: EVALUATION AND RESEARCH

APPROVED BY:

Donald R. Marcotte March 13, 1979
Adviser Date

Colleen Brun

Abraham F. Citron

Dedicated to my father, Bernard D. Rosenberg, 1892-1970, whose inordinate courage, wisdom, patience and tolerance provided the guiding lights that made this dissertation possible.

ACKNOWLEDGEMENTS

I am very appreciative of the continual support and for the valuable comments of my advisor, Dr. Marcotte, and the rest of the members of my committee, Dr. Brown, Dr. Raider, who replaced Dr. Beres when the latter was granted a sabbatical, and Dr. Citron, who replaced Dr. Klass when he took a leave of absence. The selflessness of my family permitted me the considerable amount of time needed for this project. Their contributions to the proof reading and the sundry assorted tasks required for the completion of this manuscript are gratefully acknowledged. I am also deeply indebted to the Wayne State University Computing Center for providing the Seminars that made it possible for me to computerize this dissertation.

PREFACE

With the birth of my first child in 1949, I found myself confronted, first hand, with a myriad of developmental processes that previously had been merely words in print. Within the next five years I was blessed with two more sons and, as they grew, my interest in and concern with their exposure to positive and stimulating educational forces sent me to the libraries to swallow up semi-popular literature on developments in the field as well as to join the various PTA organizations in the schools that my sons attended. I was hoping in these ways to learn more about the schooling process and, at the same time, to play a role in catalyzing administrators and faculty toward generally perceived needed reforms.

Because I seemed to be more informed than most parents in regard to educational processes and because I was willing to assume leadership roles, I soon became heir to a shockingly large number of complaints from a large body of parents, and heir to the knowledge that many of these children were perceived by both parents and teachers as having learning problems. Not having experienced similar learning problems with my own children and having been made aware of the sometimes severe psychological trauma that these families were undergoing, my sympathy was aroused and I began to focus on school inadequacies, as well as the educational stimulators that I had previously been searching out. The literature and discussions with parents and

teachers led me to believe that the learning problems were related largely to genetic components in the students, poorly trained teachers and inadequate stimulation on the part of the parents. The prevailing sentiment at this time was that many of the problems could be alleviated by some of the major reforms that had been gaining momentum with the advent of Sputnik. These included new teaching methodologies (viz., team teaching, the open classroom, non-graded classes, and programmed learning); curricula developed by experts; in-service and updated training programs for teachers; and even newly developed management techniques such as PERT, PPBS, MBO, and MIS. The literature in education bombarded those of us who were reading it with facts which indicated that all of the above could contribute immeasurably toward improving the learning process for a large majority of students.

Becoming aware as a PTA member of the apparent resistance of school administrators and faculties to any change process, I joined a Better Schools Committee. Such committees worked through Boards of Education and therefore, it seemed, had greater potential for introducing reforms. However, even this committee was not able to make significant headway in making the schools more responsive to the needs of its constituents. It was then that I decided to move into the classroom, and returned to school to obtain the necessary educational prerequisites for teaching chemistry, my college major, on a high school level. I wanted to learn experientially about the schooling process

as well as hopefully to contribute to exciting students in pursuit of knowledge for their own sakes, and welcomed the opportunity to move into a school populated by low achieving, low income students. I was also hoping that I might have a better chance of initiating change as a member of the system.

Reading the required literature in the educational courses that I had taken at Columbia Teachers College, where I received my Masters degree in Science Teaching, made me aware that of those persons who had been successful in getting school systems to adopt the aforementioned reforms, few had been successful in overseeing the successful implementation of such reforms. The large majority of schools used these reforms in such watered down fashion as to make them, in effect, inoperable. It was not until I was part of the system that I began to understand how many of the very structures of the system serve to undermine the introduction of such reform as well as to contribute negatively to the learning process itself (highly rated teacher traits and skills, improved parental relationships, understandings and skills at educational stimulation, improved teaching methodologies and technical aids, and genetic components notwithstanding). In 1971 I put to paper my studied reflections.

One of the structures that seemed to be detrimental to both the learning and teaching processes, and which militated against reform as well, was letter grading. The need to study this phenomenon in depth resulted in my

enrolling in a doctoral study program. With a major in Evaluation and Research, I was hoping to utilize the strategies and tactics that I had learned through my studies in the sciences in such manner that they would enable me to make a unique contribution to educational research as well as to ease some of the needless psychological trauma of students as they are 'belted' through our educational institutions. I perceived that it was possible to develop a theoretical framework for measuring cognitive achievement, based on empirical and supportive research, that could be analogous to that developed for the electron which suggested how its mass might be measured. (Through systematic studies of the research findings, those properties of the electron which enabled first its charge/mass ratio, then its charge and finally its mass to be measured, were revealed.)

The demand for an equitable marking system had made its appearance at the turn of the century and while such demands have waxed and waned in intensity with the educational tides, the demand has appeared continuously over the ensuing years, and has been made by countless educators. My literature search on the subject of grading pointed to two possible basic factors that were playing a role in preventing the development of the kind of theoretical framework needed for a generally acceptable marking and reporting system. These factors seemed to be:

1. the theoretical bases from which these new systems were being launched were faulty and
2. the systems developed were not acceptable to

both critics and proponents of letter grading.

My knowledge of the history of science had taught me that for a theory to be acceptable, it must be readily applicable; my knowledge of the history of the reform movement in education had taught me that for needed educational reform to have any chance of implementation, it must have widespread support among educators. Using such knowledge as a frame of reference I then proceeded to develop a theoretical framework that was derived from those factors which seemed most responsible for the criticism of letter grading as well as those factors which seemed most responsible for its wide-spread acceptance. That the factors come from both sides of the scoreboard had to be a sine--qua--non, the crucial point of departure and the basis for a theoretical framework that had the greatest probability of producing a marking and reporting system that would work. Thus the title of this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

Table

CHAPTER I

INTRODUCTION

Statement of the Problem

The problem on which this study focuses is the development of a marking and reporting system that could be acceptable to both critics and proponents of letter grading systems.

The tasks fundamental to tackling the problem were:

1. to delineate through extensive library research those factors most responsible for the acceptance of letter grading as well as those most responsible for the widespread criticism of letter grades;
2. to analyze reporting systems presently in use in light of these factors;
3. to summarize knowledge obtained from the above two tasks that a theoretical framework might be developed which encompasses the requirements of both proponents and critics of letter grades;
4. to search the literature for devices and mechanisms that can best serve the above

framework; and

5. to set up a feasibility study to test out these devices and mechanisms.

When tasks one through four were completed, the following emerged as having the greatest probability of meeting the criteria developed: cognitive objectives, mastery strategies, a mastery criterion score of 80-85% correct answers and a mastery cognitive profile.

Background and Significance of the Problem

The literature in education is replete with consistent and continual dissatisfaction with letter grading. According to Gronlund, teachers rank it as one of the most important issues of major concern to them.[1] However, while concern for the low validity found in the use of letter grading has been a part of the educational literature since such practises were first introduced at the turn of the century, letter grading continues to remain an integral part of the educational process. Attempts to replace it with pass-fail systems, check-lists of content objectives, written evaluations and the like have largely failed. A 1970 study by NEA revealed that nationwide 72-83% of a sample of public elementary and secondary school teachers were still using letter grades.[2] Most colleges and

---

[1]Norman E. Gronlund, Improving Marking and Reporting in Classroom Instruction (New York: Macmillan Publishing Co., 1974), p. 1.

[2]National Education Association, "Marking and Reporting Pupils Progress," Research Summary 1970 S-1 (Washington, D.C.: NEA Research Division, 1970.)

universities continue to rely largely on letter grades for evaluating student performance in the classroom.

The general consensus in the literature as to the reasons for the widespread acceptance of letter grading, a conclusion supported by reviewers of marks and marking systems, falls into the following six categories:[1]

1. easy to record;
2. easy to average;
3. easy to interpret;
4. good predictors of college achievement;
5. required by colleges as part of the entry process; and
6. needed as motivators.

The opposition to letter grading relates to the following four categories;

1. strong evidence of the lack of consistency, hence indications of poor reliability;
2. lack of clarity as well as agreement on what letter grading purports to measure; hence low validity;
3. can have serious negative side effects; and
4. fails to provide constructive communication.

This view is also supported by the aforementioned

---

[1]Norman E. Gronlund, op. cit., 1974, pp. 1-20: Encyclopedia of Educational Research, 3rd ed., s. v. "Marks and Marking Systems." by A.Z. Smith and J.E. Dobbin: Encyclopedia of Educational Research, 4th ed., s. v. "Marks and Marking Systems," by R. Thorndike, Howard Kirschenbaum, Rodney Napier and Sidney B. Simon. Wad-ja-get? (New York: Hart Publishing Co., 1971).

reviewers.[1]

Summarized, reviews of the literature indicate that present letter grading systems need replacement with an evaluative system that;

1. has clearly delineated objectives, thus increasing the probability of high validity;
2. can produce evidence of internal consistency and consequently a high degree of reliability;
3. is administratively functional;
4. is a constructive communicator;
5. reduces pressures that produce serious negative side-effects; and
6. provides motivation.

Further extensive reviews of the literature revealed consistent and long-standing support, both research-based and empirically derived, for the above framework, suggesting strongly that these regularities provide a soundly based theoretical framework from which to proceed.

Upon reviewing research studies demonstrating the low validity of letter grading, the differential criteria used in the evaluative process seem to be the prime culprit. Most letter grading emerges as the result of an amorphous set of criteria developed by each individual evaluator on the basis of a set of scores obtained via the use of an implicit or explicit group of content objectives and on the basis of evidence gleaned from affective and psychomotor

---

[1]Ibid.

behaviors. Such evaluation lacks the kind of rigidity that an agreed upon framework has, such as a ruler. If one is interested in developing an evaluative process which has the kind of high validity found in instruments like rulers, a relatively rigid framework of educational objectives is necessary. Early researchers, in particular, have promulgated this idea.[1] But it seems that the correlation between grade-point average in high school and college, i.e., 0.54-0.60, plus the administrative functionality of letter grading resulting in the institutional inertia which Thorndike emphasizes[2] have been limiting factors in the improvement of grading practises.

To develop a system, then, which has some chance of being widely accepted these limiting factors must be tackled, and probably could be in the following ways:

1. by more wide-spread recognition that a correlation of 0.54-0.60 only accounts for 29-36% of the variance related to such measures, leaving most of the variance unaccounted for;
2. by developing more awareness of the fact that tables on the accuracy of prediction of the correlation coefficient predict, amongst a thousand cases, an accuracy of only about 30-64%;[3] and

---

[1]Ibid.

[2]Encyclopedia of Educational Research, 4th ed., op. cit., p. 766.

[3]Stephen Isaac and William B. Michael, Handbook in Research and Evaluation, (San Diego: Robert R. Knapp, 1974), p 149.

3. by easy access to clearly delineated educational objectives, which are administratively functional.

The matter of definitive educational objectives had been addressed by such early researchers like Starch when he recommended "definite, objective measures of educational products".[1] However, explicit performance objectives in education have not been widely used until recent times as Ebel points out. He attributes their increasing use to Tyler, the advent of teaching machines, and programmed and individualized instruction.[2] Such objectives have been found to be useful in educational management, classroom procedures, and for feedback and accounting purposes where they have been used to aid in the establishment of clear-cut, operational goals so as to reduce these goals to fewer interpretations. When used as a tool for evaluating achievement, however, they evoke controversy. For used as they have been to make explicit content and task domains and to elucidate non-achievment factors they become:

1. unwieldy and time consuming, if clearly delineated, for they end up being large in number;
2. too rigid a framework for the teaching process--violating some of the basic tenets of good teaching and learning practises which

---

[1]Daniel Starch, Educational Measurement, (New York: Macmillan, 1918), p. 1.

[2]Robert L. Ebel, "Behavioral Objectives: A Close Look," Phi-Delta-Kappan, November 1970, p. 171.

encourage exploration of unanticipated areas of interest; and

3. part and parcel of an evaluation system which is not comparable across classrooms and frequently not even within classrooms.[1]

Many researchers, over the years, such as Rugg, Bass, Odell, and Wrinkle, have suggested that achievement objectives such as those related to cognitive competencies be separated from non-achievement objectives which are related to affective behaviors. These latter factors were perceived as contributing to the low reliability of letter grades.[2] They certainly contributed to the forty-nine different factors that Johnson found could influence grades.[3]

Even though concern by the educational establishment for the development of the whole human being has been on the increase since the forties,[4] educators like Ebel and Gronlund indicate that societal demands persist in making cognitive achievement the major focus of schooling. If cognitive achievement is the crucial focus of the schooling

---

[1]Stephen Isaac and William B. Michael, op. cit., pp. 164-165.

[2]Harold O. Rugg, "Teachers' Marks and Marking Systems," Ed Adm Sup 1 (November 1915); B.M. Bass, "Intrauniversity Variation in Grading Practises," Journ Ed Psych 21 (1930): 48-52; C.W. Odell, "High School Marking Systems," School Review 33 (1925): 346-54; William L. Wrinkle, Improving Marking and Reporting Practises in Elementary and Secondary Schools (New York: Holt, 1947), p. 9.

[3]Franklin W. Johnson, "A Study of High School Grades," School Review 19 (1911): 13-24.

[4]Encyclopedia of Educational Research, 3rd ed., op. cit., p. 787.

process and one wants to increase the reliability of evaluating that process, then separating out cognitive achievement factors from non-cognitive achievement factors or non-achievement factors should contribute considerably in that direction.[1]

Many educators have a kind of ephemeral understanding that they are honing in on different cognitive skills when they frame a question. They learn empirically that sheer knowledge of content material does not necessarily lead to comprehension or application of the content. But they fail to make explicit what is implicit in their understandings, i.e., when a body of content is learned, it can be learned at different levels of cognitive functioning. One would suspect that this fact has also contributed to the unreliability of grading practises. Content and cognitive ability to deal with the content are two different sides of the same coin, somewhat like mass and energy. We have commonly focused on content, but it has been demonstrated that we learn content at varying levels of cognitive functioning. The cognitive skills and abilities acquired in dealing with a given content may be more crucial aspects of the learning process in terms of pinpointing future life styles or areas of most probable societal contributions. If we turn the coin around and measure the cognitive skills and abilities acquired in dealing with the content of a course,

---

[1]Robert L. Ebel, Measuring Educational Achievement (New Jersey: Prentice-Hall, 1965), p.39; Norman E. Gronlund, op.cit., 1974, p.11.

our objectives would become cognitive oriented rather than content or task oriented, and of greater generality but sufficient specificity to measure cognitive achievement. Such objectives are available and can be found in the Taxonomy of Educational Objectives edited by Bloom et al.[1] They comprise six major groupings of intellectual skills and abilities: Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation. Their authors, an august group of over thirty college and university professors in collaboration with countless test constructors, curriculum workers and teachers, claim that these categories represent a hierarchical order of educational outcomes consistent with research findings and represent such a level of generality as to cut across content areas, teaching methodologies, and educational philosophies.[2]

Sound evidence concerning the validity of these cognitive constructs is not available however, according to Kropp, Stoker and Banshaw.[3] Research done by them nevertheless lends support to the hierarchical nature of these objectives as well as gives some support to their generality. If the validity of these cognitive objectives could be supported, then these objectives would have none of the disabilities found in working with content-oriented

---

[1]B.S. Bloom (ed), et al., Taxonomy of Educational Objectives--The Classification of Educational Goals. Handbook I: Cognitive Domain (New York: David McKay Co., 1956).

[2]Ibid., pp. 16-18.

[3]R. P. Kropp, H. W. Stoker, and W. L. Bashaw, "The Validity of the Taxomomy of Educational Objectives," The Journ of Exper Ed 34 (Spring 1966): 69-76.

objectives. These cognitive objectives are:

1. only six in number, making them administratively functional;
2. allow for flexibility in teaching modes and content areas; and
3. can be comparable between and within classrooms, if the authors' claims for them are borne out.

By using the content of subject matter as the medium, i.e., the intervening substance through which cognitive skills are acquired rather than the measuring tool, and by making the cognitive objectives the tool for evaluating whether the learner has acquired knowledge and skills in a particular area and level of content, one should have an evaluative process that could satisfy the first four criteria mentioned in the summary as needed for an acceptable system (1. high validity, 2. high reliability, 3. administrative functionality, and 4. constructive communication).

In such a system, one would need to criterion-reference traditional test items to the six major objectives, so that evaluation could be employed related to achieving a particular level of content in a given cognitive objective. Scores on such tests could be averaged as is traditionally done and a cognitive profile set up on which these scores could be recorded (see Appendix A). Thus one could have normative scores, i.e., scores based on one's performance within one's group. Such scores based on achievement

relative to Bloom's cognitive objectives on a cognitive profile should prove to be more reliable and valid as an evaluative process than present letter grading, and certainly more constructive in terms of communicative value and as administratively functional as letter grading. However, such a profile would not reflect an individual's maximum capacity to achieve, nor would it contribute significantly, according to research findings, to the last two criteria of the summary (5. reduce negative side-effects, and 6. provide motivation).

There is common knowledge as well as research evidence that factors such as fatigue, personal pressures, anxiety, motivation and interest will affect test scores in the short run, as will maturation and learning modes over a longer period. To in fact maximize the reliability, then, of an achievement evaluation system, one must minimize the effects of as many sources of variance as is possible. Proponents of mastery learning claim that such maximization is more possible with mastery strategies than with traditional teaching methods. They claim that close to 90% of most student bodies can learn a subject to a high level of mastery,[1] if given sufficient time and appropriate learning aids.[2]

Their claims stem from a sizable body of research which

---

[1]Benjamin S. Bloom, "Mastery Learning," in Mastery Learning, ed. James H. Block (New York: Holt, Rinehart and Winston, 1973), p. 48.

[2]James H. Block,"Introduction to Mastery Learning: Theory and Practise," in ibid., p.5.

is based on the Carroll postulate.[1] These proponents also claim that such strategies tend to improve student attitudes toward learning, students' self-concepts and mental health by pointing to research which indicates that significant relationsips exist between these factors and achievement.[2] If such is indeed the case, then not only should mastery strategies maximize achievement, but they should also serve to meet the last two criteria of the theoretical framework: i.e. reduce some of the pressures that produce negative side-effects; and be an effective motivator.

With mastery the only criterion for achievement, a simple cognitive profile could be generated on which one has only to record date of mastery in the six different cognitive objectives at any given level of content (see Appendix A). One would need, to determine mastery, not only tests criterion-referenced to the six major cognitive objectives, but a cut-off score to separate the masters from the non-masters. Such a profile would reflect achievement objectives, but not non-achievement objectives, would focus on cognitive behaviors, and would be a measure of success not failure. It would provide recognition of a student's ability to master some subject matter at some content level while differentiating the cognitive skills and abilities learned at that level. It would be an absolute standard by

---

[1] Degree of Learning = f(1. Time allowed 2. Perseverance /3. Aptitude 4. Quality of Instruction 5. Ability to Understand Instruction).

[2] Idem, "Affective Consequences of School Achievement," in ibid., pp. 13-26.

which the student is appraised as an individual, and would keep open the possibility of demonstrated mastery at some future date.

With educators like Johnson and Johnson pointing to the immorality of placing students in a predominantly competitive normative structure, which by its nature has to be a failure experience for the majority of students,[1] and with an increased willingness on the part of an educated and enlightened electorate to bring schools into the courts, it behooves the educational community to concern itself more vigorously with the development of a marking and reporting system that has greater validity and reliability than the presently used letter grades. On the basis of research findings investigated in the pursuit of this dissertation, most of which supports the low validity of letter grades, it would seem that educational institutions are highly vulnerable. For it could probably be demonstrated effectively that letter grading systems are fraudulent to the extent that by misrepresentation they can serve to limit an individual's chance for economic survival and security.

Hypotheses

1. General Hypothesis

A marking and reporting system based on mastery strategies, a mastery criterion score of approximately 80-85% correct answers, cognitive objectives and a cognitive

---

[1]David W. Johnson and Roger T. Johnson, "Instructional Goal Structure: Cooperative, Competitive, or Individualistic," Rev Ed Res 44, (Spring 1974): p. 234.

profile is feasible and could serve as a marking and reporting system which would meet the requirements of both critics and proponents of letter grades.

2.--Research Hypotheses

1. Tests can be devised containing test items criterion-referenced to the three major cognitive objectives as specified in the Taxonomy of Educational Objectives: Cognitive Domain: Knowledge, Comprehension, and Application.
2. Mastery strategies coupled with a mastery criterion score of approximately 80-85% correct answers tend to maximize achievment by reducing:
   a. temporary sources of variance such as fatigue, personal pressures, anxiety, motivation, and interest; and
   b. long range sources of variance related to maturation and learning modes;
3. Cognitive objectives can be shown to have the kind of construct validity that can:
   a. cut across content areas;
   b. represent a hierarchy of learning skills; and
   c. reveal deviant patterns of learning behaviors.
4. A mastery cognitive profile could emerge which would reflect a pre-determined level of

mastery in a particular cognitive objective at a particular level of content difficulty at a given point in time and which could prove to be a marking and reporting system that closely meets the requirements of both critics and proponents of letter grades.

3.--Statistical Hypotheses

1. If test items can be categorized into the first three major cognitive objectives with a large measure of validity, then objective item congruence amongst independent judges should be achieved, i.e. an analysis of variance of the scores reflecting the percent agreement of the judges' choices with the author's would support the null hypothesis of no significant differences at the 0.01 level of confidence. (This level of confidence or significance was chosen to avoid making a Type I error. i.e. rejecting a null hypothesis of no significant differences when it is true. In this analysis it was important not to conclude falsely that a difference does exist when in fact it does not).

2. If mastery strategies coupled with criterion scores of 80-85% correct answers tend to eliminate variances that interfere with achievement, then a comparison of the average percent error found on the one-shot Anderson-

Fisk Chemistry Test compared to the average percent error found on tests designed for the mastery strategy, which test for the same cognitive objective, should reveal significant differences i.e. t-tests for correlated samples should indicate significant differences at the 0.05 level of confidence. (Items on the Anderson-Fisk Chemistry Test are categorized by the publishers into the first three major cognitive objectives and therefore can be evaluated in terms of similar levels of content and cognitive objectives as the tests involved in the mastery strategy).

3. If cognitive objectives have the kind of construct validity that can make them useful for a) cutting across content areas, b) differentiating a hierarchy of learning skills, and c) revealing deviant patterns of learning in relation to cognitive skills and abilities:

   a. then means of errors covering different content areas within the same cognitive objective should provide rank orders that are highly correlated, i.e. a Spearman rho correlation should indicate significant correlations at the 0.05 level of confidence.

   b. then t-tests computed from pair-wise

contrasts of means of the average errors on tests related to different cognitive objectives but similar areas of content should reveal significant differences at the 0.05 level of confidence; and

c. then in a given area of content significant differences should occur between rank orders derived from raw scores of tests designed to test for different cognitive objectives. i.e. an analysis of variance of the rank orders will reveal significant differences at the 0.05 level of confidence on sets of tests criterion-referenced to the first three cognitive objectives in similar areas of content.

Definitions-

1. Formative evaluation:[1]

Perceived as an integral part of the teaching-learning process, it is used to provide immediate and continuous feed-back information regarding a student's progress in an instructional unit.

2. Summative evaluation:[1]

An assessment of a student's achievement at the end of an instructional unit, generally based on a

---

[1]James H. Block, op. cit.

one-shot test.

3. Mastery:[1]

Defined in terms of a specific set of major objectives which the student is expected to achieve at the completion of a unit of instruction.

4. Mastery strategy:[1]

A teaching and learning strategy employed to move the learner toward mastery of an instructional unit. Such strategies are based on the theory that achievement is a function of time and appropriate learning aids. Therefore they usually include repeated test taking in the same areas of content (formative tests) and different learning modes before summative evaluation is utilized.

5. Mastery testing:[2]

Such testing involves the use of a cut-off point. The most effective according to available research in terms of maximizing achievement and minimizing negative attitudes on the part of learners seems to be in the 80-85% correct answer range.

6. Criterion-referenced testing:[2]

A criterion-referenced test is one composed of items keyed to a set of behavioral objectives. (Ivans, 1970)

---

[1]Ibid.

[2]U.S. Department of Health, Education, and Welfare, The Evaluation of Mastery Test Items, By Robert Brennan, January 1974, pp. 1-7.

7. Norm-referenced testing:[1],[2]

   A norm-referenced test is one designed to operationally discriminate among subjects with regard to some underlying construct so as to make distinctions among students.

8. Criterion score:[1],[2]

   As used in mastery testing, the cut-off score which separates the masters from the non-masters.

9. Normative score:[2],[3]

   A number assigned to an examinee to provide a description of his performance in relation to some group as determined by a particular test.

10. Criterion-referenced measure:[1]

    An absolute standard of quality, i.e., a student's achievement measured independent of other students' scores.

11. Norm-referenced measure:[1]

    A relative standard, i.e., evaluation in terms of relative position in a group.

12. Content objectives:

    Objectives specifically related to content of a subject involved in learning cognitive skills.

---

[1]Ibid.

[2]In the literature the terms criterion-referenced testing and norm-referenced testing are often used in juxtaposition to each other, thus blurring the distinctions between tests criterion-referenced to some type of objectives and those that are not, and criterion and normative scoring which may or may not be used in either of the above kinds of tests.

[3]R. L. Ebel, op. cit., p. 463.

13. Cognitive objectives:

Objectives specifically related to cognitive skills involved in dealing with the content of a subject.

14. Affective behaviors:[1]

Those behaviors related to interests, attitudes, appreciations, values and emotional sets of an individual.

15. Psychomotor behaviors:[1]

Those behaviors related to individual muscular and motor skills.

16. Systems:[2]

An organized assemblage of interrelated components designed to function as a whole to achieve a predetermined objective.

---

[1]D. R. Kratwohl, B. S. Bloom, B. B. Masia, Taxonomy of Educational--Objectives--The Classification of Educational--Goals--Handbook II: Affective Domain (New York: David McKay Co., 1968). p.7.

[2]R. W. Hostrop, Managing--Education for Results (Illinois: ETC., 1973), p.245.

CHAPTER II

A REVIEW OF THE LITERATURE

Part I

Towards The Theoretical Framework

Thorndike (138) defines a mark as:

1. a single summary statement,
2. covering achievement in some substantial segment of the educational enterprise,
3. given by an instructor,
4. for the purposes of record and report.

He distinguishes a mark from a score, which he says merely expresses performance in relation to a single set of defined and limited tasks, whereas a mark is derived from a set of scores.

Researchers agree that a mark represents the teacher's perception of pupil achievement based on a combination of evidence selected by the teacher; that it is used to develop a permanent record of academic performance which can become available to potential employers and educational institutions; and that marks are used for selective processes. Considering the subjective, selective, and permanent nature of marks, it is little wonder that concern

as to their reliability and validity, as well as negative concomitants, became evident almost as early as their inception.

Letter grading, however, has persisted for almost eighty years, in spite of extensive evidence that it is not only inadequate as a measure of academic performance, but can have such negative effects on a student as to mitigate against the learning process itself. There have been several fairly good reviews of the literature on the subject of marks and marking, such as those by Crooks (35), Ayer (5), Smith and Dobbins (126), Thorndike (138) and Kirschenbaum, Napier and Simon (80). They do not, however, give the reader a sufficiently broad perspective of the research done, nor are they organized into the specific areas related to those factors with which this dissertation concerns itself. Believing that a definitive review could contribute importantly to a problem which has persisted for too long a time, believing that it should be established that much of what is known about grading practices is no longer in the hypothesis stage, but has withstood the tests of time and experimentation, and believing that the problem must be looked at in line with a theoretical framework derived from the literature, an in-depth review of the literature has been undertaken.

The review has been organized in keeping with the concerns of this project into the following categories: history, validity, reliability, administrative functionality, communicative constructiveness, negative

side-effects, motivational value, alternative systems, mastery learning and objectives. Part I starts with the history of letter grading and then reviews the literature in those categories developed for the theoretical framework; Part II examines alternative systems in the light of the theoretical framework; and Part III includes those categories related to the educational devices chosen as having the highest probability of meeting the criteria established by the theoretical framework.

Only those studies whose findings were supported by research designs which seemed adequate in terms of samples and methodology were included in this review. As sizeable a number of studies were reviewed as possible to give support to the validity and reliability of the framework based on the premise that a large number of consistent findings suggest a higher degree of reliability than one highly significant finding (69).

1. --History

It appears that the earliest record of a report card was in 1840 in Horace Mann's Common School Journal. According to Mousley (103), it was merely a device for eliciting support from parents for improving the achievement behavior of their children and apparently contained a word or two indicating the teacher's reflection of academic behavior, e.g. approbation, censure. There was also evidence of a tendency in the nineteenth century to make report cards decorative and to use such as rewards for

superior achievement. With the extended use of the McGuffey Readers and Spelling Books, it became the practice to send home report cards, frequently at the end of the school year, which listed the grade level and the page completed in these books. By 1847 the first age-graded school had been established in Quincy, Massachusetts and by 1860 most of the city areas in the United States had so organized their schools. The impetus for graded classes had come from increasing demands for more and different training in the schools in keeping with the needs of a growing nation, and the subsequent increased need for efficiency in the rapidly burgeoning schools. Horace Mann's Seventh Report in 1844 added a shove to that impetus as Mann was impressed with Prussian schools, and had included in his report high praise for the structural efficiency of the age-graded classes that he had found in that country. Thus with the size of the schools increasing, written messages gave way to percent scores, with some letter grading. These were arrived at through the use of teacher or Board of Education designed tests.

However, the use of one uniform curriculum in the age-graded schools produced three strata of students: a group for whom the work was too easy, a group that the curriculum served well in that it kept pace with their cognitive maturation, and a group which, failing to make progress, fell farther and farther behind each year. To tackle the learning problems involved in dealing with these different groups, differentiated curriculae, differentiated schools,

tracking and "coaching" were instituted. All these new structures served to emphasize the problems inherent in using the same letter and percentage grading systems with each of these groups. Finkelstein in 1913 (48), Kelly in 1914 (76) and Rugg in 1915 (118) articulated some of the fundamental problems that concerned educators about the marking practices. They posed such questions as

1. should marks indicate performance, ability or accomplishment?
2. should marks reflect an average standard of achievement for "normal" children of a given age group?
3. should marks reflect a distribution of ability around a standard?
4. do we have as many standards of marking as there are teachers?

Research undertaken at that time to test out these new percentage and letter grading procedures that were emerging as marking systems indicated great variability in the distributions of marks as well as the actual assignment of a mark to any one paper. Mean variations in marks given to the same student, in the same subject, and on the same exam as pointed out by Rugg in 1915 (122) were running as high as 15%. This report was based on the work of eleven different investigators, who had sampled from 500-26,000 students. And Eells (45) pointed out in 1930 that the differences in assigned grades of A's and failures varied from 2-10%, and C's from 38-50% even though teachers were claiming that they

were assigning grades based on the normal curve distribution.

By the turn of the century new statistical and experimental procedures had been developed and these were used to examine the problems of grading practices and to test the degree of precision that could be expected by their use. Grades within and between schools were correlated and percentage grading was attacked more analytically, resulting in demands for increased standardization of evaluation measures. Towards this end, some researchers focused on the use of the normal curve, others on weighting factors. But studies which examined the effectiveness of such processes found that such attempts failed to substantially increase the reliability and validity of assigned marks. By 1915 Rugg (120) reported that there were thirty-two published reports bearing on the question of the reliability of grading and by 1918 twenty-three more had been published which also showed striking variability of teachers' grades.

Those studies that focused on the unreasonableness of expecting percentage differences of 0.5-1% to reflect real differences in students' achievements led to the abandonment of percentage grades in favor of a four to seven symbol system, with the A,B,C,D,E, system becoming the most popular. By 1932, Billet (10) reported that of the 258 schools he sampled, even though there were one hundred different marking systems among them, 80% were using letter grades or their equivalents.

Educators, understanding the need to introduce more

standardization into the grading process to reduce its subjective nature, perceived that increased standardization of testing procedures might substantially improve the reliability of grading practices. In 1904 Thorndike's (137) treatment of test construction in his book on educational measurement stimulated activity in this field. The appearance of the Terman Revision of the Binet Test in 1916, introduction of the school survey, periodicals in educational measurement, and the organization, in 1912, of the first Educational Research Bureau plus wide-spread testing of recruits for the army that supported World War I were among the significant consequences of such emphasis. The rapid increase in the number of objective and standardized tests provided many more different kinds and larger numbers of criterion measures with which research projects on the reliability of grades could be launched.

While it had been perceived that objective testing in the classroom would improve the reliability of grades and that standardized testing could also serve to aid in making marking more reliable, such developments hardly affected the reliability and validity of grading. Segel (120) summarized such findings in 1934 after looking at the work of twenty-three different investigators, who had attempted to predict college success on the basis of high school grades. He found that the average correlation was of the order of 0.55, with a range of 0.29-0.77. These findings simply added more fuel to the concerns over the unreliability of grading practices, since objective and particularly standardized

tests were more precise measures of cognitive achievement. (Odell (108), who had looked at standardized tests in 1930, found that almost two thirds of them had coefficients of reliability of 0.8-1.00, with one third ranging from 0.9-1.00.) Such findings encouraged more school systems to attempt alternative grading procedures and even one school system, that of Newton, Massachusetts, to abandon grading altogether, in the year 1933.

The developing psychology of the thirties had its impact on society and as a result educators began to focus on the personal and social aspects of schooling, questioning the desirability of stressing the simple acquisition of knowledge. This moved researchers into focusing on the effects on grades of such variables as personality and sex of both teacher and pupil. Greater emphasis emerged on separating out achievement factors from non-achievement factors and more report cards reflected these concerns, having separate sections for the subject matter and separate sections for evaluation of affective behaviors. By 1935 Wrinkle (149) reported that adjustment to the concept of doing away with letter grading was taking place most rapidly in the elementary schools because it was here that the greatest distance from academic domination by higher institutions existed and it was in the elementary schools that one found the less academically inclined teacher. However, in 1947, when he published his book on "Improving Marking And Reporting Practices", he concluded, after ten years experience as part of a team attempting to look at

alternative grading procedures, that there was no one right answer for all schools, and that each school had to develop its own alternative based on its own objectives. His book served also to initiate renewed attacks on the low reliability and validity of letter grading.

By this time the advent of programmed learning and computerized instruction had made possible more individualized instruction and brought on increased interest in marks based on absolute measures. The concept of developing pre-established objectives was getting new emphasis and the development of check lists of content objectives was experiencing renewed favor among reformists. Then, in 1957, the Russians launched a satellite into space by the name of Sputnik. Being the first launch of its kind it offended the American ego and jarred an already vunerable community into looking at schooling practices with more systematized eyes. This brought forth both a spurt in innovations as well as retrenchment of letter grading as a "spur and a whip" to recalcitrant learners - to quote a report published by the National School Public Relations Association (106).

The 1960's was the decade of student power and with it came more demands for the introduction of innovative grading practices into many more schools. There was more widespread understanding that the so-called "objective test" was really a subjectively constructed test scored more or less objectively and that even standardized tests were merely cross-sectional or longtitudinal samples of an arbitrarily

selected universe of possible items.

Some universities experimented with the pass-fail system in selected subject areas. Once again there arose the question of the need for any grade at all. Supporters of non-graded K-12 schools such as Goodal and Anderson (57) were foremost in this regard. Yet in spite of all the increased support for innovative marking systems, a study done by NEA (105) in 1970 on a sample of public schools indicated that close to 80% of these schools were still using letter grades or their equivalents. However, a survey conducted by the American Association of Collegiate Registrars and Admission Officers in 1970-71 of 1,696 of its members, found that among the 96% of the schools using traditional letter grading, 46% were using some alternative non-traditional grading practices as well.

The testing and grading of students still encounters much talk, some experimentation, and continued research, but to date no alternative grading system has found generalized support. Thorndike (138) believes the lack of sensitivity on the part of reformers to institutional complexities is at the seat of the problem. While this aspect of the problem must be given due consideration, the fact is that the problem has not been approached with the kind of scientific logic that has worked successfully for the "hard" sciences. By using the measuring instruments of these disciplines as models, this author believes a highly reliable and highly valid grading system can be developed that encompasses the simplicity and breadth of letter grading as well as

alleviates some of letter grading's negative concommitants. For unless and until such a system of marking and reporting can be developed, educational institutions remain in the position in which Cattell (26) found them in the 1900's: that of the "grocer who lets each of his clerks give to customers without weighing and without knowledge of market prices what he believes to be a dollar's worth of sugar or tea".

2.--Validity

Validity refers to that characteristic of a measurement which enables it to measure what it purports to measure, i.e. the mental and physical image of the measure must be the same for all its users. Almost all researchers who have studied marking practices concede that one of the causes of the considerable variation in teachers' marks is the varying concepts, or mental images, if you will, that markers have concerning what a letter grade represents. Even while the physical representation of a mark, that is the symbols used, are the same for all users, the mental image is not. Among those researchers who have addressed the problem of the validity of letter grading most concisely are Rugg (118), Johnson (73), Odell (108), Adams (1), Travers and Gronlund (143), Vredroe (146), Kirby (79), Haagen (61), Chansky (28), Thorndike (138) and Milton and Edgerley (96).

In 1915 Rugg (118) stated that each teacher has consciously or unconsciously set him or herself up as a designer of educational yardsticks, involving his own scale

of marks in the subjects taught by him and applied it more or less rigorously to his pupils. Thus, he concluded, there exist as many standards of marks as there have been teachers. In 1925 Johnson (73), who sent out a questionnaire to forty-three principals and teachers to determine the bases of their marks, came up with forty-nine different ones.

Odell (111) in 1928 reported at least fifteen features that teachers consider when assigning marks. Among them were such factors as: a pupil's capacities and abilities to perform and memorize, and his/her attitudes, initiative, application, speed, attractiveness, dress, use of the english language, and neatness. Other factors mentioned included the relative importance of competent answers, a determination of what constitutes a correct answer, the decision as to whether to allow partial credit for an answer, and the attitude of the teacher toward marks. Odell also pointed to such facts as that some teachers give relatively high marks believing that they are more encouraging to the learning process, while some teachers give relatively low marks believing that they stimulate greater effort on the part of students, as well as to the facts that some teachers believe in eliminating those students having difficulty keeping up with the class, while others hold the opposite view, believing that keeping slow learners within a class has the effect of stimulating them towards greater effort. Interacting with all these variables, Odell points out, is the added variable that even

the same teacher can vary in his/her opinion from time to time depending on his/her mental or physical condition.

In a study on when teachers fail pupils, Adams (1) in 1932 reported that, in Elkhart County in January of 1930, an analysis of the replies revealed twenty-nine different major reasons for failing students. He had elicited written statements from forty-one classroom teachers accounting for the percentage of failures in their classes the first semester by sending out the request for such information from the superintendent's office. One-third of the replies were categorized as reasons over which the teacher had no control; the remaining two-thirds were related to such items as the student not being able to work up to the standards set by the teacher. However, the teachers never offered any evidence as to what their standards were, nor ever referred to such scores as those derived from standardized tests, even though such were available. The sole criteria for the standards were simply the judgmental statements of the teacher, with some of the teachers admitting to using fear of failure as an inducement to better work.

Travers and Gronlund (143) in 1950 attempting, on a limited basis, to study the variables that faculty members use in assigning grades, used ten instructors from each of five disciplines, who had taught both graduate and undergraduate students in the fall of 1948, making for a sample of fifty male instructors. While they considered their study limited because it was relatively unsophisticated, their sample of instructors seemed

representative and of fairly good size. The authors concluded that there was no one single concept of what a mark should represent, and that those who teach the social sciences tend to place more emphasis on factors such as amount of progress being made by a pupil than those who teach the physical sciences.

Interviewed administrators, teachers, parents and pupils over a period of four years in 328 schools, confirmed for Vredevoe (151) in 1953 that many teachers differ in interpretation of achievement as well as the values assigned to letter grades, with few schools even reviewing the standard of work that a letter grade symbolized.

Three chance factors that affect grades in ways that are "far from trivial" were dug out by Kirby in 1962 (79). He found them to be:

1. grading practices of the instructor,
2. cutting point error, and
3. guessing.

Summing up, in a more general way than the other researchers had, the factors that determine a grade Haaggen (61), at a conference held on grading systems in 1963, concluded that the instructor, the institution, the student, and society all operate in determining what a grade shall represent; That a grade was really a multivariate interacting complex representing the judgment of the grader based on all of the above influences.

The findings of Chansky (28) supported all the aforementioned conclusions, when in 1964 he reported that,

in answer to a questionnaire submitted to forty teachers as to whether literal marks represented a qualitative or quantitative difference in achievement, he obtained responses representing a ratio of fifty to fifty. And that when he analyzed the way marks were used, he found that the grades from which a GPA is determined are derived from capricious judgments and volatile criteria. He further pointed to the fact not previously emphasized that in compiling an average one loses sight of the extremes, for a student may perform very well in one aspect of a course and do mediocre work in another aspect, but because of the averaging out of scores for grading purposes these facts are not reflected in the letter grade he/she might receive under such circumstances. Chansky also cites the work of Marshall who determined, as others before him, that such teacher tendencies to use marks to enforce discipline, cajole or patronize students also play crucial roles in determining marks, and to the work of Battle who found that a portion of a student's mark can be explained in terms of congruence of the student's values with those of the teacher.

Thorndike (138) who had been studying measurement since marks were first introduced at the turn of the century could contribute little more in 1970 but reiterate what Odell had pinpointed in 1928 re the meanings of marks. Thorndike noted that in practice a mark does not merely represent pure competence, but factors such as industry, effort, class participation, neatness, mechanical correctness, docility and cleanliness.

Perhaps the best summarization of what all the others had said about what a letter grade represents is that of Milton and Edgerley. In 1976 they determined that the accumulated evidence concerning letter grades indicates that:

1. letter grades are unidimensional symbols reporting multidimensional phenomena, and
2. the letter grade symbol, by itself, reveals nothing about the quality of the tests through which it is derived.

Studies supporting such statements but which deal with specific and unique aspects of the problem are those of Crawford (33), Barton (7), University Of California (145), Hughes (67), Gould (57), Bolmeier (16), Wrinkle (150), Glaser (55), Aiken (3), Rose (3) and Kelley (75).

Crawford in 1930 (33) reported that, in studying 50,000 marks of 4,985 freshmen who graduated in February and June during the years 1926-1932 at Yale University, he found that the value of grades varied considerably for some departments from year to year, with experience making for a wider distribution of grades.

Barton (7) in 1925 elicited some interesting responses to a questionnaire administered to 1,513 students in four different states, specially chosen by the authorities in each of their respective schools as representative in terms of ability, achievement and socio-economic strata. Their ages ranged from thirteen to twenty-one. In his responses he found that 43.4% of the students did not think that marks

gave them or their parents a true estimate of their accomplishments, 40.9% felt that their teachers did not carefully consider their marks before assigning them, 75.9% felt that some teachers were "harder" markers than others, and 46% thought that their marks were unfair when compared with those of other pupils. At least 50% of the pupils did not attach much value to the mark.

In a report put out by the University of California on education at Berkeley (150), the authors reported on the results of a questionnaire submitted to a random sample of 2,576 returning students. In answer to the question as to how well the students thought their grades reflected their actual knowledge and understanding of the subjects studied, only 3.4% answered "very well", while 49.2% answered "fairly well", and 41.8% answered "only slightly well" and five percent said "not at all". The percent of honor students represented by each group was as follows: very well-3.6%, fairly well-55.8%, and only slightly well-55.8%. The authors point out that these percentages of honor students tend to negate the possibility that such a factor as "sour grapes" was influencing the responses.

In 1930 Hughes (67) attempted to ascertain the ingredients that went into determining school marks. He developed profiles of an average student, an honor society student and a non-honor student of 120 I.Q. or better by using grades, Terman I.Q. scores, Stanford Achievment Test scores, and pooled ratings from a scale that he had developed for measuring affective behaviors as criterion

measures. He found that the average honor student compared to the average non-honor high I.Q. student rated approximately ten points lower in I.Q. scores, but thirty points higher in affective behaviors such as persistency, initiative-aggressiveness, respect for authority, cooperation, leadership and trustworthiness. The non-honor student of high I.Q. was rated significantly lower than the average student in respect for authority.

In studying the widespread lack of uniformity of standards by which letter grades were determined, Gould (57) in 1932 reported that, in answer to a questionnaire that he submitted to 125 schools in forty-eight states, he received replies which indicated that 63% of the teachers were using what they believed to be an absolute marking scale, but with varying cut-off points, while most of the others reported that they were using a relative marking scale. Those using relative marking scales were, however, using curves of varying shapes due to the varying distributions of the number of students assigned the varying grades. The author points out that the replies, which represented forty-seven states and a good distribution of school districts, indicated that there was little uniformity of standards by means of which a pupil's progress was being measured.

Bolmeier (16) did a particularly interesting study in 1943 that demonstrated that personal characteristics of students do indeed feed into grading patterns. He assembled a group of twenty-four school officials, representing twelve different schools, of whom a majority had been former

teachers and asked them to assign grades to a series of six case studies that he had drawn up. Among the more interesting cases were the following:

1. a student slightly below average in mathematics and science achievement, but who put in a good deal of effort to deliver high quality work and turn in assigments on time. He received 1 A, 13 B's 3 C's, and 7 D's.
2. a student with a high I.Q., but poor attitudes. He scored high on tests but was a discipline problem. He recived 3 A's, 13 B's, 4 C's and 4 D's.
3. a rich, female student who was an all A student but who had missed a week of school and in an attempt to maintain her high GPA, because her parents had promised her some gifts, cheated and was caught. She received 2 A's, 4 B's 4 C's, 5 D's and 9 F's.

In 1947 Wrinkle (150) examined the records of four classes that he taught in a general course in secondary education for four successive quarters. He used the same test for each class at the beginning of each quarter and the same test at the end of each quarter for each class. In the first instance he was interested in measuring background knowledge and in the second instance he was interested in measuring comprehensive achievement. He secured from the students' records their percentile scores on two standardized tests. In analyzing the various scores, he

found wide variation in achievement levels between the classes with reference to the four measures, as well as great variation among students within any given class. However, if the marks had been assigned on the basis of the average achievement of all the students in any given class (as is usually the case), a student who received a score of 130 on an end quarter examination would have been given a B had that student been in the fall and winter quarter, and an A had that student been in the spring quarter, and a C had the same student been in the summer quarter. Wrinkle concludes that for any mark to be interpreted correctly, the achievement level of the class must be known.

Supporting the findings of Wrinkle were those of Glaser (55) and Aiken (3). Glaser in 1963 reported that teachers adjust grades to make their distributions more reasonable; that there is a remarkable similarity between grade distributions in high school and those in college, which generally have a different quality student; that grade distributions do not go up even when the quality of students admitted does, for grades tend to be a measure of comparative achievement.

In 1963 Aiken (3) wrote that he believes that in spite of what teachers say, they usually grade with reference to existing ability level of the class, either intuitively or statistically. He cites the case of the University of North Carolina Women's College. The powers-that-be wanted to improve the quality of the student they were admitting and decided to select their students on the basis of scores

obtained by assigning weights to predictor variables such as verbal SAT, mathematical SAT, and a converted two digit score of rank in high school class upon graduation. The weighting factors were determined by the use of regression equations. There were statistical indications that the correlation of these factors with freshman grade-point average could be increased to 0.70 by these techniques. What was found in actual practice was that the faculty simply shifted its standards toward the ability level of the class and that while the quality of the students admitted had increased, as indicated by an increase in mean entrance scores of from 59-61, there was no accompanying increase in criterion mean, i.e. means derived from college grades. Thus one could not interpret the fact that the quality of the student had increased, if one examined only the grades that these students were being assigned.

Rose (3) in 1952 reported, in studying twenty-two college departments for six semester grading periods, evidence that departments having the most students tend to give the lowest average grades and that grades tend to be lower in required courses. They also found that there was a negligible relationship between the student's departmental rank and the student's estimate of course difficulty, or estimate of what grades he/she or the average student would make in the course.

Kelley (75) in 1958 reported that he looked at the discrepancies between the grades given by the instructor for 565 males and 469 females who had completed twelve courses

at Michigan State in the general education program and their scores obtained on the final examination in the same course. Looking at the extreme cases, which amounted to a total of 128 cases, he came up with three groups: those who received higher marks than they should have, those who received lower marks than they should have and those who received marks commensurate with their abilities. In an attempt to understand the variables that might account for such discrepancies, the author had administered an Inventory of Belief. From this inventory, the author concluded that the group receiving higher grades than they should have were characterized as conforming, rigid, and insecure, and the group receiving lower grades than they should have were characterized as lacking in motivation and indifferent. The author suggests that grades might not be an accurate description of mastery of subject matter.

While differences in teachers' standards are the chief causes for variation in marks, Ayer (5) in 1933 pointed out that such qualities as penmanship and sex can play a role in influencing marks. Even prior recall has its effect on the grading process. Studies done by Shepherd (122) and Lauterbach (87) support these statements.

In 1929 Shepherd demonstrated how penmanship can affect marking practices. He submitted the exact same composition written by an eighth grade pupil to 225 teachers. The paper was duplicated in both good and poor quality penmanship as determined by the Ayres Handwriting Scale. The poor quality penmanship paper was graded three weeks after the first

paper. After six months the experiment was repeated but the poor quality penmanship paper was graded first this time and the good quality penmanship paper after a three week interval. Based on the Harvard-Newton Scale, the composition was determined to have a grade of 71.9%, but the means for the first group of papers was 77.896 for the good penmanship paper with a standard deviation of 10.75 and 66.066 for the poor penmanship paper with a standard deviation of 10.60. When the grading was repeated, an interesting reaction was discovered. Both of the group means were higher, even though the direction of the grading differences was the same. The group means this time were 77.927 for the poor penmanship paper, with a standard deviation of 8.25, and for the good penmanship paper 87.949 with a standard deviation of 7.10. While the paper of higher quality penmanship was rated in each case on the average ten points higher than the paper of low quality penmanship, the grades on both papers were affected by prior recall.

Lauterbach (87) studied the distribution of marks of fifty-seven teachers on sixty eighth grade compositions written both on a typewriter and in longhand. The papers had been selected at random, and the typewritten copy was an exact duplicate of the longhand paper. The papers were divided up into sets of thirty, also chosen at random, and all the teachers were either grade or english teachers. The ranges on both papers were from 8-79 percent, the median and means for both papers being exactly the same. However, the

typewritten papers received larger numbers of marks in the 95-100 range and larger numbers of lower marks. The author suggested that the variations might be due to the fact that errors are more prominent on a typewriter than when written in longhand. The rank order correlation for the longhand papers was 0.70 with an uncertainty of 0.10 and for the typewritten papers was 0.64 with an uncertainty of 0.012.

Following are several studies that have indicated that girls in general get higher grades than boys and that women teachers show a preference for girls . The Cocking and Holy (31) study in 1927 at the University of Iowa demonstrated that girls tend to get higher marks than boys. They found no significant differences in the scores obtained on the Thorndike Intelligence Test in their sample of every third freshman at the State University of Iowa making up 107 girls and 159 boys. However, when the authors examined the students' marks, marked differences appeared in the means of marks, both on the University and High School levels, favoring girls. On the University level the marks varied 4.4% from the mean, while on the high school level, the authors found a variation of 10.2% from the mean.

Lentz (89) reported on a study in 1929 in which the marks of 188 girls and 202 boys in grades two through six in a midwestern suburban system were compared with scores on Stanford Achievement Tests. The marks were similarly distributed in the second grade, even though the boys scored higher on the achievement tests. While in achievement tests boys generally did better than girls in three out of five

grades, in four out of five grades the girls got superior marks.

Maney (92) reported in 1933 that in studying the marks given at Transylvania College for the ten-year sessions of twenty semesters running from the years 1921-1931, in all cases the average grade was higher for women than for men. These were marks that had been given by six male instructors who had been in continual service for the greater part or all of that time. While the differences in mean values for three of the professors were negligible, in the other three cases the differences were five to seven times the probable error. The author also points out that the range of mean values was greater for the women than for the men, being 0.13 for the men, but 0.30 for the women.

Supporting such findings that boys generally achieve higher scores on standardized tests than girls is a study done by Eells. In 1937, Eells (45) reports having had administered an examination standardized for the ninth through the twelfth grade level in 198 representative secondary schools. He tested 20,000 juniors, approximately equally distributed in terms of sex. He found that the boys had higher mean scores than the girls, the differences increasing with grade levels.

Carter (26) found that, not only can a student's sex influence a grade, but the sex of the teacher can be a determining factor as well. In 1952 he reported on a study in which he examined the students' scores in a Quick Scoring Otis Test as well as an Algebra Test. These were students

in a public school in a western Pennsylvania city. He compared the scores for the above tests with the students' final grades, and found that the girls had significantly higher grades in Algebra even though there were no significant differences in the scores obtained on the tests between the boys and the girls. The mean correlations of teachers' marks with the Algebra Test for boys was 0.59 while for girls it was 0.45. When intelligence as determined by the Otis Test was partialled out, the mean correlations were 0.47 for boys and 0.36 for girls. On further examination he found that women teachers tended to give higher marks than men teachers. While Carter had found that in general boys were given lower marks than girls, marks assigned by men teachers were even lower than marks assigned by women teachers. His findings supported the findings of those reported by Garner in 1935 (26), by Swenson in 1937 (133), by Douglass in 1938 (42), by Shinnerer in 1944 (123), by Newton in 1942 (26), by Edmiston in 1943 (44), and by Lobaugh in 1942 (26). However, one study done by Yates in 1934 (26), did report no sex differences in grading.

Degree of fatigue and boredom on the part of the teacher can also influence letter grading. Dexter (41) found this to be the case in a study on which he reported in 1935. He recruited thirty advanced students planning to teach penmanship and had them practice scoring papers. Then he asked them to arrange a convenient three hour period to score a set of 400 papers, presuming that they would set

aside a time when they would not feel under pressure. He found that in 75% of the cases there was a constant increase in the average deviation from an earlier period of scoring to a later period between the same items as well as different items, with a tendency toward either increasing severity or leniency. It was the better students who seemed to move in the direction of greater severity.

That last statement finds support in a study done by Rocchio and Kearney (116), who reported in 1954 that they found significant relationships between MTAI, (Minnesota Teacher Attitude Inventory) scores and failure rates in 1954. They interpreted these findings to mean that teacher attitudes and failure rates are related. They suggested that those teachers who think in terms of subject matter are more likely to fail students then those who think of pupils as pupils.

The evidence that letter grading is not used to measure the same things all the time and that it does not mean the same to all its users is overwhelming. While one may take issue with specific aspects of the research that has been reported on, nevertheless the findings have consistently and abundantly pointed in the same direction. The probability of obtaining such an abundant consistency by chance are so slight that one can only concur with the Rugg statement made in 1915 that there are as "many standards for marks as there are teachers". The evidence is patently clear that letter grading falls far short as a valid measure of student performance. To use it as if it were an acceptable measure

for evaluating the many and diverse aspects of studenthood is an anachronism unworthy of our age.

3. Reliability

An instrument can be invalid, nevertheless reliable, for reliability refers to that characteristic of a measurement which makes possible consistent results of the same measurement, no matter who uses it, within acceptably established limits of precision. It is possible to have an instrument measure the same value each time, but not be valid, when consistent intervening variables effectively reduce the instrument's validity. For example, a meter stick made of metal that contracts in cold temperatures might measure the same value for a meter in that cold climate each time it was used, but while it would be measuring a meter precisely i.e. reliably, it would not be measuring it accurately i.e. validly. Thus while we have seen overwhelming evidence that letter grading is not a highly valid measure it could still be a reliable measure, if the variables affecting its validity were consistent. A look at the research evidence, however, does not support such a possibility. The large number of research investigations, undertaken almost since the inception of these grading measures, provide considerable evidence that letter grades are not highly reliable as measures of academic performance.

One of the first attempts to determine the reliability of marks was that by F.Y. Edgeworth in 1889 (43). He was

then a professor at the University of Oxford and was so bold as to advertise for competent people to rate a paper on the quality of its Latin prose. His advertisement brought forth twenty-eight "highly competent examiners". Their ratings ranged from 45-100 points. Similar findings were reported on the marking of examination papers by such researchers as Jacoby (72), Ruggles (123), Starch and Elliot (134) (135) (136), Gray (60), Kelley (77) Bolton (18), Hulten (68), Eells (45), Tieg (139), Lawson (88), Penfold (111) and Moslemi (102). They all found substantial variability among teachers in grading papers, even as the size and kinds of samples and kinds of investigators they used varied.

Jacoby (70) in 1910 studied the ratings of six astronomy professors on a set of eleven astronomy papers rated on a scale of ten. He found that the average divergence was 1.5 points. While this does not sound like a significant deviation, one must remember that on a scale of one hundred such a deviation would be equivalent to 15 points. If one looks at the individual markings of the raters, one finds four judges passing four papers that the two other judges failed. Ruggles (119) in 1911 had twenty sixth-grade geography papers rated by eleven graduate students in Teachers College. He found that their average deviation from the median scores was 12.15 points, with as much variation between judges as there was of the marks on the twenty papers.

Studies that had great impact on the educational community were those done by Starch and Elliott (134) (135)

(136) at the University of Wisconsin in 1912 and 1913. These studies covered such diverse areas as English, Mathematics and History and used larger numbers of raters than previous studies.

In English, 142 qualified first year english teachers from accredited high schools were used. They were asked to rate two different english examinations written by two different pupils, who had just finished the first year of high school English. Their ratings deviated from the mean by an average of 4.5 points. The range of ratings on one paper was from 65 to over 95 points, while on the other paper the range of grades was from 50 to over 95. Five teachers did not pass one paper and twenty-two teachers did not pass the other paper.

In studying marks in Mathematics, 118 raters were used to rate a final geometry examination. Analysis of the variability of their grading practices found them to have an average deviation from the mean of 7.5 points. The range of these ratings was from 29 to over 90 points, with fifty-four of the teachers failing to pass the paper.

Seventy history teachers were used to study marking practices in History. They were asked to grade a final United States history examination and their average deviation from the mean was 7.7 points with a range of 43-92 points. Approximately forty teachers failed to pass the paper.

Starch's explanation for the great variability in Mathematics, generally perceived by many as a more precise

subject than English or History and therefore less prone to subjective grading procedures, was that "greater certainty of correctness contributes to stricter marking".

Starch (131) further had ten instructors at the University of Wisconsin English Department grade ten written papers of a final examination in Freshman English and found that, even though efforts had been made within the department to have as much uniformity as possible, the grades assigned by these teachers varied as widely as grades assigned in different institutions. He, also, had seven instructors regrade a set of their own papers after a long interval of time. The results indicated that while the mean variation can be reduced (and in this case the reduction was 2.2 points) the marks in some cases continued to vary as much as 40-50 points.

In a similar study of variability of grading practices, Gray used sets of tests given in the areas of Mathematics and English. However, he had a small number of raters - only five "competent" teachers besides the class teacher rate these papers. Differences were found of 20.7 points on the average between judges A and F in rating the mathematics papers, and of 29.7 points between the averages of judges B and D in rating the english papers. One judge failed all but one paper and another judge passed all but one paper. The average deviation in Mathematics was 7.1, while in English it was 9.2. Note that unlike Starch and Elliott, Gray (58) found that the average deviation in Mathematics was less than the average deviation in English.

Concern that some of these variations could be due to the lack of effort to standardize the judgments of judges led Kelly (76) to compare the marks of teachers in New York State with Regent Examiners on the state regent examinations in the years 1889-1895 and 1911-1913. It seemed to him that such built-in controls as the facts that these raters had been at their tasks for several decades and that their judgments were being critically compared would serve to force greater standardization of judgments. He found, however, that on the average teachers passed approximately 10% more students than did the regent examiners. In analyzing 1913 data, he found great variation from subject to subject in percent failed by regents that teachers passed, ranging from 1.9% in Music to 25.7% in Mathematics, with the regent examiners consistently failing more students than did the teachers. In examining the distribution of differences between teachers' marks and regents' marks on the same set of papers for thirty-six schools, he found that 16% of the papers passed by the teachers were failed by the regents.

In another study Kelly (76) had a uniform arithmetic test given by all fifth grade teachers to their respective classes in schools in Orange, New Jersey. The teachers rated their own papers. Then one of the mathematics teachers, considered very systematic, was chosen to develop a rating scheme for the papers. All the grade teachers were asked to rate the papers again, including the mathematics teacher, but this time in line with the scheme developed by

the hand-picked teacher. This technique reduced teacher variation from a maximum of twenty to ten points. However, individual differences among teachers were still apparent. Generally, though, these differences were within the 5% range and this technique increased congruence with the teacher judge from 0.06% to 60%.

Bolton (18) in 1927 took issue with the findings of earlier researchers which supported the low reliability of letter grades. He believed the great diversity in marks was due to variations among the raters in experience, training, knowledge and responsibility to the subject matter. He set up a study in which he very carefully chose twenty- two teachers, and gave them instructions in the content of the examinations they were to rate. He then presented them with a set of twenty-four papers that were selected randomly from sets of tests which had been constructed by the teachers themselves and administered to sixth grade mathematics classes. He did not inform the teachers that they would be participating in an experiment. The average variation of the teachers/pupil ranged from 1.4-10.5 points based on a 100 point scale. Bolton perceived that these findings, like past findings, gave evidence of the uniformity of grading practices. He pointed to such facts as that 86% of the cases varied not more than 10% and that approximately 61% varied no more than 5%. He believed that this indicated an accuracy sufficient for determining whether to pass a student on to a higher level of learning and after all "what more is necessary". He found that the greatest variations

existed among students at the lower ends of the achievement scale. These findings were consistent with the findings of previous researchers.

Bolton, wanting to compare the grading for each of the ten questions used in the examination, selected three of the papers "at random" for analysis, and found that the average deviation ranged from 0-2.6 points, each question having been assigned a total of ten points. Bolton perceived these findings, also, as indicating striking uniformity among grading patterns of experienced and responsible teachers. What he seems to have forgotten is that such a range on a 100 point marking scale would be equivalent to a range of 0-26 points. This means that when he says that 95% of the deviation was less than 3 points, on a 100-point marking scale, which is the type of scale in general use, this would be equivalent to 30 points.

Bolton also took issue with the study by Starch in which ten instructors mark ten papers. Starch found that the teachers' average deviation from the mean was 5.3. Bolton, however, points out that two papers contributed considerably more variation than the other eight; that 83% of the variation was less than 10%; that all of the variations greater than 13% were given by one instructor; and that if one eliminates two instructors, the mean variation ranges from 2.6-5.7, instead of from 2.6-12.3.

While Bolton's point that there is much more uniformity than diversity in many of these studies is valid, it does not support, as he suggests, relatively high reliability of

letter grading. There still exists much too much diversity in this evaluative system -- a system which can have profound effects on people's lives. For example in the Starch study that Bolton examined, on all the papers with the exception of one, eliminating even the two instructors who made the greatest contribution to the mean variation, the range of variation of marks was 10 points or more. Bolton thus does not succeed in presenting evidence that his technique of carefully selecting homogeneous groups of teacher raters improves considerably the range of variations.

Studies done by Lawson (88) and Penfold (111) support this point. Lawson (88), in teaching a sample of teachers of unusually similar background and training in 1940, gave them as a class assignment three specially prepared papers. He asked them to rate the papers and told them that they would be graded on their rating ability, but did not tell them that he was running an experiment on them. The results were as follows: the first paper had a range of grades from 0-90%, the second paper had a range of grades from 20-95%, and the third paper had a range of grades from 10-100%. While Lawson's findings hint of graders still in the learning stage, Penfold (111) got similar results with trained graders. Penfold in 1956 reported that, while in the British system where their School Certificate Graders undergo intensive training in grading essay type examinations to insure greater reliability, examiners often disagree to a significant degree with each other as well as

being often inconsistent even in their own grading procedures. When on two different occasions he had examiners mark 165 junior high school papers, he found similar results even though he had introduced a carefully devised analytic marking scheme.

Research done on the regrading capabilities of teachers also suggest that homogeneous groupings of teacher raters does not necessarily improve the range of variations in teachers' marking practices. These were done by Hulten (68), Eells (45) and Tieg (139).

Hulton in 1923 (68) found that 15 out of 28 teachers rating five compositions in December and again in February, failed the same child in February that they passed in December. Eells in 1930 (45) had a relatively larger sample of teachers. He used ninety-one "largely experienced" teachers. He had them regrade the same material after an interval of eleven weeks, and found amongst them Pearson r's of 0.25-0.51 with probable errors of 0.006-0.008.

With one teacher and ten papers, Tieg (139) reported in 1931 differences in grades ranging from 5-25 points when the papers were remarked over a two month interval. While he found that the mean grade didn't vary much, probably indicating that the teacher was marking on a curve, changing from 80 to 78, six of the papers had variations of over ten points.

Rater correlation, however, can be improved as evidenced by the work of Kelley and Moslemi. Like Kelley, Moslemi (105) reported in 1975 that he was able to improve

rater correlation. He ran a study with three judges on 412 compositions and he was able to obtain correlations as high as 0.947 by utilizing the following techniques: (1) clearly articulated criteria, (2) a rating scale, (3) using a pretest to determine rater reliability, (4) a trial to provide experience, and (5) a review of every 25th composition to monitor judges. Such procedures are not practical, however, and hardly can be expected to be applied in classroom settings.

Critics of the aforementioned findings were wont to point out that in the course of any marking period, variations found on individual papers tend to adjust themselves out over a series of graded papers. Therefore, the final or summative grade is a more valid representation of a student's achievement than any single grade on a single paper. However, studies which look at summative marking patterns between schools and within schools of the same or different grade levels or subjects, including among departments and within departments and even between teachers teaching the same subject, tend to negate this point of view. One finds such studies as far back as the early 1900's and while they become more sophisticated through the years, their findings, vis-a-vis the reliability of letter grading, remain essentially similar.

Studies which compared marking patterns in different schools are those of Miles (94), Carter (76), Alexander (76), Roberts (115), Thompson (136), Dearborn (38), Smith (125), Segel (120), Bixler (12), Wood (148), and Lindquist

(90).

Miles (97) had found that, in the 106 cases in which he averaged the last four years of elementary school grades of students in the elementary schools in Iowa city in 1910 and correlated them with the average of grades received in at least two years of high school for a period of twelve years, he was able to come up with an average correlation of 0.71.

Carter in 1911 (76) who had looked at grades of students who had completed eighth grade classes fed by three elementary schools in Milwaukee, Wisconsin, found that 2/3 of those students from school B had summative marks that fell within the range of the lowest third of those students assigned to school A, and that 2/3 of those students' summative marks from school C fell within the highest range of those summative marks assigned by school B. To determine if there were any real differences in ability of the students coming from the three different elementary schools, Carter studied the rank of these students in the algebra course given at the high school. He found that a larger percentage of school B students were able to maintain their original rank or keep it, indicating that each school was using different marking standards.

Alexander in 1912 (76) studying the variability of teachers' marks in thirty-one schools had found that one-quarter of the teachers failed 8-20% of their students in English, Mathematics, History and Latin while another one quarter failed none.

Roberts (115) reported that in 1917, in studying all

the grades given by eighty-two teachers in Missouri for a period of six years, he found wide variability in the percent of students passed as well as wide variability in the distribution of grades given by individual teachers. Thompson (136), in 1955, reported similar findings in studying the grading practices of thirty-one instructors in freshman English. He found that their mean variation in grade ranged from 0.20-4.20, with almost one-third of them varying significantly.

The Dearborn (38), Smith (125) and Petit (112) studies correlated high school grades with college grades. Their primary interest was in developing prediction measures for college admission purposes. Dearborn (38) claims to have discovered that he could predict 75% of the college students' ranks from such measures. In studying the high school and undergraduate marks of 472 students fed by six cities to the University of Wisconsin in the years 1900-1905, he found correlations of 0.80. However, Smith (125), who in 1910 compared the high school marks and college marks of 120 Liberal Arts students who graduated from the University of Iowa, found a Pearson coefficient of only 0.53, while Petit (115), who in 1912 studied the averages of high school and college freshman grades at Columbia College, found correlations of 0.63.

Those studies were among those done in the early nineteen hundreds. Their findings were suspect to later researchers, who felt that the variations found in these studies might be spurious since they were based on results

obtained from tests that were considered too subjective. (The objective and standardized testing movement was just beginning to have some impact on the educational scene at that time.) But even later studies, done after the testing movement had become an integral part of the schooling process, show similar variability.

Trabue (147) had found in 1924 that the percentage of failures in five large high schools of Northern New Jersey, whose student populations were similar in terms of socio-economic backgrounds and the kinds of teachers to whom they were exposed, ranged from 8-27%. He concluded that these variations had to be due to differences in teacher standards and studied marks assigned by teachers in the same subject in the same high school and substantiated his hypothesis.

Gilky (53) reported in 1929 that, in comparing college grades with high school grades, assigned on the basis of scores obtained on regent's examinations, for all the students who graduated fron New York College for Teachers in the years 1921-23, he found correlations on the two sets of records of 0.498-0.51. The author perceived the correlations as being low and suggested that they might be low because it was more difficult to obtain high grades in college.

Segel (120), who examined six universities and over 10,000 students, found an average correlation of average high school and average college grades of 0.52 with a range of 0.35-0.66. Segel attempted to increase the correlation by studying only those grades given in subjects in which

objective testing was utilized. While his range decreased from 0.47-0.64, his average correlation did not change significantly. On examining other studies which compared general college scholarship with average high school marks, he found that the median of the average correlations of twenty-three different investigators was 0.55 with a range of 0.29-0.77.

And in 1936 Bixler (12) reported that in looking over studies done since the twenties, one finds the same familiar picture. This statement was supported by Wood (151), who had published a good deal of material on measurement on education. He pointed out that studies done by the College Examination Board, the Secondary Education Board, the American Council on Education, as well as state and regional agencies had all emphasized and re-exposed the lack of comparability of marks given by different schools.

By 1963 sophisticated computer and electronic scoring devices were available, and Lindquist (90) made use of them to study a large population of schools and colleges in Iowa City. He had hoped that through the use of regression equations he could improve predictions of college success. He scaled both his high school and college grades by using the lines of "best fit". He found that his correlation did not improve significantly when he scaled either his high school or college grades in line with his regression equations. His median correlation between original high school grades and scaled college grades was 0.629 as compared with the median correlation of 0.621 that he had

obtained by comparing unscaled high school grades with unscaled college grades. If one looks at his range of correlations, one finds that while the scaled correlations are somewhat higher, their range is larger. The scaled correlations range from 0.501-0.736, while the unscaled correlations range from 0.485-0.682.

Lindquist points out that the improvement in median with-in school correlation, obtained by using a scaling technique, was only 0.008 for the with-in school correlation of high school and college grades for the 608 schools in his study. He does, however, mention the results obtained by Bloom and Peters in a similar study which, while it encompassed a much smaller sample (23 secondary schools), nevertheless did show a significant improvement in correlation, going from 0.54 to 0.77 - a gain of 0.23 points. Lindquist suggests that only when there exists wide differences in grading standards can one expect large improvements with such scaling techniques. However, it is notable that even with scaling to account for differences in ability levels of students the median correlation rose to only 0.77.

While the issue of whether the variations under study are related to the variances in students rather than variances in teacher standards remains a moot question in many of these studies, the findings of studies done within schools and particularly within departments, which have a higher probabilty of controlling for confounding variables, generally tend to support the contention of the authors of

the above studies that the variations were indeed due to teacher standards. Variablity of marking practices within schools show much the same pattern in variability as between schools.

Studies that looked at variations in marking practices within schools but between departments were done by Meyer (97), Johnson (71), Crawford (33), Heilman (62), Bass (8) Kirby (79) and Temple University (135).

Meyer (97) collected the marks of forty seasoned professors for a period of five years at the University of Missouri. The data came largely from the College of Liberal Arts. He demonstrated in 1903 that the percentage of assigned letter grades between departments ranged from 1% A's in Chemistry to 55% A's in Philosophy. His findings had such an impact that they fostered a Missouri Plan. (This plan was an attempt to develop a more uniform marking system. It was based on a complicated system of marking by ranking and the use of a normal curve distribution. The author, however, admits that it did not remove many of the inequities of grading).

Johnson (71) demonstrated considerable amounts of variation within schools, among departments, and among teachers within the same department in percent of assigned letter grades when he investigated marks given by various departments in the University High School of the University of Chicago from 1907-1909. He found that failures in Mathematics and English outnumbered those in History and Science and German, while A's were three times more frequent

in Greek than in English.

Crawford (33), who studied 50,000 marks in about five subjects for 4,985 freshmen in six successive classes at Yale in the years 1926-1932, found also that department standards varied considerably as judged by the means of the grades or the percentage of students who passed.

In 1931 Heilman (62) studied students' averages for successive quarters at Colorado State Teachers College and found that repeated computations of correlations yielded coefficients of about 0.60. Since he felt that such a correlation represented poor agreement and couldn't be explained on the basis of differences in ability of students, he proceeded to investigate the reliability of class room tests. He found that 1/3 of the tests were highly reliable, 1/3 of the tests fairly reliable, and 1/3 of the tests inadequate. In analyzing what constituted a highly reliable test, he found that the tests teachers prepare vary widely in length and difficulty, and that to get the most satisfactory degree of reliability from a test, it had to have approximately 300 items.

Bass (8) also studied means of grades. In looking at 396 means of 139,659 grades assigned during the four semesters of the years 1947-1949 at Temple University, he found different departments differing significantly from the course level average for all departments in the mean grade they assigned.

A report put out by Temple University (135) in 1968 found that similar introductory courses in the College of

Education and the College of Liberal Arts gave out dramatically different percentages of grades. The Liberal Arts College gave out 30% D's and F's, while the College of Education gave out only 2% D's and F's for the same courses.

Bass (8) had also reported that the mean grade changed significantly at Temple University from level to level -- above and beyond changes due to departmental variations or semester fluctuations. He found a mean grade for freshmen and sophomores of 2.43, and a mean grade for juniors and seniors of 2.87. For graduate students he found the mean grade was 3.48.

Kirby (79) reported similar findings when he studied the grades of lower division instructors in large institutions of good reputation. He reported in 1962 that the average GPA range was 1.82-3.88, and that the upper division grades had a smaller range and a higher mean than the lower division grades. From his findings Kirby determined that one's grade could change as much as two letters, depending on one's instructor. These findings are, of course, subject to the argument that early in the college game the poorer students get weeded out, leaving the upper divisions with a better quality student. No studies were found that dealt adequately with this issue.

Perhaps the following case identified by Kelly (76) best sums up how such variations come about. Kelly points to the singular case of large differences in the number of failures between the years 1910 and 1911 in one school in New York City. The number of students failing decreased by

500 in one year. Kelley attributes this remarkable feat entirely to the presence of a new principal, whose philosophy of education fore-closed the possibility of large numbers of failures.

Even the studies done on grading practices within departments which have the advantage over all the previous studies mentioned of having the greatest probability of controlling for confounding variables such as differences in abilities and skills of students show the same kinds of correlations when the reliability of letter grades is examined. Such studies were done by Finkelstein (48) Chapman and Hills (29), Ohlson (109), Temple University (135), and Taylor and Constance (134).

Differences of approximately 25% in the number of students exempted from the final exam were found by Finkelstein (48) in 1913 between two different instructors teaching the same subject but in different terms to practically the same body of 250 Cornell students.

Chapman and Hills (29) in 1918 supported these findings. They, too, found wide variations in the percentage of students passed and the distribution of grades of college instructors, even within the same department, where they found one instructor giving 400% more E's than another.

Ohlson (109) in 1927 looked at the percentage of assigned grades in the Everett (Washington) High School of 200 boys and 306 girls. He found that correlations within departments ranged from 0.25 in English to 0.12 in

Vocational Departments.

Around 1968 when Temple University (135) did a study of the marks assigned by different instructors for the same course, they found that one instructor awarded 20% of his students A's while another gave no A's. These authors determined that two-thirds of the students taking that course would have found themselves receiving an unsatisfactory grade simply on the basis of having been assigned one instructor rather than another.

In 1933 Taylor and Constance (134) felt that comparisons of successive quarters or successive semesters, which many researchers had been studying, could be low because of interactive effects due to such intervening variables as:

1. faculty judgments can become relatively clouded by previous judgments in successive quarters,
2. fluctuations in student interest and effort are probably greater between successive quarters,
3. similar programs of work are least apt to be pursued in successive quarters.

They, therefore, compared grades within the same department between alternate quarters, and used a weighting system based on the mean deviation of the grade in any given class. Their correlations ranged from 0.58-0.90, with higher correlations for grades received by women. While they did seem to improve their correlations by this device

of looking at alternate quarters and the inclusion of a weighting factor to account for differences in ability levels, the authors felt compelled to suggest that some of the correlations might be spurious due to the fact they they were aware that professors discussed grades within departments. Thus they perceived that some of the ratings may have been based on reputation rather than performance. When they analyzed successive quarters a year apart, they found correlations similar to those of other researchers who had examined successive semesters such as Toop, McPhail, Kernauser, Cleeton, Crawford, and Wood--all of whom had correlations of the order of 0.66.

With the development of standardized testing, researchers had a new criterion measure to use in comparing letter grades. These tests were generally found to be more reliable than teacher made tests. (Some educators even went so far as to suggest that the scores received on these tests be used to replace letter grading.) Among those researchers who did studies comparing the scores on such tests with letter grades were Ohlson (109), Segel (120), Gilkey (53), Bixler (12), Wood (148), Twerlinger (144), Hills, Klock and Bush (67), and Klugh and Bierley (83).

Ohlson (109) in 1927 reported on a study in which he correlated grades with Terman I.Q. scores and found correlations of only around 0.38. His sample included 200 boys and 306 girls from the Everett (Washington) High School. These compared with the findings of Jordan (73) who did a similar study, except that he used Army Alpha scores

instead of I.Q. scores, and came up with correlations of about 0.321.

Segel (120), who correlated grades with mental ability scores, reported in 1934 that he found correlations ranging from 0.27-0.65. He also studied the correlation between scholastic aptitude tests and average college grades and found that these ranged from 0.29-0.60. Subsequent studies supported such correlations.

Gilkey (53) in 1929 reported correlations between scores on "intelligence" tests and marks at colleges such as Columbia, Brown, Stanford, and the Universities of Wisconsin, Chicago, South Dakota, and California ranging from 0.27-0.66. And in 1936, Bixler (12) found that a high school grade of 85 from fifty high schools could mean a score of from 75-180 on a scholastic aptitude test.

Wood (148) in 1939 studied fifteen students, comparing their grades with a standard achievement score in the same subject. He found only one instance in which the grades and scores were comparable. In six cases he found the grades and scores separated by ten points and in eight cases, the grades and scores were separated by between 40-50 points. Wood's data revealed that the student who had received the highest mark on that national achievement test, had received a grade of only 72 on his report card, while a student who had tested in the 29th percentile had received a grade of 73 on his report card.

Widely varying correlations ranging from -0.1-0.65 in one school and correlations ranging from 0.25-0.70 in the

other school, were found by Twerlinger (144). In 1968 Twerlinger correlated the teacher assigned marks in two public schools in Nashville, Tennessee with scores obtained on an Otis Quick Scoring Test. He confined his sample to the first class taught each day by thirty-eight teachers. Of significant interest is the fact that the median of the correlations of average class score and average assigned grade in both of the schools was 0.57. This finding led the author to suggest that ability level of a class is not reflected in the class mark. He also found that Mathematics and Social Studies had the widest range of correlations, with Biology having the smallest.

Multiple correlational analyses were undertaken by Hills, Klock and Bush (67). They used the scores on the verbal and mathematics sections of the scholastic aptitude test. They had collected data on samples that came from publicly supported institutions of higher education in Georgia. Separate prediction equations, for the classes entering in 1958 for each of the six institutions involved, were set up. Predictor correlations were determined on the basis of the scholastic aptitude test scores and the high school and first year college GPA's of those students who entered these publicly supported institutions in the year 1957. The predictor correlations ranged from 0.46-0.82; the predicted correlations ranged from 0.31-0.82. The average multiple correlation yielded by the 1957 data was 0.65.

Klugh and Bierley (81) did a similar study, using as their predictors four years of high school grades, and a

score obtained from an ability test. The student GPA at the end of the college semester was used as a criterion measure. They eliminated courses carrying only one-half credit per semester. They studied all the students who entered Alma college in the fall of 1956 and 1957, controlling for sex and year entered. Their multiple correlations ranged from 0.661-0.782 with women having the larger multiple correlations.

The general conclusion one reaches from all these variant studies is that no matter what type of school or grade level or subject one examines, no matter how uniform departmental standards seem to be, or the student population in terms of ability, socioeconomic background, geographical location, or other relevant factors, on the average, grades can be expected to have a reliability coefficient of approximately 0.60. This means that on the average 36% of the variance between students is accounted for by letter grading, while over 60% remains unaccounted for, making letter grading a measurement or evaluative process of low reliability. Such abundant scientific evidence in almost any other enterprise would have produced appropriate changes during the eighty years that this problem has been studied. Educators are simply working with a system of evaluation which has a level of precision that ignores the scientific developments of the last 300 years and the accumulated knowledge of the last eighty years presented herein.

4.--Administrative Functionality

As mentioned in the section on the history of letter grading, grading was introduced to improve the efficiency with which a pupil's progress was being reported in the schools. It came to take on other functions as well, as the schools found themselves having to sort out students for awards, jobs and college admission procedures. All researchers who have concerned themselves with marking practices have spoken of this aspect of grading. And in this regard grading serves administrators well.

This administrative functionality is related to the ease with which letter grading can be averaged, recorded, and interpreted. For it is these qualities that make letter grading readily useable, particularly in a time-bound setting, for the many different administrative functions that have become an integral part of educational systems. The administrative functions that letter grading is regarded as fulfilling are:

1. reportorial, i.e. providing a simple device by which students, educators, parents and potential employers can have some reflection of the student's relative performance at the grading institution,
2. one of guidance in educational and vocational matters, and
3. one of selection by which awards, college entrance, placement and promotion can be accomplished.

Starting with Finkelstein in 1813 (49), and through the

years with Trabue in 1924 (142), Crooks in 1933 (35), Wrinkle in 1947 (150), Adams and Togerson in 1964 (2), Miller in 1966 (95), Thorndike in 1970 (138) and Gronlund in 1974 (60), to name just a few of the researchers who have concerned themselves with marking systems, these administrative functions, which letter grading serves so well, were spoken of as being prime functions of letter grading. The awareness that this administrative efficiency of letter grading was more important to the educational community than letter grading's function as an evaluative tool doesn't seem to become a prominent issue until around 1933. It was then that Crooks (34), in writing on the problems of marks and marking, stated that the efficient clerical administration of the marking system is one phase of the marking problem of marking systems that requires further study, while Ayer (5) in the same time frame was admonishing his readers to consider the important administrative and pedagogical values attached to school marks.

Those who continued to survey the problems related to letter grading practices reached very much the same conclusions. In 1935 Wrinkle (155) referred to grades as the most effective and efficient device for serving the administrative functions of placement, promotion, transfer and graduation, even as he took cognizance of the fact that letter grades were not wholly adequate as measures of educational outcomes. And a University of California report, published in 1963 at Berkeley (145) studying methods

of evaluating students, pointed out that one of the mainstays of traditional letter grading is the numerous administrative purposes, both within and without the institution, that are served by it.

Miller in 1966 (95), in attempting to analyze why letter grades have persisted on the educational scene for such a long time, in spite of strong evidence as to their low reliability and validity, came to the conclusion that their administrative functionality played a most crucial role, and that the greatest recommendation for traditional grading was its administrative efficiency.

Responses to interviews sent to a nationwide sampling of subscribers of Education U.S.A. (106), indicated in 1972 that one of the major reasons for the usefulness of traditional grading practices is that they are a convenient way in which to sort out those students in high school and college for awards as well as selection procedures.

The heart of the problem was probably exposed by Thorndike (138), who after a lifetime of studying grading practices, said in 1970 that "the literature on marks and marking over the last fifty years seems to have missed the mark" because "it has been insensitive to the very real limits of time and precision of judgments and skill in assessments within which a typical teacher operates". But it was Gronlund (60) who summarized the matter most concisely when he stated in 1974 that the advantages of traditional grading were that it was easy to use and thus convenient for maintaining school records, and that it

allowed for ease of averaging and provided fairly good predictions of future achievement.

Supporting Gronlund is a statement in a 1976 Change policy paper on the Testing and Grading of Students (96) which concluded that the purpose of evaluation seems to have become the assigning of letter symbols largely for record keeping purposes.

The literature on marks and marking systems leaves little doubt that the most pressing rationale for hanging on to a grading system that has been exceedingly well documented with sufficient research to demonstrate its low validity and reliability, whether one calls it an evaluation system or a measurement system as has been suggested by some concerned parties, is related to the facts that letter grading is perceived as the most administratively functional system yet devised and that letter grading serves primarily as a record-keeping tool and not primarily as the evaluative tool it is touted to be. Thus for any alternative system to compete with letter grading, it must compete favorably in these regards.

5.--Communicative-Constructiveness

Constructive communication between school and student between school and parent and between school and other educational institutions is consistently mentioned in the literature on grading as being a crucial function of grading. However, letter grading's role as a constructive communicator has been suspect according to the references

that follow.

Wrinkle (156) in 1947 found that the number one fallacy in the use of letter grades in reporting achievement is that anyone can tell from the mark assigned what the student's level of achievement is or what progress the student is making. Chansky (27) pointed out in a 1964 article that one of the many facts that marks obscure is the student's varying abilities in any one subject. He suggests that a student can be talented in one aspect of a subject, but less talented in another, yet because marks tend to be an averaging of achievement, they obscure areas of excellence. A publication by the National School Public Relations Association on Grading and Reporting said in 1972 that if the purpose of traditional grades is to communicate, grading could stand improvement.

While most educators had perceived the communicative aspect of marks as one of their more important functions, Twerlinger in 1971 (144), in clarifying the concept of communication, negated the communicative role of marks as a separate and single function apart from the administrative, guidance and informational functions of marks. Twerlinger perceived that in actual practice, all of these functions were simply special cases of the more general role that marks play -- which is that of a communication system. He further defines a communication system as one that sends out, by its many transmitters, messages that have the same interpretation for its many receivers, as well as one that employs a set of symbols that have the same meaning to all

its users. Twerlinger specifies that unless such is the case, even the most carefully designed system will fail to serve as an effective vehicle of communication.

It is difficult to comprehend that marks can serve as an effective vehicle of communication since the many studies on the low reliability and validity of letter grading simply attest to the fact that what marks really mean can not only be different for different people, but can even change in meaning from time to time for the same person. Research studies on the communicative aspect of marks has been meager. However, two interview studies, one by Barton and the other by Birney, and a case study by Bolmeier do contain findings which relate to this issue.

The Barton study reported in 1925 (7), that of 1,513 pupils, ranging in age from 13-21 and of various degrees of ability, achievement and socio-economic status, interviewed, only 42.8% felt that marks gave both them and their parents a true estimate of what they had accomplished. The Bolmeier (16) case study reported in 1943 indicated quite clearly that marks are interpreted in varying ways according to one's preconceived notions of what a mark should represent. And the Birney study found in 1965 (11) that students of the 1959 class at Amherst seemed to agree that marks "tell little".

While there is hardly any research that deals specifically with the issue of the communicative value of letter grading, the abundant evidence on the low validity of marks supports the contention of those herein referenced

that letter grading as a communication system fails to adequately convey a precise understanding of what is being evaluated.

6.--Negative-Side-Effects

Thorndike (138) in his 1970 article summed up what most researchers perceived as the negative side-effects of letter grading practices. They were: (1) the dehabilitating aspect of chronic failure of those trying to meet a standard for which they are not ready, (2) the undue competitiveness and the resulting anxiety that ensues, (3) the widespread cheating and dishonesty that seem to result, and (4) the distorted educational value patterns which make the appearance rather than the substance of learning the important aspect of learning. Thorndike's summation merely reflects what has been said by such educators as Odell (111), Hillbrand (66), Mason (66), Crew (66), Smith (131), DePencier (40), and Johnson and Johnson (74) over the years in regard to the negative concomitants of letter grading.

The use of marks as an incentive for learning was deplored by Odell (111) in 1930. He found that their use often encouraged overwork as well as widespread cheating. Hillbrand (64) emphasized some of the generalized reactions to letter grading in 1931 when he singled out the remarks of some prominent authorities. He quoted Mason, the president of the University of Chicago, who said in 1928 that marking is a "hinderance to genuine learning", and the president of AAUP, Crew, who in 1930 stated that marks interfered with "a

free and easy meeting of student and teacher" and "diverted a student's attention from the main purpose of learning". Numerous studies were cited by Smith (131) and others in 1942 which indicated that teachers believe that the competitive features of marking have been developed to the extent that they threaten pupils, and in 1951 a number of studies which concerned themselves with the anti-social attitudes and behaviors bred by the competitive aspects of marks, were reported by DePencier (40). In 1974 Johnson and Johnson reflected the concerns of many, (72) when they questioned the ethics of placing an individual in a predominantly competitive structure where the vast majority of the students must continually experience failure.

The developing psychology of the 20's encouraged research on the effects of achievement vis-a-vis aspiration levels, self-image, attitudes and anxiety, since empirical evidence seemed to indicate that one's perception of one's achievement can have considerable influence on one's affective behaviors.

The effect of success and failure on aspiration level was looked at by Child and Whiting in 1949 (30). He had 151 men, taking a course in psychology that he taught, write three descriptive incidents in which they: (1) experienced only frustration, (2) experienced frustration, but achieved goals anyway, and (3) achieved goals with no appreciable frustration. Child and Whiting then did systematic analyses of these events and found that success gradually leads to the raising of one's aspiration level, and that the stronger

the success the greater the probability of a rise in aspiration level. He found just the opposite for failure and that failure is more likely than success to lead to withdrawal in the form of avoidance of setting any aspiration level.

Tending to support such findings were those studies that looked at the relationships between self-image and achievement, such as those done by Kurtz (85), Shaw (121), Brookover (23) and Torshen (140,141). Kurtz (85) published a report in 1951 in which he found that positive achievers were not only happier than negative achievers in traditional classroom settings, but when he looked at the characteristics of positive and negative achievers, he found that, in general, positive achievers rated higher on such characteristics as relationships with peers, physical and mental well-being, academic inclinations and aspirations, and relationships at home. This data was obtained by interviewing 200 students, their parents and their teachers from a midwestern city. Kurtz suggests that these findings are in line with the Lecky theory that students' opinions of themselves influence their achievement in school.

Shaw in 1960 (121) analyzed the long range effects of grading. He obtained data on the GPA's of students, 36 males and 17 females, that had been selected as representing underachievers, and 36 males and 45 females classified as overachievers. This sample had been selected from the upper 25% of a larger group of 6000 students in two fairly representative high schools, whose ability levels had been

predetermined by testing. He found that while significant differences in GPA's appeared in grades one through three among the males, the differences increased in significance at each grade level up to the tenth grade. At this point in time the significance decreased somewhat, nevertheless it remained significant at the 0.01 level. In the case of the females, the underachievers had higher GPA's in grades one through five, though not significantly higher. By grades six through ten the achievers begin to get significantly higher GPA's increasing in significance each year. The findings suggest that whatever the variable responsible for underachievement, it becomes an increasingly negative force.

In 1964 Brookover (23) reported on a study of 1050 seventh graders, half male and half female, in which he found that: (1) general self-concept and academic performance were positively and significantly related (0.57 for males and 0.57 for females, even when I.Q. was controlled), (2) that specific self-concepts of ability were significantly better predictors of specific subject achievement than was general self-concept, and (3) that general self-concept was positively and significantly related to a student's perception of how a few significant persons evaluated him/her.

Torshen's (140, 141) 1973 data, from a sample of 318 fifth grade students of varying socio- economic classes, revealed through multiple regression analyses that norm-referenced grades assigned by teachers were significantly related to the students' self-concepts and mental health.

Her variables consisted of twelve indices of self-concept, five of mental health, norm-referenced grades, achievement scores and other measures. From an earlier study reported in 1968, she had concluded that teachers' evaluation of students' cognitive achievement have a greater influence upon students' self-concepts than do their objective achievement test evaluations. With the former she found an r of 0.41 and with the latter an r of 0.33.

One finds, however, that it is difficult to interpret studies of this sort because of the lack of clarity as to which variable is cause and which is effect. However, when one looks at studies like Feather's (47), Weiner's (147) and Modu's (99), one finds that they tend to support the fact that changes in affective behaviors and cognitive achievement can indeed result from one's perception of one's cognitive abilities.

Feather in 1965 (47) investigated the relationship between an individual's orientation towards a task, his expectation of the task and his initial experience with the task in terms of success or failure. To do this, Feather set up a rather interesting experiment. He had seventy-two college students work at tasks consisting of fifteen anagrams. The first five anagrams were unsolvable and given to half of the subjects; the second five were easy anagrams and given to the remaining half. The rest of the anagrams were of approximately 50% difficulty. All the students were given the latter, but half were told that the anagrams were easy, while the other half were told that the anagrams were

harder than most. His results indicated that prior success or failure influences an individual's expectations of later success and actual performance. Supporting Feather's findings were those of Weiner (147), who reported in 1968 that the effects of continued success and continual failure do affect the persistance of certain type students. Using a sample of sixty male students, he tested for their anxiety levels and need for achievement. The upper and lower quartiles were then subjected to tasks, and half of the students were told that 70% of college students tested were able to complete the tasks within a specified amount of time. The other half were told that only 30% of college students tested were able to complete the tasks in the given amount of time. The first group was allowed to complete every task before being told that the time was up, while the second group was told that the time was up before they had completed the task. Weiner found that subjects high in achievement undertook more trials in the failure condition than they did in the success condition. Subjects in the lowest quartile of achievement, however, persisted longer in the success condition.

Modu (99) reported in 1969 on an investigation of the relationship between affective characteristics such as aspiration level, life-goals, interpersonal competencies leadership, and grades. He explored the extent to which perceived changes in cognitive achievement influence these variables and found a significant relationship between these variables and grade discrepancies. The study involved 2,433

students from sixteen colleges and universities, and these students were studied for a period of a year. He found that the relationship held across sex and persisted even when differences in academic aptitude and student's satisfaction with college choice were controlled. Changes in self-ratings and leadership qualities were most noticeable, but even changes in interest were found to be closely associated with cognitive change.

Supporting the anxiety producing aspects of some of these findings are studies done by Barton, Phillips, and Osterhouse. Barton in 1925 (7) reported on the results of a questionnaire administered to 1,513 pupils in four different Eastern schools chosen by authorities as representative in terms of ability levels, achievement levels and socioeconomic strata. Forty-four and nine-tenths percent of the girls said that they suffered considerable strain from marks. Thirty-seven and three-tenths percent of the total sampled said that they were frightened by marks, twenty-one percent said that marks made them angry, and approximately sixteen percent said that they were indifferent to marks.

Phillips reported in 1962 on the relationship between anxiety and achievement and the interactive effects of sex and class on that relationship. He studied 759 7th grade students in Texas and found evidence that subjects of low anxiety seemed to achieve at a higher level than those with high anxiety, with sex and social class having an interactive effect. Middle class males and lower class females demonstrated lower achievement results with an

increase in anxiety. Lower class males seemed to be the only group that increased achievement with an increase in anxiety.

In 1975 Osterhouse found similiar results when he found tha high anxiety level in the classroom appeared to debilitate exam performance more than a low level of classroom anxiety. He reported a significant linear trend in amount of inner level anxiety and environmentally induced anxiety which he found combine to effect performance on exams of moderate and high test anxiety subjects.

A study done by Bostrom, Vlandis, and Rosenbaum (19) in 1961 found that marks can even affect attitudes towards such social problems as legalized gambling and socialized medicine. On a sample of three groups of twenty students each from the University of Hawaii, matched in terms of sex, age, college class and scholastic apptitude, but varying widely as expressed by cumulative grade-point average, they found that good grades serve a reinforcing role in significantly changing attitudes concerning the afore-mentioned problems in contrast to poor or no grades.

A few studies have attempted to investigate the amount of cheating many educators claim letter grading generates. Knowlton and Hamerlynck in 1967 (84) reported on a study that they did in which they found no fewer than 81% of the students admitting to cheating in college. Forty percent admitted that they cheated in some form or another regularly. This sample was drawn from both rural and urban universities.

Fala (46), who had interviewed 5000 students, reported in 1968 that at least half of them admitted to incidents of cheating. Specifically he found that the highest incidence of cheating was among the weak students, men, career-oriented majors and those in school with other than academic interests. Supporting this finding was the work of Bowers (19) who had done a national survey and who reported in 1968 that at least fifty percent of his responding students admitted to cheating.

Barton (7) had found through his questionnaire that, depending on the high school, from 3-34.6% of the students admitted to forging their parent's signature on the report card. These incidents were most frequent at the age of sixteen and girls were more addicted than boys.

The findings of these studies related to the effects of letter grading present a consistent pattern which gives credence to the empirical observations of many educators that letter grading does tend to play a destructive role in the learning process vis-a-vis aspiration levels, self-image, attitudes and values.

7. Motivational Value

The important role that motivation plays in students acquiring knowledge and skills has long been recognized as a crucial determinant in the learning process. A number of educators have perceived letter grading as fullfilling such a role. In reviewing the early literature on the subject, one finds Odell in 1930 (111) listing motivation as one of

the major functions of marks, even while there were still those educators like Forman (49) who were deploring the fear attitude fostered by marks. Continued concern is expressed over the years concerning the inefficiency of letter grading-induced motivation, as indicated by Fraser (50) in 1937, nevertheless a review by Norsted in 1938 (107) stated that the consensus of opinion supports the motivational aspect of marks.

In 1947 Wrinkle (150) took up the cudgels against the use of marks as motivators. He reported that while marks were indeed used as such to threaten and encourage both slow and able learners, using them for motivating students in these ways was merely a temporary substitute for what should be the real motivators - interest and values. Wrinkle also determined that one of the reasons students generally do not pursue the learning process after they leave school was because they had been conditioned to what was a temporary motivator rather than a more intrinsic and permanent one.

Nonetheless over the years grades continued to be perceived as motivators even as grades were being dropped in the growing number of nongraded schools that were being established. Adams and Torgerson in 1964 (2), Miller in 1967 (95), and Twerlinger in 1971 (149), all attest to the perception of grades as motivators.

Twerlinger (144), however, does make a distinction vis-a-vis the motivational aspects of grading that previous educators did not. For Twerlinger, the motivational function of grading is not so much a purpose of evaluation

as it is a consequence of evaluation. While he perceived that an evaluation system could be a constructive motivator in the learning process, he did not perceive letter grading used as it was in most schools as being constructive. He found the process replete with anxiety and undue pressure, and like Wrinkle, rewarding artificial or extrinsic values rather than the intrinsic joy that can come from the very learning process itself. Twerlinger also saw the letter grading process as decreasing the effectiveness of a teacher, because it forced the teacher to set unrealistic expectations for students. He argued that it was inappropriate to set expectations for students not consonant with their preparation and/or intellectual maturity.

By 1974 we do find an educator, Gronlund (62), publishing a grading review that does not include motivation as one of the major functions of letter grading. Gronlund perceived that the degree to which a marking system could serve as a motivator depended to a large extent on the way it was used. His experiences and research knowledge led him to the conclusion that indications of good progress can be reinforcing and that low marks can also result in increased effort, but only when they follow some positive evaluation of progress and point to specific areas in need of improvement. He perceived constant feed-back as crucial to contributing to student motivation, and saw such feed-back as providing the kind of short-term goals that make it possible for a student to focus on areas of weakness. This finding is supported by a study done by Page (113), who

reported in 1958 that, even on objective tests, comments made by the teacher on such tests led to higher achievement scores than tests scored with no comments. This study had been run on seventy-four selected secondary school teachers from three school districts and 2,139 unknowing students.

A limited amount of research has been done in the area of the motivational value of letter grading. Among those who did interview studies were Barton (7), Tieg (139) and Burke (24). In 1925 Barton (7) reported, that in interviewing 1,513 students in four major cities in the eastern part of the United States, he found that 51.2% said that high marks made them work harder, while almost 66% said that they believed a low mark made them work harder. When asked if they would work as hard with no marks, 55.2% said no and 39.4 % said yes. However, Tieg (144) reported in 1931 that his interviews of students revealed that 90% of them believed that they worked harder because of good marks and that 90% believed that they worked harder because of poor marks. Perhaps by the 1960's students had become more sophisticated about marks because Burke (23) reported on a study in 1969, done at the University of Minnesota, in which he found that only 7.9 % of the students thought that grades were helpful in giving them extrinsic motivation.

Studies which tried to hone in on more specific variables relating to motivational factors were those done by Bostrum, Vlandis and Rosenbaum (19), Birney (11) and Heist (63). Bostrum, Vlandis and Rosenbaum (17) studied three groups of twenty students each, matched in terms of

sex, age and college class at the University of Hawaii. They found that differences in attitude, motivation and academic drive were related to academic success. The implications of these findings would seem to indicate that there are motivational factors in individuals developed by forces other than grades. Supporting such a conclusion is the empirical finding of Goodlad (56), who directed a nongraded elementary school which did not use letter grading. He claimed in 1963 that empirical observation of children functioning in such a setting found no evidence to suggest that these youngsters were any less motivated than those in traditional classrooms.

The study by Birney in 1964 (11) on a previous class at Amherst suggested that failing or near failing grades spurred greater effort on the part of failing or near failing students and that higher grades seemed to be more related to course interest. His findings further suggested that in a course of low interest, high grades seem to lessen study, while in a course of high interest, high grades seem to stimulate effort. In the study done by Heist (63), he reported in 1965 that the effect of low grades on bright students was unpredictable.

The question of whether letter grading has the kind of motivational value some people attribute to it remains moot. The evidence tends to suggest that a sizable proportion of students are not motivated to do the kind of serious work that academia has established as its objective, but the evidence of research is meager in this regard. A more

definitive answer awaits more extensive research.

Part II

Alternative Marking Systems

Present evidence indicates that while several alternative marking and reporting systems have been and are still being used, particularly in the lower level grades and in the more affluent communities, the large majority of educational institutions continue to use letter grading. A rash of people have written on the pros and cons of the various alternative systems. People such as Gronlund (62), Terwilliger (149), Kirschenbaum, Napier and Simon (82), Thorndike (143), and Gilman (55), to name a few, are among those whose writings one might turn to for such information.

This section, therefore, will look at the perceived advantages and disadvantages of some of the more commonly used alternative systems in the light of the criteria of the theoretical framework developed in this dissertation. The alternative systems will be analyzed in terms of their validity, reliability, side-effects, constructive communication, motivational value and administrative functionality. For the sake of greater clarity the category "administrative functionality" has been subdivided into its five original summary categories (see page 4).

Two major types of alternative systems are in use. They are:

1. those that use no letter grading. (These include written evaluations, either by teacher or pupil, parent-teacher conferences, contract performances, check-lists of learning objectives and two-symbol systems such as P/F, C/NC and S/U), and
2. those that use letter grades but base them on criteria other than the usual normative evaluation procedures in which students' performances are evaluated in terms of class norms. (These include contract performances, check-lists of learning objectives, mastery learning units and performances based on norms such as local, national, or a student's own achievement test norms.)

Table 1 is presented as a short summary of the extent to which the alternative systems meet the criteria of the theoretical framework. A plus sign was assigned if the evaluation system was considered to rate well in any given category, and a negative sign if the system was considered to rate poorly relative to the category. A question mark was used when there did not seem to be enough evidence to make an evaluation.

TABLE 1

GROSS EVALUATIONS OF ALTERNATIVE SYSTEMS IN LIGHT OF THE THEORETICAL FRAMEWORK

	CATEGORIES*									
SYSTEM	1	2	3	4	5	6	7	8	9	10
NO GRADES										
Evaluations	-	-	?	?	+	-	-	+	-	-
Conferences	-	-	?	?	-	-	-	-	?	+
Two-Symbols	+	+	-	-	?	-	-	-	-	-
Check-Lists	-	-	+	?	+	+	+	+	+	+
Contracts	-	-	+	?	+	+	+	+	+	+
LETTER GRADES										
Check-Lists	+	+	-	-	+	-	-	-	-	-
Contracts	+	+	-	-	+	-	-	-	-	-
Norm-Based	+	+	-	-	+	-	-	-	-	-
Mastery	+	+	-	-	+	-	-	-	-	-

*Categories: 1: easy to record, 2: easy to average,
3: easy to interpret, 4: good predictors of college success,
5: needed for college entry, 6: highly valid, 7: highly reliable,
8: few side-effects, 9: good communicator,
10: good motivator.

These ratings were derived from the following analyses:

1. Easy to Record

Marking systems based on scales that encompass two to five points have the advantage over all other type systems of being administratively simple to record, for they provide for the issuance of one simple symbol in each subject being evaluated. All the other types of systems in use involve detailed reports.

2. Easy to Average

While theoretically any scale from two to five points can be averaged, the five-point marking scale differentiates students more than other scales. It is also more readily amenable to a ranking system than smaller scales. This makes the five-point system the most efficient for the many selective processes in which all educational institutions are involved. Systems which use no scaling devices would have to develop some kind of rating scale for averaging and ranking purposes, if these two functions were perceived as necessary.

3. Easy to Interpret

It is a common perception that letter grades are easy to interpret. There is no doubt that the conceptual meanings of the symbols themselves are clear. Two symbol systems like P/F are precise. They are either/or systems which indicate that

either you have fulfilled the basic requirements of a course, or you have not. The five-symbol systems stretch out the either situations into four categories, which are also conceptually clear. In the either category an A stands for excellent work, a B stands for good work, a C stands for fair work, a D stands for poor work and in the or category an F stands for failing work. These meanings are readily understood by most people. However, the abundant evidence on validity and the evidence on letter grading's value as a communication system certainly indicate that letter grading has different meanings not only for different people but even for the same persons at different times. Wherein, then, lies the problem?

While letter grading is conceptually clear, its operational meanings are vague. And it is this vagueness that is responsible for its ambiguities and hence low validity (see section on validity). It is the great variations in the specific performance objectives from which these symbols are derived that are the crux of the problem. Only in an evaluative process where the summative evaluation is criterion-referenced to precise objectives and these objectives become a part of the summative process, can the process be made patently clear.

Evaluations based on check-lists of objectives, contract systems, or mastery strategies do have operational definitions based on clearly defined objectives, and can be made patently clear. However, letter grading often accompanies such systems. While the probability is high that the validity of the summative letter grading process would increase when derived from such objectives, the interpretation of the grades beyond the classroom could hardly be expected to improve (see the section on objectives in the literature review).

As for the written evaluations or conferences, several authors have suggested that such evaluations tend to become too subjective, creating extensive opportunities for misinterpretation.

4.--Predictors of College Success

The only evaluation system for which correlations for predicting college success seem to have been attempted has been letter grading based on class norms. In the section on reliability in the literature review, the evidence indicated that on the average one could expect correlations of high school and college grades to run about 0.60. Such a correlation can account on the average for only 36% of the variance among students, leaving on the average 64% of the variance unaccounted for.

Thus, while it is legitimate to say that such correlations are the best predictors for college success that we have, we can hardly say that they are good predictors. There are obviously other factors which contribute to college success that are not being picked up by grades and it behooves us in the names of science and humanity to seek these out.

5.--College Entry Prerequisites

The general consensus of opinion among most parents and high school administrators has been that letter grades are needed for the college entry process. Yet one survey of thirty colleges found an overwhelming response to accepting students without grades, and another study found support among even prestigious colleges to accept students without grades (79).

The American Association of Collegiate Registrars and Admissions Officers did a study of 1,301 of its members, who represent one-half of the institutions listed in the Educational Directory of Higher Education in the years 1970-71. They found that the two and four year colleges would welcome high school applicants with non-traditional grading (21). While this may or may not reflect the majority of colleges in the United States, it is a large number. It becomes obvious, then, that any evaluation system which

can provide enough evidence for college admission officers to make good predictions, not only for their successful matriculates but for their graduates as well, has the possibility of replacing letter grading.

6.--Validity

Many researchers have effectively demonstrated that letter grades can have almost as many different operational meanings as there are people using them. The consensus of opinion amongst those who have studied letter grading in depth seems to be that the validity of letter grading can increase only when educational outcomes are clearly delineated. The systems that show promise of providing such objectives are mastery learning systems, check-lists of objectives and contract systems, provided they are not obfuscated by letter grading. Written evaluations and conferences could also be based on a pre-determined set of objectives. However, they are too likely to deteriorate into subjective analyses, a problem that must be avoided if one wants an evaluative process which is highly valid.

Performances relative to some norms run into the problem of the validity of the tests on which the norms are established. The controversy that is currently swirling around the content validity of tests provides evidence of this contention

(95). The fact that these systems are used in conjunction with letter grades also makes them poor prospects for an evaluation system that is highly valid.

While the validity of some of these systems might be increased by introducing clearly delineated objectives, the use of content objectives lacking in generality would mitigate against the kind of validity needed for a standardized measure, i.e. one that can be used in any classroom. One could probably establish "internal" validity within the framework of a particular performance or area of content with such concise verbalizations, but certainly not "external" validity which allows generalization to all populations. It is this step in the process of developing an evaluation system that remains to be taken. And it is why this author chose to examine the use of the cognitive objectives for an evaluation system.

7.--Reliability

The section in the review of the literature on reliability presents abundant evidence that letter grading is not a highly reliable measure. The inclusion into such a system of pre-established content objectives as is the case with mastery systems, contract systems, and check-lists of objectives could serve to increase the reliability

of such a system. But once again, it must be pointed out that while such a system might be highly reliable for any given person in any given situation, it could hardly be expected to cut across content areas or even across classrooms teaching the same subjects. Good teaching practices allow for sufficient flexibility in choosing content so that the teaching process can be relevant to the needs of any given class. The kinds of specific and cumbersome objectives in general use, while a step in the right direction, tend to be restrictive in this regard.

8. Negative Side-effects.

To reduce negative side-effects, according to the review of the literature, it would seem one needs a system which gives a student the opportunity to maximize his/her achievement, thereby bolstering a student's self-esteem. For doing so seems to result not only in improving self-image, but to result in reducing undue anxiety, in raising aspiration level, in weakening anti-social attitudes and behaviors, and reducing cheating. Check lists of objectives, contract performance, and mastery learning systems would seem to have the most potential for providing such a possibility with their potential being even greater without the use of letter grading.

9. Constructive Communicators

The literature review section on the communicative constructiveness of letter grading indicates that for an evaluative process to be constructive it must provide feedback. Most students do not perceive letter grading as adequate in this respect. The only types of systems herein mentioned that have the potential of doing just that are those in which pre-established objectives are indicated on report cards, or are discussed in conferences.

10.--Good Motivators

The problems related to the motivational aspects of present alternative systems have not been studied in any systematic way. However, the available evidence indicates that the feedback nature of systems like check lists, contracts, and mastery strategies does encourage the kind of intrinsic motivation that makes for the kind of highly motivated student who learns for the sheer joy of learning.

Part III

Related Educational Mechanisms

1.--Mastery Learning

The concept of mastering learning materials has been

the essence of all learning. In the early days of the United States, when the learning process was largely reading, writing and arithmetic, and delegated to teachers or tutors in schools, they determined on an individual basis when a student was ready for a higher level of learning. Generally this seemed to be done when the teacher or an educational board perceived through oral or written testing procedures that a lower level of learning had been mastered. The student, according to the literature, was then given a higher level reading or spelling or arithmetic assignment. However, as the industrial revolution opened up new vistas for the American community, more of its members wanted to enjoy the fruits of such learning processes and opted for a place in the rapidly emerging public schools. With large numbers of students now in classrooms, educators had to search for more efficient devices by which students might be processed or moved, if you will, to higher learning levels. Moving students in groups according to chronological age was appealing for its efficiency as was the use of the five symbol letter grading system, which finally emerged. The latter, generally A,B,C,D,E symbols, was a rather gross evaluative process which required little time and skill. It proved to be the path of least effort on the part of the teacher and gave the appearance of working well, for it worked on the average for about 36% of the students (see the section on reliability). In such a system students could be moved into higher levels of learning even though, as Morrison (104) so eloquently stated, they only half-learned,

or even barely learned a given body of content and skills. Some educators such as Eliot (36) and Morrison foresaw the pitfalls of these types of administratively efficient devices. Eliot had warned that they would move schools in the direction of mediocrity.

While some school systems did attempt to hold on to the concept of mastering learning such as the Dalton School System in Massachusetts, which instituted in 1919 a plan known as the Dalton Plan, the Winnetka School system in Illinois, whose superintendent, Washburne, developed the Winnetka Plan in 1922 and the University of Chicago Laboratory School where Morrison developed a plan in 1926, these plans did not survive. Block (13) believes that the reason for their demise was the lack of adequate technology to support systems that required an inordinate amount of dedication and effort.

With the development of modern technology, however, and Skinnerian theories on learning by conditioning, educators began to look anew at the prospect of reintroducing into the classroom the concept of mastering learning. The development of learning by conditioning had led to programmed instruction and these innovations plus the growth of computers prompted the introduction of two individualized instruction projects in 1960. These came to be seen as useful tools for attaining mastery of subject matter with little extra effort and time on the part of the teacher. One was in Pittsburgh and the other in Stanford. Both were individualized prescribed instruction projects and led the

way to the introduction of many more such projects in the 70's.

However, a model of school learning, published by Carroll in 1963 (25), seemed to give the most support to the return to requiring subject matter mastery in the classroom. It explicitly related via a clear cut mathematical formula the relationships of aptitude, quality of instruction and the ability to understand instruction, as well as the amount of time the student was willing to spend on the learning process and the total learning time he/she was allowed to spend on school learning. Carroll through this model had redefined aptitude to learn as a function of the amount of time required to learn a given task to a given criterion level under ideal instructional conditions. Such a definition had the effect of stating that the degree to which a student learned was a function of the ratio of time actually spent in learning to the time needed to learn the material to a given level of mastery. What Carroll, in essence, was saying was that many more students than is usual could master a given set of objectives, if given sufficient time and appropriate learning aids.

Bloom (13), in putting to practice Carroll's ideas, defined mastery in terms of a major set of course objectives that a student is expected to achieve at the completion of a unit of instruction. He perceived that a mastery stategy could be designed for use in the structured classroom by using feedback mechanisms such as diagnostic or formative tests, as defined by Scriven (13), and clearly articulated

instructional objectives, as well as alternative learning devices. Following Bloom's approach, several researchers introduced seemingly successful mastery strategies into classrooms of varying subject matter and chronological age levels. According to Block, the results of forty major studies indicated that approximately three-fourths of the students learning under such conditions achieved the same standard of mastery as one-fourth did under conventional group-based instruction.

The precise manner in which the mastery strategies were carried out varies, but their common base is the clarification of the educational objectives of the unit under study, pre-established mastery criteria, auxilliary instructional aids and formative evaluation procedures before any summative evaluation is made. The summative evaluative procedures used with these strategies also varied. Some instructors merely evaluated their students in the traditionanl normative pattern, using a single final exam, which they letter graded. Others, while using traditional letter grading, based it on the number of units mastered. Some even developed a system where both techniques were employed.

In looking at the studies that deal with the effectiveness of mastery strategies in the classroom, one finds that they fall into three categories. There are those that compare one year's class with another; those that attempt to be more rigorous by setting up a control class running concurrently with their experimental groups; and

those that not only utilize a concurrent control group, but attempt to match their students on some relevant variables. Criterion measures for mastery seem to be around 80-90% correct answers.

Studies representative of the first type were done by Arasian (4), Keller (74), Mayo, Hunt and Tremmel (93), Moore, Mahan and Ritts (100). Araisian (4) in 1967 applied a mastery strategy to a graduate level ten-week course in test theory consisting of thirty-three students. In the previous year's class only 30% of the students had received a grade of A. However, in the following year's mastery class 80% of the students received a grade of A on a parallel exam. Arasian, also, claimed that students used their time more efficiently under the aegis of a mastery strategy. In 1968 Keller (76) reported that in two courses in general psychology of two hundred students each that the introduction of a mastery strategy increased the percentage of A's and B's to 65-70%. The courses had been taught in successive years, and each time the strategy was applied the percentage of A's and B's increased, while the percentage of D's and F's decreased. Mayo, Hunt and Tremmel (93) in 1968, using a mastery strategy in a six-week University summer course in statistics, found that 65% of their mastery group of seventeen students as compared with 3% of their control group achieved a grade of A on the same final exam. Moore, Mahan and Ritts (100) reported in 1968 that in a mastery philosophy course given a year after its traditional predecessor, 4/5 of the experimental group received A's and

B's, compared to 3/5ths of the control group which received A's and B's on the same exam.

Studies that used concurrent control groups were done by Collins (31), Gentile (53) and Block (13). Collins studied freshmen learning mathematics in college. The study involved fifty liberal arts students in two modern algebra courses and approximately forty engineering and science students in two calculus courses. A mastery strategy was employed with one of the algebra and one of the calculus courses. Reporting in 1970, Collins (32) found that the percentage of students in the algebra mastery course that achieved A or B increased to 75%, as compared to 30% of those students in the traditional course. In the calculus classes 65% in the experimental group achieved mastery as compared to 40% for those who achieved the same level of mastery in the control group. The author claims that the introduction of a mastery strategy into the courses served to eliminate for all practical purposes D and F grades. Gentile (52) reported in 1970 on a study of a group of students in an introductory educational psychology course. While he used a somewhat different mastery strategy than some of the previous researchers mentioned, he found significant levels of increased understandings in the mastery course as compared to the levels reached in the traditional course for comparable material.

Block reported in 1970 (13) on a study that he did which indicated that individual differences at entry into a mastery program are not reflected in summative tests, as is

the case with a traditionally taught group. This finding was a spin off from a study that he did primarily for the purpose of determining the most effective mastery criterion score. What he had discovered was that a student's resources at entry into a control unit of no established unit mastery level played a large role in his/her final achievement, as well as in learning throughout the sequence. However, in those units where a pre-established mastery level was introduced, he found that resources played a decreasing role in the learning process. He had used ninety-one eighth graders each of whom were taught three sequential units of elementary matrix algebra. The students were randomly assigned to five groups, each learning their units to a different pre-established level of mastery. The 95% mastery level produced maximal cognitive learning but had a long run negative effect on attitudes and interest for students, while the 85% level of mastery produced maximal interest and attitudes and only slightly less effective cognitive learning. He also found that maintenance of a high level of mastery can make student learning more efficient.

Moore, Mahan, and Ritts (103) and Kim (79,80) used matched groups, thus adding more control to their designs than the previous researchers mentioned, nevertheless they came up with similar results. Moore, Mahan and Ritts (102) reported in 1968 that seventy students matched in terms of aptitudes and randomly assigned to either experimental or control groups revealed differences in mastery achievement.

The groups represented the subjects of biology and psychology. In both of these subjects the experimental group was found to be a standard deviation above the control group in the final exam.

Kim (77,78) did a series of studies, in Korea, in which he claims to have found that 74% of the experimental group compared to 40% of the control group achieved mastery, and that mastery learning was most effective for students with a below average I.Q. His sample consisted of 272 seventh graders paired in terms of I.Q. and mathematical achievement. Half of them were assigned to an experimental group, and half of them were assigned to a control group. Each group was taught a unit in simple geometric figures. The report was published in 1969. To attempt to verify the findings he set up another study a year later involving a much larger group of students, 5,800 seventh graders in mathematics and english courses. After eight weeks of learning, he found that while the results varied widely for the different schools in which the study was carried out, on the average, 72% of the students in experimental english groups reached the mastery criterion as compared with only 28% of those in the traditional groups and 61% of the students in the experimental mathematics groups achieved mastery as compared with 39% of those in the traditional groups. Kim perceived that the fluctuations he found were in large measure due to improper utilization of mastery techniques.

Researchers such as Collins (32), Swanson and Denton

(132), Decker (41), and Green (59) attempted to investigate which variables are most effective in mastery strategies. Collins (32) did a study with six mathematics classes from approximately twenty-five junior high schools. He reported in 1970 that specifying the instructional objectives and formative testing procedures is of great importance and go a long way toward improving student performance and that diagnostic problems and review perscriptions were so effective that the introduction of alternative learning resources seemed to be superfluous. Swanson and Denton's study in 1976 (132) with fifty-three eleventh and twelth grade chemistry students tends to support these findings. They found that remediation positively influences cognitive achievement and retention, and that alternative materials and activities that are teacher directed provide more optimum learning than mere repetition and review of materials under study.

Decker reported in his study of mastery strategies in 1976 that I.Q. was not a significant determinant of student performance and that instructional strategy can affect achievement. He had employed four different types of instructional strategies in an attempt to identify crucial variables related to mastery strategies. He found that students with unlimited testing opportunities performed better than those without such opportunities - in every case that he studied. Other variables that were significant were that students with two-week deadlines as compared to those with semester deadlines and student with advisor input, as

compared with those that had none, performed better. Green (59), in using a mastery strategy to introduce physics at M.I.T. in 1969, found that tutors worked better than technological gadgets.

Studies which indicate that mastery learning situations seem to improve students' attitudes and interests were done by Green (59), Gentile (52) and Biehler (9). Green in his study found that students in mastery classes seemed to enjoy their courses more than students in traditional courses. Gentile (52) found that on identical course evaluations, 74% of the mastery students compared to only 21% of the other students indicated thay enjoyed taking the course. The mastery students rated the course as one of the best that they have ever had. Biehler (9), in reporting in 1970 on a mastery strategy that he had utilized for teaching a course in introductory educational psychology, found that, when students were given the option to choose between the traditional type course or a mastery course, over 90% registered for the mastery course.

Two studies were found, however, that reported no statistically significant differences in affective outcomes between mastery stratagies and traditionally taught courses. One of them found no significant differences in achievement, either. One was done by Brolund and Smith and the other by Meyers. Both were reported in 1975. The Brolund and Smith study did find a small gain for the mastery students in achievement.

The indications of these studies tend to support: (1)

the Carroll postulate that learning is a function of time and remediation, (2) Bloom's prediction that a mastery strategy can be employed in the traditional classroom setting, (3) the propositions that students, if given a choice, prefer mastery strategies to traditionally taught courses, and 4) that student interest and attitudes toward learning seem to improve when the students are given conditions which offer them an opportunity to prove themselves capable of mastery of the subject matter to which they are exposed.

2.--Objectives

Early in the 1900's one finds concern for the lack of clear operational definitions of what a mark should represent. Starch (131) reflected on the "need for definite measures of education", and Ruch (117) stated that "factors considered in a marking system must be defined in detail". Educators like Pressey (117) began to see that the trouble vis-a-vis marks was not the test, but the lack of clearly defined goals.

According to Smith and Dobbins (132), the thirties brought increased concern for the problem with emphasis being placed on separating out achievement factors from non-achievement factors, and some insistence on the use of absolute measures in place of the relative measures in large use.

The imperative of introducing into classroom practice clearly defined objectives not only continued on into the

forties and fifties, but was given momentum by people such as Lamson (86), Tyler (127) and Bolmeier (17).

Lamson (86) emphasized in 1940 that "a mark should be a measure of the extent to which students have obtained the objectives of a course". Bolmeier (17) participating in the planning of a "progressive system of reporting" in 1951, held the line for developing objectives germaine to a given course. But it was to Tyler that most of the credit goes for the rapid interest in developing clear cut objectives. For it was he who spelled out the general procedures for the formulation of such objectives and insisted that such objectives needed to be classified into major types, if the objectives were to be useful for practical treatment.

This work of Tyler had a very crucial spin-off. It led to an informal meeting of college examiners in 1948. During this meeting interest was expressed in the development of a theoretical framework for improving communications among examiners through a system of clarifying goals for the educational process. This meeting turned out to be merely the first in what became a series of meetings toward such an end. The result was a brilliantly conceived Taxonomy of Educational Objectives of which Handbook I: Cognitive Domain was published in 1961 (15), and Handbook II: Affective Domain was published in 1964. These books plus Mager's "Instructional Objectives", which was published in 1962 (91), facilitated the use of clearly delineated goals for teaching and evaluative purposes, as did the work of Gagne in 1965 (51), who emphasized that to properly evaluate

instruction a domain of performance must first be defined, Popham in 1969 (113) who advocated increased use of measurable objectives, and Gronlund in 1970 (59), who defined evaluation as the systematic process of determining the extent to which educational objectives are achieved by pupils.

Perceiving that many educators would prefer to be selectors rather than generators of instructional objectives, an Objectives Exchange was established in California in 1968. Those who founded the exchange also had hope of it becoming a national depository and development agency not only for instructional objectives, but also for related measurement devices.

But in spite of these developments the use of instructional objectives in classroom practice met with a great deal of resistance. Isaac and Michaels (69) sum up the objections to behavioral objectives in their Handbook in Research and Evaluation. Much of the resistance seems to be related to the fact that the kinds of classroom objectives being used were too specific and thus painstaking and tedious to deal with, as well as the fact that such specific objectives were perceived as tending to force teaching into an inflexible mold.

Thus while performance objectives were recognized as crucial for improving the validity and reliability of evaluative systems, they did not take hold because of their cumbersome nature and their restrictions on content. Perhaps the problem lies in the fact that the admonishments

of Baker and Tyler had not been given the kind of serious consideration they deserved. Baker (6) emphasized the need for the use of objectives that demonstrate a "generality" that can cut across classrooms; Tyler (131) understood the importance of classifying objectives in such a manner as to make them practical. The content objectives that have been in general use have neither of these characteristics.

Objectives having such characteristics are, however, available. They are found in the taxonomic scheme developed by Bloom et al (14). This is a classification scheme that can be used in a practical manner, and that has the kind of generality needed to cut across classrooms, leaving open the matter of content. The major objectives involved are only six in number, making them administratively manageable. Their authors claim that they can cut across content areas, making them generalizable and thus usable in any classroom. This generality gives them the potential for being used as standardized measures of educational outcomes. It is claimed that these objectives represent a heirarchy of learning skills and behaviors, making them discriminatory as well.

Another aspect of these objectives that we may have been ignoring is the implied wisdom of these cognitive objectives. The message that emanates from them is that content can be learned and utilized at various levels of cognitive achievement, and that it is equally important to evaluate the cognitive level at which a person is functioning as well as to determine the actual content of

knowledge that he/she has accumulated. The latter skill, according to Bloom et al, represents only the first and most basic of the major cognitive skills and abilities that man can acquire.

If indeed these six major cognitive objectives have the kind of construct validity accredited to them, then the content of any course could be measured at the varying cognitive levels. This can be accomplished by criterion-referencing tests to these specific levels. This is possible simply by framing a question related to a particular body of content in such a manner as to elicit one of the six major cognitive responses (see Appendix C). If these objectives are also hierachical in nature, it would then also be possible to discriminate varying cognitive skills and abilities. Such a discriminatory device could be a very useful for improving both educational and guidance processes.

The work of Kropp, Stoker and Banshaw (86) lends support to the claims of a hierarchical nature and "generality" for the cognitive objectives. In 1966 they reported that on investigating the nature of the heirarchical structure of the cognitive objectives developed by Bloom et al (14), that their study findings, based on a decreased mean score with increasing complexity of cognitive objectives, supported the heirarchical hypothesis, and that a simplex analysis of an intercorrelation matrix, while less definitive than the first analysis, offered some support. The imputed generality of the cognitive processes was also

investigated through the use of circumplex and factor analysis. Their results were unclear, but the findings tended to support such an hypothesis. They believed that complex interactions of content and process contributed to the lack of clarity.

Along with the development of instructional objectives interest in criterion-referenced testing and absolute measures increased. For with the use of specified performance standards it now became incumbent on those working with such to criterion-reference their tests to their objectives, and to establish an absolute performance standard against which a pupil's learning could be evaluated (13). The introduction of mastery learning strategies and individualized instructional packages and contract learning techniques spurred such development. (Brennan (21) points out that most criterion-referenced testing is closely associated with some kind of instruction.) Increased interest in the use of absolute measures was also the result of the perception by many that such measures are more appropriate for classroom evaluation procedures in that with their use one can evaluate individuals according to their own patterns of learning and measure the extent of their progress.

The accumulated evidence, thus, both inductive and deductive, suggests that specific objectives, accompanied by an absolute measure, have the most potential for improving the validity and reliability of the evaluative process vis-a-vis learning outcomes; that with such objectives

criterion-referenced tests are needed; and that the six major cognitive objectives developed by Bloom et al are the kinds of objectives that hold out the promise for an evaluative process that can replace letter grading.

REFERENCES

1. Adams, W. "Why Teachers Say They Fail Pupils." Ed Adm Sup, 18:594-600;1932

2. Adams, G.S., and Torgerson, T.L. Measurement and Evaluation, Psychology and Guidance N.Y.: Holt, Rinehart and Winston, Inc., 1964.

3. Aiken, L.R. "The Grading Behavior of a College Faculty." Ed Psychol Meas 23:319-322;1963.

4. Arasian, P.W. "An Application of a Modified Version of John Carroll's Model of School learning." Unpublished Master's thesis, Univ. of Chic., 1967.

5. Ayer, F.C. "School Marks." Rev Ed Res 3:210-204;1933.

6. Baker, E. "Defining Content of Objectives." (Los Angeles Vinicit Assoc. P.O. box 24714, 1968.)

7. Barton, Jr. W. A. "Pupil Reaction to School Reports." Sch Rev 33:771-780;1925.

8. Bass, B.M. "Intrauniversity Variation in Grading Practices." J Ed Psychol 21:48-52;1930.

9. Biehler, R.F. "A First Attempt at a 'Mastery Learning' Approach" Ed Psychol 7:7-9;1970.

10. Billet, R. Provisions for Individual Differences, Marking, and Promotion. U.S. Office of Ed Bull 1932, No. 17. Nat Surv Sec Ed, Monograph No. 13, GPO 1933.

11. Birney, R.C. "The Effects of Grades on Students in College Grading Systems." J of Higher Ed 35:96-98;1965.

12. Bixler, H.H. "School Marks." Rev Ed Res 6:169-73, 247-48;1936.

13. Block, J. H. ed. Mastery Learning, N.Y.: Holt, Rinehart and Winston, Inc. 1971.

14. Bloom B. S. ed. Taxomomy of Educational Objectives: The Classification of Educational Goals, Handbook I, Cognitive Domain. N.Y. David McKay Co., Inc. 1956. 207 p.

15. Bloom, B.S., Hastings, T.J., and Madaus, G.F. Handbook on the Formative and Summative Evaluation of Student Learning. 1971. N.Y.: McGraw-Hill, Inc. 905 p.

16. Bolmeier, E.C., "What's in a Mark?" Sch Ex 62:25;1943.

17. Bolmeier, E.C. "Principles Pertaining to Marking and Reporting Pupil Progress." Sch Rev, 59:15-24;1951.

18. Bolton, F.E. "Do Teachers Marks Vary as Much as Supposed?" Ed 48:23-39;1927.

19. Bostrum, R.N., Vlandis, J.W., and Rosenbaum, M.E. "Grades as Reinforcing Contingencies and Attitude Change." J Ed Pschol 52:112-115;1961.

20. Bowers W. Student Dishonesty and Its Control in College N.Y.: Bur of Appl Behav Sci, 1964.

21. Brennan, R. " The Evaluation of Mastery Test Items." Natnl Cen Ed Res and Dev. Wash, D.C. 1974 255p. Eric Doc ED 092 593.

22. Brolund, N.S., and Marshall, P. "Cognitive and Affective Outcomes of PSI Mastery Programs as Compared to Traditional Instruction." Paper presented at annual meeting of the Amer Ed Res Assn (Wash., D.C.1975.) Eric ED 108 985.

23. Brookover, W. B., Shailer, T., and Paterson, A. "Self-concept of Ability and School Achievement." Soc of Ed 37:271-78;1964.

24. Burke, R. "Student Reaction to Course Grades." J Exp Ed 37:13-16;1969.

25. Carroll, J.B., "Model of School Learning." T.C. Rec 64:723-733;1963.

26. Carter, R.E. "How Invalid Are Marks Assigned by Teachers?" J Ed Psychol 43:218-28;1952.

27. Cattell, J. McK. "Examination, Grades and Credits." Pol Sci Mon 66:367-378;1905.

28. Chansky, N.M. " A Note on the Grade Point Average in Research ," Ed Pschol Meas 24:95-99;1964.,

29. Chapman, I.C. and Hills, M.E.. "A Statistical Study of the Distribution of College Grades." Pedagogical Seminary 23:204-210;1916.

30. Child, I.L., and Whiting. "Determinants of Level of Aspiration: Evidence From Everday Life." J of Abnorm Psychol 44:303-14;1949.

31. Cocking, W.D. and Holy, T.C. "Relation of Intelligence Scores to High-School and University Marks." Ed Res Bull 6:383-384;1927.

32. Collins, K.M., 1970 "A Strategy for Mastery Learning in Modern Mathematics." Unpublished study, Purdue Univ. Division of Math Sciences.

33. Crawford, A.B. "Rubber Micrometers." Pop Sci Mon, 66:367-378;1905.

34. Crew, H. "President's Address - Annual Meeting." Am Ass'n Univ Prof Bull 16:103-11;1930.

35. Crooks, A.D. "Marks and Marking Systems, A Digest." J of Ed Res 27:259-72;1933.

36. Cubberley, E., The History of Education. MA: Houghton Mifflin Co., 1922, 607-708 p.

37. Cureton, L.W. "The History on Grading Practices." Meas in Ed, Natl Con on Meas in Ed, 1971.

38. Dearborn, W.F. " Relative Standing of Pupils in High School and in the University." Univ of Wis Bull no. 312.

39. Decker, D.F. " Teaching to Achieve Mastery by Using Retesting Techniques." 1976, Eric Doc ED 133 002

40. Depencier, I.B. "Trends in Reporting Pupil Progress in Elementary Grades." El Sch J 51:519-528;1951.

41. Dexter, E.S. "The Effect of Fatigue or Boredom on Teacher's Marks." J of Ed Res 28:664-667;1935.

42. Douglass, H. R. and Olsen, N.E. "The Relation of High School Marks to Sex." Sch Rev, 45:283-288;1937.

43. Edgeworth, F.Y. "The Element of Chance in Examinations." J Royal Stat Soc 1890, p. 460-75 and 644-73.

44. Edmiston, R.W. "Do Teachers Show Partiality?" Peabody J Ed 20:234-238;1943.

45. Eells, W.C. "Five-Point Grading Systems." J Ed Psychol, 21:128-135;1930.

46. Fala. M. A. Dunce Cages, Hickory Sticks, ard Public

Evaluations: The Structure of Academic Authoritarianism. The Teach Asst Assoc, Univ. of Wis, (1968), 11-12.

47. Feather, N.T. "Effects of Prior Success and Failure on Expectations of Success and Subsequent Performance." J Person Soc Psychol, 3:287-98;1966

48. Finkelstein, I.E. A Marking System in Theory and Practice. Baltimore: Warwick and York, Inc., 1913.

49. Forman, W.O. "The Gradeless Era in High School." J Ed 111:501-502;1923.

50. Fraser, M.G. The College of The Future. N.Y.: Col. Univ. Press, 1937.

51. Gagne, R.M., Essentials of Learning for Instruction. Ill: Dryden Press. 1974, 164 p.

52. Gentile, J.R. "A Mastery Strategy for Introductory Educational Psychology." Unpublished Materials, State Univ. of N.Y. at Buffalo, Dept. of Ed Psych, 1970.

53. Gilkey, R. "The Relation of Success in Certain Subjects in High School to Success in the Same Subjects in College." Sch Rev 37:576-588;1929.

54. Gilman, D." Alternatives to Tests, Marks and Class Ranks. " Curr Dev Cen, Sch of Ed, Ind. State Univ. May 1974 Eric Doc ED 095 210.

55. Glaser, "Instructional Technology and Measurement of Learning Outcomes." Amer Psychol Aug 519-21, 1963.

56. Goodlad, J.I. and Anderson, R.H. the Non-Graded Elementary School Rev. Ed., Harcourt, 1963, 248 p.

57. Gould, G. "Practices in Marking and Examination." Sch Rev 11:142-146;1932.

58. Gray, C.T. Variations in Grades of High School Pupils. Warwick and York. 1913. p.120.

59. Green, B.A. Jr., A Self-Paced Course in Freshman Physics. Cambridge, MA: M.I.T., Ed Res Cen, 1969.

60. Gronlund, N.E. Improving Marking and Reporting in Classroom Instruction. N.Y.: Macmillan Publ. Co., 1974, 57 p.

61. Haagen, C.H. "The Origins of a Grade, in College Grading Systems: A Journal Symposium," J of Higher Ed 35:89-91;1964.

62. Heilman, J.D. "The Reliability of College Teacher's Classroom Tests." Ed Adm Sup 117:535-543;1931.

63. Heist, P. "Personality Growth in the College Years." College Board Review 56:25-32;1965.

64. Hills, J.R. Klock, J.. and Bush, M. "The Use of Academic Prediction Equations with Subsequent Classes." Amer Ed Res J. 2:203-206;1965.

65. Hills, J.R. "Predictions of College Grades for all Public Colleges of a State." J of Ed Meas. 1:155-9;1964.

66. Huddleston, E.M. "Measurement of Writing Ability at the College Entrance Level: Objective vs Subjective Testing Techniques." J Exp Ed 22:165-207;1954.

67. Hughes, W.H. "Analyzing the Ingredients of Teacher's Marks." Nation's Sch 6:21-26;1930.

68. Hulten, C.E. "The Personal Element in Teachers Marks." J Ed Res 12:49-55;1925.

69. Isaac, S., and Michael, W.B. Handbook in Research and Evaluation. CA: Knapp, 1974. 186 p.

70. Jacoby, H. "The Marking System in the Astronomical Course at Columbia College." Sci 31:819;1909.

71. Johnson, F.W. The Administration and Supervision of a High School. Ginn, 1925, 402 p.

72. Johnson, D.W. and Johnson, R.T. "Instructional Goal Structure: Cooperative, Competitive, or Individualistic." Rev Ed Res 44:213-240;1974.

73. Jordan, D.S. The Trend of the American University. CA: Stanford Univ. Press, 1929.

74. Keller, F.B. "Goodbye, Teacher." J Appl Behav Anal, 1:78-79;1968.

75. Kelley, E.G. "A Study of the Consistent Discrepancies Between Instructor Grades and Term-end Examination Grades." J Ed Psychol 49:328-34;1958.

76. Kelly, J.K. Teachers Marks. N.Y.:Columbia Univ., 1914.

77. Kim, H., et al, A Study of the Bloom Strategies for Mastery Learning. Seoul: Korean Institute for Research in the Behavioral Sciences. (In Korean), 1969.

78. ---------- The Mastery Learning Project in the Middles Schools. Seoul: Korean Institute for Research in the Behavioral Sciences. (In Korean), 1970.

79. Kirby B.C., "Three Error Sources in College Grades.", Journ of Exp Ed 31:212-218;1962.

80. Kirschenbaum, H., Napier, R., and Simon, S.S. Wad-ja-get? N.Y.: Hart Publ. Co., 1971.

81. Klugh, H.E., Bierley, R. "The School and College Ability Test and High School Grades as Predictors of College Achievement." Am Ed Res J 1965, 22 p.

82. Knowlton, J.Q., Hamerlynck, L.A. "Perception of Deviant Behavior: A Study of Cheating." J Ed Psychol 379-385;1967.

83. Krathwohl, D.R., Bloom, B. and Masau B. A Taxonomy of Educational Objectives: The Classification of Educational Goals, Handbook II. Affective Domain, N.Y.: McKay Co. Inc., 1964.

84. Kropp, R.P., Stoker, H.W., and Banshaw, W.L. "The Validity of the Taxomony of Educational Objectives." J Exp Ed 34:69-76; (Spring 1966).

85. Kurtz, J.J., and Swenson, E.J. "Factors Related to Over-Achievement in School." Sch Rev 59:472-480;1951.

86. Lamson, E.E. "The Problem of Adequate Evaluation of the College Student's Achievement ." Ed Adm Sup 26:493-507;1940.

87. Lauterbach, C.E. "Some Factors Affecting Teacher's Marks J Ed Pschol 19:266-271;1928.

88. Lawson, D.E. "Teachers' Marks - Tragic and Absurd." Ed For 4:175-179;1940.

89. Lentz, T.J. "Sex Differences in School Marks with Achievement Scores Constant." Sch Soc 29:65-68;1929.

90. Lindquist, E.F. "An Evaluation of a Technique for Scaling High School Grades to Improve Predication of College Success." Ed Psychol Meas 23:623-645;1963.

91. Mager, R.F. Preparing Instructional Objectives. 2nd ed. CA: Fearon, 1975.

92. Maney, C.A. "Sex Bias in College Marking." J High Ed, 4:29-31;1933.

93. Mayo, S. T., Hunt, R. C., and Tremmel, F., 1968. "A Mastery Approach to The Evaluation of Learning Statistics.", Paper presented at the annual meeting of the Natn'l Counc. on Meas. in Ed, Chicago, Ill.

94. Miles, W.R. Comparison of Elementary and High School Grades. Studies in Ed, 1. no. 1.

95. Miller, S. Measure Number and Weight: A Polemical Statement of the College Grading Problem. Cen. of Res. on Learn. and Teach., The Univ. of Mich., Ann Arbor, March 1967.

96. Milton O., Edgerly, J.W. The Testing and Grading of Students. Change Magazine and Educational Change, 1976.

97. Meyer, M "The Grading of Students." Sc 28:243-252.

98. ---------- "Experiences with the Grading System of the University of Missouri." Sci 33:661-67; April 28, 1911.

99. Modu, C.C. "Affective Consequences of Cognitive Changes." Unpublished Ph.D. dissertation, Univ. of Chicago, 1969

100. Moore, J.W., Mahan, J.M., Ritts, C.A., 1968. "An Evaluation of the Continuous Concept of Instruction with University Students." Paper presented at the annual meeting of the Amer. Ed Res. Ass., Chicago, Ill.

101. Morrison, H.C. "Studies in High School Procedure." Sch Rev 29:106-118;1921.

102. Moslemi, M.H. "The Grading of Creative Writing Essays." Res Teach Eng 9:154-161;1975.

103. Mouseley, W. "Report Cards Across the Nation." Phi Delta Kappan, 53:436-437;1972.

104. Myers, R.R. "The Effects of Mastery and Aptitude on Achievement and Attitude in an Introductory College Geography Course." Dissertation, Univ. of Georgia, 1975. 195 p. ED 120 035.

105. National Education Association. " Marking and

Reporting Pupil Progress." Res Sum. 1970, S-1, Wash., D.C. NEA Res Div 1970.

106. National School Public Relations Association. Grading and Reporting. 1972.

107. Norsted, R.A. "To Mark or Not to Mark?", J Ed, 121:8-84;1938.

108. Odell, C.W. Educational Measurement in High School. Century, 1930, 641 p.

109. Ohlson, D. "School Marks vs. Intelligence Rating." Ed Adm Sup 13:90-102;1927.

110. Page, E.B. "Teacher Comments and Student Performance." J Ed Psychol 49:173-181;1958.

111. Penfold, D.M.E. "Symposium: The Use of Essays in Selection at 11+." Brit J Ed Psychol 26;128-136;1956.

112. Petit, W.W. "A Comparative Study of New York High School and Columbia College Grades." Master's Essay, T.C., 1912.

113. Popham, J.W. "The Uses of Instructional Objectives." Fearon Pub/Lear Siagler, Inc. Calif. 1970. 135 p.

114. Pressey, S.L. "Fundamental Misconceptions Involved in Current Marking Systems." Sch Soc 21:736-738;1925.

115. Roberts, A. "A Study of the Marking System of Teachers of the Everett(Washington) High School." Ed Adm Sup 3:485-497;1917.

116. Rocchio, P.D., Kearney, N.C. "Teacher and Pupil Attitudes as Related to Non-promotion, Secondary School Pupils." Ed Psychol Meas 10:244-252;1956.

117. Ruch, G.M. The Objective on New-Type Examinations. Chicago:, Scott, Foresman and Co. 1929, 478 p.

118. Rugg, H.O. "Teacher Marks and Marking Systems." Ed Ad Sup 2:117-142;1915.

119. Ruggles, A.M. Grades and Grading. Teacher's College, Masters Essay. 1911.

120. Segel D. "Predicting College Success." U.S. Office of Ed Bull 1934 no. 15. GPO. 1938, 48 p.

121. Shaw, M.C., McCuen, J.T. "The Onset of Academic

Under-Achievement in Bright Children." J Ed Psychol 51:103-108;1960.

122. Shepherd, D.E.M. "The Effect of the Quality of Penmanship on Grades." J Ed Res 19:102-105;1929.

123. Shinnerer, M. C. "Failure Ratio: 2 Boys to 1 Girl." Clear House 18:264-270;1944.

124. Skinner, A.F. "The Science of Learning and the Art of Teaching." Harv Ed Rev, 24:86-97;1954.

125. Smith, F.O. "A Rational Basis for Determining Fitness for College Entrance." Univ of Iowa, Studies in Ed., vol 1. no. 3. 1910.

126. Smith, A.Z. and Dobbin, J.E. "Marks and Marking Systems." The Encylopedia of Educational Research Harris, C.W. and Maris, R.L. (ed.), N.Y.: Macmillan, 3rd ed. 783-791, 1960.

127. Smith, E.R. et al. Appraising and Recording Student Progress. Harper, 1942. 550p.

128. Starch, D., and Elliott, E.C. "The Reliability of Grading High School Work in English." Sch Rev 21:442-457;1913.

129. _____"The Reliability of Grading Work in Mathematics.." Sch Rev 21:254-259;1913.

130. _____ "The Reliability of Grading Work in History." Sch Rev 21:676-681;1913.

131. Starch, D. Educational Measurements. N.Y.: The Macmillan Co., 1918, 202 p.

132. Swanson, D.H., and Denton, J.J. A Comparison of Remediation Systems Affecting Achievement and Retention in Mastery Learning, 1976. Eric Document, ED 131 037.

133. Swenson, C. "The Girls are Teacher's Pets." Clear House 17:537-540;1943.

134. Taylor H. R., and Constance, C. L. " How Reliable Are College Marks?" In Res and Higher Ed Bull. 1931. No. 12. GPO, 5-14, 1913.

135. Temple University, Report of the College Education Ad Hoc Committee on Grading Systems, 41-48, 1968.

136. Thompson W.N. "A Study of Grading Practices of 31

Instructors in Freshman English." J of Ed Meas, 49:65-68;1955.

137. Thorndike, E.L. An Introduction to the Theory of Mental and Social Measurement. (Revised) N.Y.; Teachers College. 1913.

138. Thorndike, R.L. "Marks and Marking Systems." In Encyclop Ed Res ed by R.L. Ebel. 4th ed., N.Y.: The Macmillan Co., 1969.

139. Tieg, E.W. Tests and Measurements for Teachers, Houghton, 1931, 470p.

140. Torshen, K.P., 1968. "The Relation of Classroom Evaluation to Students' Self-concepts." Unpublished manuscript, Univ. of Chicago, Dept. of Ed

141. ---------- "The Relationship of Evaluations of Students' Cognitive Performance to Their Self Concept Assessments and Mental Health Status." !973, Eric Doc Ed 074 424.

142. Trabue, M.R. Measuring Results in Education. N.Y.: American Book Co., 1924.

143. Travers, R.M., and Gronlund, N.E. "The Meaning of Marks." J Higher Ed 21:369-374;1950.

144. Twerlinger, J.S. Assigning Grades To Students. Ill.: Scott Foresman, 1971.

145. University of California @ Berk., Report on Methods of Evaluating Students at the Univ. of Calif. - Berk., Oct., 1965, p.13.

146. Vredroe, L.C., Lindecamp, C.D. "How Shall We Make Recording and Reporting of Pupil Progress More Meaningful?" Bull Natl Assoc Sec Sch 37:179-185;1953.

147. Weiner, B. "The Effects of Unsatisfied Achievement Motivation on Persistance and Subsequent Performance." J. Pers 33:4428-42;1965

148. Wood, B.D. "The Need for Comparable Measurements in Individualizing Education." Ed Rec Sup No. 12, 20:14-31;1939.

149. Wrinkle, L. "The Story of an Experiment in Marking and Reporting." Ed Adm and Sup., 23:481-500;1937.

150. ---------- Wrinkle, L. Improving Marking and

Reporting Practices N.Y. Rinehart and Co., Inc., 1947, 120 p.

CHAPTER III

METHODS AND PROCEDURES

Data-Collection-Procedures

The needs of the study required:

1. a random sample of students studying the same content at the same level of difficulty;
2. tests criterion-referenced to cognitive objectives;
3. a mastery strategy;
4. a criterion score of 80-85% correct answers;
5. a mastery cognitive profile; and
6. a standardized chemistry test.

-1--Sample

The use of a random sample of students studying the same content at the same level of difficulty would allow for generalizations about the population from which the sample was drawn. However, a compelling need of the study was to test the feasibility of the selected educational devices in the classroom. This involved not only the use of a mastery strategy, but also an attempt to validate the generality and taxonomic nature of the cognitive constructs herein

suggested. The mechanics of obtaining a random sample on which this strategy could be employed with a high degree of control seemed beyond the province of this author. Thus the author's own General Chemistry classes were chosen for the feasibility study.

The sample consisted of two classes of students in the eleventh and twelfth grades who had elected to take chemistry for many diverse reasons. The group contained twenty-four students, eleven males and thirteen females, representing a wide variance in achievement. Their PSAT scores ranged from the 3rd percentile to the 71st percentile of all the students in the United States who had taken that test. (In large measure these are students who are pursuing a college career.) The mean of the sample's score was at the 24th percentile, while the median was the 21.5 percentile, indicating that the scores were positively skewed. Compared to a normally distributed population of students, this sample represented the lower end of the achievement scale. However, the members were not chosen in any systematically biased manner. Thus while not a normally distributed or randomly selected sample, the lack of systematic bias would lead one to consider that any generalization suggested by statistical analyses of the sample might be accepted as being highly probable. The fact that the sample has a wide range achievementwise gives it a representativeness crucial for differentiating the cognitive constructs and for studying the effects of a mastery stategy in maximizing achievement. (Isaac and Michael support such reliance on

controls by individual differences and less reliance on random sampling and formal statistical controls for seeking empirically established principles and methods in relation to people of particular types.)[1]

The sample size, while relatively small, was large enough to test null hypotheses. With a small sample the researcher can stay close to the data, an important factor in studying feasibility, and one can eliminate concern for statistical signficance due merely to large sample size.

The classrooms were traditionally structured, meeting one period (45 minutes) per day every day. Learning units had to be divided up into approximately six-week periods to coincide with card-marking. Card-marking was performed in traditional fashion with traditional letter grading.

-2.--Tests-

There were no known tests in Chemistry that were criterion-referenced to cognitive objectives that could be used with the mastery strategy envisioned. Tests, therefore, were developed during the summer of 1975 in the following manner:

1. the text (Chemistry by Parry et al.) was divided into six units of content to coincide with the six-week marking periods.
2. ten chapters were chosen as representing the minimum material needed for a basic understanding of high school chemistry and

[1]Stephen Isaac and William B. Michael, op. cit., P.68

labeled Level I (this included material covered the first semester).

3. eight more chapters were determined to contain material more complex than Level I, but necessary to cope with a college level course, and labeled Level II (this included material covered the second semester).[1]

Chem study tests, regent tests, and text-book developed tests were each analyzed and test items chosen that represented the above content levels. The questions from these standardized tests were then categorized into one of the six major cognitive objectives: Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation. While one finds frequently some overlap of objectives in many questions, the most predominant objective that the question tapped became the determining factor in criterion-referencing the test item. (Bloom's text was found to be very useful in training one to criterion-reference questions to the cognitive objectives.)[2] Through this process it soon became evident that most of the test items in general use for a General Chemistry course were tapping the first three major objectives in the Bloom Taxonomy. Confirmation that these first three categories were sufficient to differentially place most of the

---

[1]Robert W. Parry et al Chemistry (New Jersey, Prentice-Hall, 1970).

[2]Benjamin S. Bloom, Thomas J. Hastings, and George F. Madeus, Handbook on Formative and Summative Evaluation of Student Learning (New York: McGraw-Hill, 1971).

questions was found in the standardized Anderson-Fisk Chemistry Test.[1] The developers of this test categorized their test items into only these three objectives in their test booklet.

Arbitrarily, it was determined that high achieving students ought to be able to master twelve twenty-question tests per marking period (six weeks) or per unit of content. Aware of the need for all students to experience some achievement and aware of the fact that most of the students in the Chemistry class tested at the lower ends of the achievement scale in the PSAT tests, the objective Knowledge was further subdivided into Knowledge of Terminology and Knowledge of Specific Facts, etc. The expectation was that mastery of at least one objective would become more probable for the large majority of the students by such a subdivision.

Four twenty-question tests were then devised for each of the four objectives. With the exception of the tests for the objective, Knowledge of Terminology, each test consisted of four pages with five questions to a page in a kind of random order. This arrangement made it possible to administer different tests at the same time to any group of four students and to make up different combinations of tests for other students. Repeated test taking was to be part of the mastery strategy, as well as an option for attempting to

---

[1]Kenneth E. Anderson, Franklin G. Fisk, Anderson-Fisk Chemistry Test, Form F (New York: Harcourt, Brace & World, Inc., 1966).

establish high reliability. With no adequate theory to test for reliability of criterion-referenced tests, and supported by the Isaac and Michael statement that replication provides empirically stronger evidence than a single occurrence,[1] repeat tests became the option of choice. With a universe of eighty questions, five questions per page and a total of sixteen pages to use in repeated test-taking, it was possible to manipulate the pages so as to reduce opportunities for cheating and for sheer memorization of questions. Since the content of tests criterion-referenced to cognitive objectives play only an ancillary role, it mattered not that each item of content have an equal chance of being tested, or that the exact same content be tested with each administration of a test.

The tests were constructed in the following manner:

1. tests for the objective, Knowledge of Terminology, were simply two-page tests. The questions were fill-ins, following the general order of the text.
2. tests for the following objectives, Knowledge of Specific Facts, etc., and Comprehension were largely multiple choice, following no specific text order.
3. tests for the objective, Application, were largely of the problem type, but had the added requirement that all the work needed to solve

---

[1]Stephen Isaac and William B. Michael, op. cit., p. 144.

the problems had to be clearly organized on the test paper.

4. the tests related to the last two objectives, Comprehension and Application, were conceived as open book tests, for their test items were designed to evaluate skills and abilities, not knowledge, per se.

3. Mastery Strategy

According to mastery proponents, achievement is a function of time and facilitative review. Working with these two concepts a mastery strategy was developed which consisted of:

1. allowing students to take tests in the same unit of content for the same objective and as often as the students were willing to demonstrate the required mastery within the time-confined limits of a card-marking period; and
2. using different instructional techniques for review of subject matter.[1] This consisted of utilizing program texts, films, and relevant laboratory experiences -- the introduction of such techniques being consistent with research evidence, particularly the work of Gagne.

---

[1] R.M. Gagne and N.E. Paradiso, "Abilities and Learning Sets in Knowledge Acquisition," in Mastery Learning op. cit., p. 116.

The mastery strategy was fielded in the school year 1975-76, for the purpose of ironing out any difficulties that might arise in the traditionally structured classroom. The students that year were allowed a great deal of flexibility in test taking. They were told that they could take any test when and if they felt ready for it. Given such flexibility, many of the students put off taking tests until it was too late to demonstrate mastery in a given unit of content for any given objective within the card-marking period. As a result, the year that the study was undertaken, 1976-77, the students were required to take tests on assigned test dates and attempts were made to get the students to take as many tests as possible within the limits of the time confined setting. They were also told that if they wanted to pass the course, the first objective, Knowledge of Terminology, had to be mastered in every unit of content. However, in keeping with the mastery strategy they were allowed to take repeat tests at conveniently arranged times, and as often as they were needed to demonstrate mastery. Letter grades were to be based on the number of objectives mastered.

-4.--Mastery-Criterion Score

Evaluation with a mastery strategy can be accomplished with the foregoing techniques:

1. a summative test, normatively-scored;
2. a summative test, criterion scored; or
3. demonstrated mastery, based on a criterion

score.

Demonstrated mastery criterion-scored seemed to best meet the requirements of the theoretical framework for an improved marking and reporting system. Research by Block suggests that an 85% correct answer score maximizes attitudes and interest of students, while reducing cognitive learning only slightly.[1] The preliminary experiences of the school year 1975-1976 led, however, to an 80% correct answer criterion score, for this was found to be more administratively feasible in the time confined classroom. Mastery was then arbitrarily determined as reaching the 80% correct answer criterion score in three tests for any given content area in any given objective.

5. Mastery Cognitive Profile

This was simply generated from the kinds of objectives that were expected to be mastered in any given area and level of content. It can be a simple chart on which date of mastery need be the only recorded item. While recording date of mastery may be challenged by some as having the effect of stratifying learners in much the same manner as do normative scores, the author felt that dates should be included, but used only when needed to differentiate fast from slow learners. Such a profile has the advantage over normative scoring of leaving open the possiblity and hope of demonstrating mastery at some later date. This type of

---

[1]James H. Block, "The Effects of Various Levels of Performance on Selected Cognitive, Affective, and Time Variables," in Mastery Learning, ibid., pp. 104-106.

reporting system should not result in the type of labeling found in normative systems, for it is a record of accomplishment, not failure. Such a profile should also be simple to administrate, file and reproduce.

6. Standardized Chemistry Test

To determine the validity of the claim that mastery strategies tend to maximize achievement, a standardized chemistry test had to be found having items criterion-referenced to the same cognitive objectives under study in this thesis. A search through the The Mental Measurements Yearbook led to consideration of the Anderson-Fisk Chemistry Test.[1] It seemed to be one of the better standardized tests for it:

1. was recommended by Crawford as a well constructed test;
2. made claims to a high construct validity although the publishers provided no identifying sources;
3. made claims to high reliability which it did support with item analyses, standardization procedures and split-half reliability measures; and
4. provided standard error of measurement figures.

Most importantly it provided test items that were

---

[1]Oscar K.Buros, The Mental Measurements Yearbook (New Jersey: Gryphon Press).

categorized into the three major cognitive objectives crucially needed for comparison purposes.

Procedures for Treating the Data

In an attempt to establish construct validity of the test items, the 640 questions that made up the tests related to learning Level I were assigned numbers, in the order in which they appeared. Twenty-six numbers were then selected from a table of random numbers and test items with these numbers were compiled into one test. This test was mailed along with an item analysis sheet and an accompanying letter to ten independent judges. The judges were selected arbitrarily from amongst chemistry teachers in Oakland County school systems. The item analyis contained brief descriptions of the cognitive categories and the judges were asked to criterion-reference the twenty-six test items according to the four stated objectives, anonymously. Four of the ten judges responded. There was no attempt made to determine why the other judges did not respond, as the respondents were unknown (see Appendix B).

A table was set up containing scores which reflected the percent agreement of the independent judges with the author's choice of objective to which the test item had been criterion-referenced. An analysis of variance revealed that the null hypothesis that there are no significant differences among the judges with the author between test items or within categories at the 0.01 level of confidence could not be rejected. Thus by indirection the findings

supported some confidence in the construct validity of the tests designed for this study. The 0.01 level of confidence was deliberately chosen so as to avoid the possibility of discarding a promising lead.

To test the validity of the claims of mastery strategists that mastery strategies tend to maximize achievment, a one-shot standardized chemistry test whose items were criterion referenced to the cognitive objectives was needed for comparison purposes. The Anderson-Fisk Chemistry Test met the requirements of the study. While this test is designed to be used in traditional fashion in which content alone is tested, the test items had been categorized into the first three major cognitive objectives. Thus it was possible to analyze these test items in terms of the major cognitive objectives and in terms of content units and learning levels, similar to those used in the teacher developed mastery learning tests.

The average percent of errors were compared for similar material on both the standardized and mastery tests through the use of a correlated t-test, based on the conclusions of Gardner.[1] He states that while the assumptions for a t-test are linear relationships, normal population distributions and equal variances, t-tests are very robust and maintain their logicality even if the assumptions are violated.

The general validity claimed for the cognitive objectives based on findings that they cut across content

---

[1]Paul L. Gardner, "Scales and Statistics," Review of Educational Research 45 (Winter 1975), pp. 43-57.

areas and can reveal deviant patterns of learning, were tested in the first instance by a Spearman rho correlation and by an analysis of variance in the second instance. Rank orders and not scores were analyzed since they are independent of difficulty levels.

The hypothesis that the cognitive objectives represent a hierarchy of learning skills and abilities was tested by t-tests in pair-wise contrasts.

Limitations of the Study

The problem of developing a marking and reporting system is tackled in this project in a manner that does not seem to have been attempted before. The problem is approached through the use of a theoretical framework based on research findings and a feasibility study that tests out selected educational devices and mechanisms that seemed to best meet the criteria of the theoretical framework.

The feasibility study was carried out in the field in an available environment, but not the one suggested by the theoretical framework. Ideally an environment through which a learner moves according to his own rhythms and modes of learning and an instructional milieu in which many types of teachers and learning aids provide facilitative review would be the environment of choice. However, the learning conditions were of the type generally found in most schools--the traditional grade-level, letter grading structured classrooms. The mastery strategy was designed to fit into this time-confined structure and while the strategy

was modified in the school year, 1976-77, no attempt was made to perfect it.

The tests used were developed by the author and items from them were randomly selected and submitted to independent judges in an attempt to establish construct validity, but no attempt was made to determine item difficulty in terms of content and thus perhaps improve the tests. The tests were used as developed based on a finding of no significant variances with the author and a set of four independent judges at the 0.01 level of confidence in criterion-referencing the test items to the cognitive objectives under consideration.

To perfect both the mastery strategy and the instruments will take much more research and training of judges. Neither of these tasks, however, were perceived as being a requirement for the kind of study which simply tests out the feasibility of new uses for old mechanisms and claims of construct validity for cognitive objectives so as to suggest their viability in a new type marking system.

The use of a sample that was not random, that is positively skewed, indicating relatively few high achievement scores, and one that contains no students above the seventy-first percentile, leaves open the question of generalizability. Logic suggests that if this sample can serve to support the hypotheses herein advanced, one could proceed with a high level of confidence to further large scale, more normative and systematic studies. The sample should suffice to serve the needs of a feasibility study

whose goal is to attempt to point toward a marking and reporting system that could be acceptable to both critics and proponents of letter grading.

CHAPTER IV

FINDINGS OF THE STUDY

Literature Search-

The basic purpose of this study was to move toward the development of a marking and reporting system that could meet the criteria of those who have been critical of present day letter grading systems, while not ignoring the large measure of support for such systems. Toward such an end the literature was searched thoroughly in an attempt to isolate the major factors that were responsible for the criticism of letter grades as well as those factors which were responsible for the widespread reluctance to abandon a system of demonstrated low validity and reliability. This procedure was seen as offering the possibility of developing a theoretical framework from which to proceed. Thus a definitive review of the many variant studies undertaken vis-avis these factors was essayed - the goal being to establish the fact that these factors have withstood the tests of time and are supported by extensive research.

The major factors that emerged as prerequisites for a marking and reporting system that could have the possibility of being acceptable to both critics and proponents alike

were six in number. They were that the system had to be: (1) highly valid, (2) highly reliable, (3) administratively functional (4) a constructive communicator, (5) able to alleviate some of the negative side-effects of letter-grading, and (6) a good motivator.

The above framework, and the results of an analysis of alternative systems in general use became guides for a second literature search. This time educational mechanisms that had the greatest probability of meeting the criteria of the established framework were plucked from the literature. The following devices emerged:

1. A mastery strategy.

   The evidence indicated that such strategies tend to maximize achievement, reduce negative side-effects and have the potential for a more constructive communication system than letter grading.

2. The six major cognitive objectives developed by Bloom et al.

   If increased validity and reliability were to be achieved, the evidence pointed to the need for establishing clearly delineated objectives that define the domains of performance being measured and that can provide uniform standards without jeopardizing variety in content and teaching philosophies. For it is the vastly different interpretations of what marks represent that is largely responsible

for the relatively low validity and reliability found in letter grading systems.

Much has been written about objectives. The objectives that generally have been proposed and used, however, have been largely content-oriented. These have not met with wide acceptance for they are cumbersome, i.e. too large in number, and lacking in the kind of generality that would make them useful in any educational setting. Thus while the use of content objectives can provide a measurable domain, their use is limited in evaluative procedures.

The six major cognitive objectives developed by Bloom et al do not seem to have the deficiencies of content objectives. They are only six in number, making them administratively functional, and there is experimental evidence which supports their generality and even a hierarchical nature. This latter quality has the potential of adding a new dimension to an evaluative process in that it could aid in discriminating different learning patterns among students, thus enabling better guidance procedures.

3. Criterion-referenced tests.

If students are to be evaluated in terms of cognitive objectives, then summative

evaluative procedures must be criterion-referenced to such objectives for maintenance of a highly valid and reliable evaluative process. Bloom points out how the phraseology of a content question can be used to elicit any given cognitive objective, or learning behavior.

4. A criterion measure.

The work of Block suggests that an 85% criterion measure works best to maximize attitudes while not appreciably lowering achievement goals.

5. A mastery cognitive profile.

The use of a marking system based on an absolute measure seems to offer more potential than a relative marking system for improving self-image, for raising aspiration levels, for decreasing anxiety levels, for reversing anti-social attitudes and behaviors, and for developing intrinsic motivation. Absolute marking systems avoid the question of the ethics of relegating by definition alone a sizeable majority of students to the average or below-average portions of a scale.

A mastery cognitive profile seems to have the potential for serving as a highly valid and reliable measure of achievement, and for providing the kind of administrative

functionality needed to compete with letter grading.

Feasibility Study -- Hypotheses

The results of the feasibility study devised for two traditionally structured high school chemistry classes consisting of thirteen females and eleven males for a total of twenty-four students generally supported the above mechanisms.

The following hypotheses were investigated: (1) tests can be developed containing questions which are validly criterion-referenced to the first three major cognitive objectives developed by Bloom et al; (2) mastery strategies tend to maximize achievement; (3) the first three major cognitive objectives possess a generality that cuts across content areas, have a hierarchical nature indicating that they tend to represent an increasing complexity of skills, and are able to differentiate learning patterns; and (4) a mastery cognitive profile can be developed which is administratively functional.

Specific statistical null hypotheses were:

1. there are no significant differences at the 0.01 level of confidence between independent judges in criterion-referencing randomly selected chemistry test items to the first three major cognitives objective developed by Bloom et al. (The 0.01 level of confidence

was chosen so as to avoid making a Type I error and thus possibly finding differences where differences did not indeed exist).

2. there are no significant differences at the 0.05 level of confidence in the average percent error found on the one-shot standardized Anderson-Fisk Chemistry Test and the average percent error found on tests designed for mastery learning conditions which test for the same objective and similar units of content.

3. a. there are no significant correlations at the 0.05 level of confidence in the rank orders of each testee based on mean errors found on four sets of tests designed to test for the same cognitive objective, but in different areas of content.

   b. there are no significant differences at the 0.05 level of confidence in the means of errors made by testees on three sets of tests, each criterion-referenced to a different cognitive objective, but in the same areas of content.

   c. there are no significant differences at the 0.05 level of confidence in rank orders obtained by each testee on three different sets of tests each criterion-referenced to a different cognitive objective, but in the same

content areas.

The apriori hypotheses with which the statistical analyses were approached and their related findings were as follows:

1. the null hypothesis would be supported with a finding of no significant differences among independent judges in criterion-referencing randomly selected test items to the first three major cognitive objectives.

   A manovax computer analysis of variance of the percent of independent judges who criterion-referenced twenty-six randomly selected chemistry test items to one of the four specific cognitive objectives indicated no significant differences amongst the judges at the 0.01 level of confidence (this level of confidence was chosen to avoid concluding falsely that a difference exists when in fact it does not i.e. to avoid making a Type I error). Table 2 is a summary of the findings.

TABLE 2

ANALYSIS OF VARIANCE AMONG FIVE INDEPENDENT JUDGES

Source	SS	DF	MS	F	P less than
Item	1.446	25	0.058	1.724	0.037
Objectives	0.283	3	0.094	2.812	0.045
Error	2.517	75	0.034		

2. The null hypothesis would not be supported, for the findings would support the alternative hypothesis that a mastery strategy tends to maximize achievement.

An SPSS t-test of paired samples, which compared the average error made on 640 mastery test items with the average error made on thirteen questions from the standardized Anderson-Fisk Chemistry Test, all criterion-referenced to the same objective - knowledge - indicated a significantly higher error score on the standardized Anderson-Fisk Test at the 0.05 level of confidence. A summary of the findings can be seen in Table 3.

TABLE 3

t-TEST COMPARISON OF MEAN ERRORS ON MASTERY TESTS VS ANDERSON-FISK TESTS

Variable	N	Mean	SD	SE	t	DF	2-Tail Prob
Mastery	24	0.2854	0.087	0.018	-5.73	23	0.0001
A-F	24	0.4687	0.181	0.037			

3. a. the null hypothesis would not be supported by the finding that there are significant correlations in rank orders of testees derived from means of errors on tests designed to test for the same objective but in different units of content, thus indicating that these objectives do tend to cut across content areas.

A nonparametric correlation of rank orders run on the computer by way of the SPSS program indicated significant correlations at the 0.05 level of confidence. Table 4 summarizes these findings.

TABLE 4

CORRELATIONS OF RANK ORDERS ON FOUR DIFFERENT UNITS OF CONTENT

Variable Pair	N	Spearman Rho	Significance
A with B	24	0.7036	0.001
A with C	24	0.7783	0.001
A with D	24	0.5410	0.003
B with C	24	0.6845	0.001
B with D	24	0.5166	0.005
C with D	24	0.5165	0.005

3. b. the null hypothesis would not be supported for the findings would support the alternative hypothesis that there are significant differences in means of errors obtained on three different sets of mastery tests, each criterion-referenced to a different objective; and that the means of the errors would increase consonant with the hierarchical nature of the objectives.

An SPSS t-test of paired samples indicated that there were significant differences at the 0.05 level of confidence in the means of the errors obtained on tests with questions criterion-referenced to the cognitive objectives - knowledge (K), comprehension (C), and application (A). The means of the errors

did, indeed, increase with increasing complexity of cognitive objectives. The following table summarizes the analyses.

TABLE 5

t-TEST COMPARISONS OF MEAN SCORES ON TESTS CRITERION-REFERENCED TO THE FIRST THREE COGNITIVE OBJECTIVES

Variable	N	Mean	SD	SE	t	DF	2-Tail Prob
K	24	4.8483	1.755	0.358	-23.26	23	0.0001
C	24	10.3950	2.215	0.452			
K	24	4.8483	1.755	0.358	-41.07	23	0.0001
A	24	16.0712	2.129	0.435			
C	24	10.3950	2.215	0.452	-20.54	23	0.0001
A	24	16.0712	2.129	0.435			

3. c. the null hypothesis would not be supported indicating that there are significant differences in rank orders derived from mean errors made on tests criterion-referenced to the three different cognitive objectives, but in the same content areas, thus indicating that questions criterion-referenced to different cognitive objectives do tend to expose different learning patterns.

A manovax analysis of variance of such rank orders yielded significant differences in rank orders derived in such fashion at the 0.05 level of confidence, indicating that with the

use of cognitive objectives differences in learning patterns can be revealed. Following is a summary of the statistical analysis.

TABLE 6

ANALYSIS OF VARIANCE OF RANK ORDERS ON TESTS CRITERION-REFERENCED TO THE FIRST THREE COGNITIVE OBJECTIVES

Source	SS	DF	MS	F	P less than
Rank	2984.529	23	129.762	12.925	0.001
Obj	0.000	2	0.000	0.000	1.000
Error	461.817	46	10.040		

Feasibility Study -- Empirical Observations

To enhance the reliability of this study, replication was used in place of the generally used procedures of random and normative sampling and control groupings. For this purpose sixty-four tests were developed which were criterion-referenced to the first three major cognitive objectives developed by Bloom et al. They covered most of the content of the course. A criterion measure for mastery was established as a score of 80% correct answers. Mastery was defined as reaching the criterion measure on three out of four possible tests in any given unit of content in one of four specified objectives. (Based on research findings that success breeds success, the cognitive objective

knowledge was subdivided into knowledge of terminology and knowledge of specifics. The expectation was that starting these students off with the most basic of learning skills, an area in which the large majority of them could experience success, would help to create positive attitudes toward the mastery process.)

Thus the study was set up to allow students to demonstrate mastery in the course of the forty-week school year in four different units of content for three different cognitive objectives. But administering this number of tests per year to students who test at the lower ends of a standardized achievement scale proved (1) too arduous a project for that amount of time, and (2) that few of these students could and/or would demonstrate successful mastery for all three content areas in more than one cognitive area during the forty-week school period.

Table 7 indicates the extent to which the students were able to demonstrate the required mastery in the school year 1976-77.

TABLE 7

MASTERED OBJECTIVES*

N	KNOWLEDGE (Terminology)				KNOWLEDGE (Specifics)				COMPREHENSION				APPLICATION			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	M/4	M/3	M/3	M/6	2/13	0/6	0/8	0/4	0/4	0/0	0/0	0/0	0/4	0/0	0/0	0/0
2	M/4	M/3	M/3	M/8	M/8	0/4	0/8	0/4	0/8	0/0	0/0	0/0	0/4	0/0	0/0	0/0
3	M/3	M/3	M/4	M/3	M/7	0/6	0/8	0/4	0/11	0/0	0/0	0/0	0/0	0/0	0/0	0/0
4	M/5	M/3	M/4	M/8	M/8	0/6	0/8	0/4	1/8	0/0	0/0	0/0	0/4	0/0	0/0	0/0
5	M/4	M/3	M/3	M/5	M/8	0/6	0/8	0/4	1/9	0/0	0/0	0/0	0/4	0/0	0/0	0/0
6	M/4	M/4	M/3	M/8	M/12	0/5	0/8	0/4	0/7	0/0	0/0	0/0	0/4	0/0	0/0	0/0
7	M/3	M/3	M/4	M/7	M/6	0/6	0/8	0/4	0/11	0/0	0/0	0/0	0/4	0/0	0/0	0/0
8	M/3	M/3	M/3	1/2	M/9	0/6	0/8	0/4	0/12	0/0	0/0	0/0	0/4	0/0	0/0	0/0
9	M/6	1/6	0/4	M/4	0/8	0/6	0/8	0/4	0/6	0/0	0/0	0/0	0/4	0/0	0/0	0/0
10	M/3	XXX	M/5	2/7	0/9	0/6	0/8	0/4	0/7	0/0	0/0	0/0	0/4	0/0	0/0	0/0
11	M/7	XXX	2/7	M/6	0/11	0/6	0/8	0/4	0/7	0/0	0/0	0/0	0/4	0/0	0/0	0/0
12	M/5	M/3	M/3	M/4	M/11	0/6	0/8	0/4	0/7	0/0	0/0	0/0	0/4	0/0	0/0	0/0
13	M/3	M/3	M/3	M/6	M/6	0/6	0/8	0/4	M/13	0/0	0/0	0/0	0/4	0/0	0/0	0/0
14	M/5	M/5	M/7	M/7	0/12	0/6	0/8	0/4	0/4	0/0	0/0	0/0	0/4	0/0	0/0	0/0
15	M/4	M/3	M/3	M/5	M/8	0/6	0/8	0/4	0/6	0/0	0/0	0/0	0/4	0/0	0/0	0/0
16	M/3	M/3	M/3	M/6	M/3	0/6	2/8	0/4	M/9	0/0	0/0	0/0	0/4	0/0	0/0	0/0
17	M/7	1/1	2/4	0/0	0/10	0/6	0/8	0/4	0/4	0/0	0/0	0/0	0/4	0/0	0/0	0/0
18	M/5	M/3	M/3	1/3	M/10	0/6	0/6	0/4	0/5	0/0	0/0	0/0	0/4	0/0	0/0	0/0
19	M/6	M/5	M/7	M/6	2/12	1/6	0/8	0/4	0/6	0/0	0/0	0/0	0/4	0/0	0/0	0/0
20	M/6	M/3	M/5	M/5	0/14	0/6	0/8	0/4	0/7	0/0	0/0	0/0	0/4	0/0	0/0	0/0
21	M/6	M/4	M/3	M/5	M/10	0/6	0/8	0/4	0/9	0/0	0/0	0/0	0/4	0/0	0/0	0/0
22	M/3	M/3	2/3	1/2	1/12	0/6	0/8	0/4	0/7	0/0	0/0	0/0	0/4	0/0	0/0	0/0
23	M/7	M/4	M/6	M/5	2/12	0/6	0/8	0/4	0/4	0/0	0/0	0/0	0/4	0/0	0/0	0/0
24	M/3	M/3	M/3	M/5	M/12	0/6	1/8	0/4	0/7	0/0	0/0	0/0	0/4	0/0	0/0	0/0

*The four cognitive objectives are divided into four columns, each column representing a given unit of content. The rows represent students. M indicates that three tests were mastered in the given unit of content for the given objective at an eighty percent criterion level. A number in place of M indicates that three tests were not mastered but indicates the actual number mastered. The number to the right of the slash indicates the actual number of tests taken. The symbol, XXX, was used to indicate that these students were caught cheating.

It was found that the six-week cut-off marking periods had a dampening effect on the motivation of students to demonstrate mastery. They were acutely aware that a grade had to be assigned and they knew that the grade could not be assigned on the basis of demonstrated mastery, if most of the students chose not to do so. Such circumstances force the evaluator into one of three positions: 1) lower the criterion measure, 2) revert to a normative marking pattern, or 3) fail the large majority of the students. Under such conditions the teacher has little clout in enforcing high standards from the students, or in the case of a mastery strategy, enforcing mastery performance, since almost no teacher would take the option of failing the large majority of a class.

There seemed to be another variable that played a crucial role in the problem of motivating students towards maximum achievement. The students in the class on which the study was field-tested exhibited a higher degree of motivation to demonstrate mastery than in the class on which the study was carried out. Observations of student behavior pointed to the best students in the class playing a leading role in this regard. In the latter class, the best students were, in spite of an obvious need to excel, hostile to the learning process due to conflicting aspirations. When they chose to lower their achievement level, the rest of the class seemed to follow their lead.

Nonetheless, the mean score on the Anderson-Fisk Chemistry Test was higher for the class of '77 than for the

class of '76, and even higher than for the class of '78 which was taught in traditional fashion. Students in the classes were also evaluated summatively at the end of the school year with a teacher-developed test. This was a traditionally designed test whose items were content-oriented with no conscious concern for cognitive objectives. The items came from Chem study, regents and text-book tests. Table 8 contains the mean scores and ranges for these tests.

TABLE 8

COMPARISONS OF MEAN SCORES ON THE SAME TESTS FOR DIFFERING TEACHING STRATEGIES

		Mean Scores and Ranges			
Teaching Strategy	N	A-F	Range	Traditional	Range
Mastery '76	15	24.36	17-39	29.20	14-27
Mastery '77	24	25.30	16-36	30.57	23-42
Traditional '78	20	22.95	17-32	30.55	20-54

Observational analysis reveals that the mean score on the traditionally developed test is approximately the same for both the '77 class taught by a mastery strategy and the '78 class taught by traditional techniques, even though the class of '77 had a higher mean score on the standardized Anderson-Fisk Chemistry Test than the class of '78. Standardized test items are geared to the average chemistry class in the United States. Teacher-devised tests are almost always geared to the specific content on which any

given teacher focuses. Perhaps such findings suggest that mastery strategies tend to force students to learn a larger variety of content.

The fact that the class of '77 had higher mean scores than the class of '76 on both the traditional teacher developed test and the traditional standardized Anderson-Fisk test could be due to differences in the administration of the mastery tests. In the '76 class, the students were given a great deal of leeway. They were allowed to take tests at their own discretion. Some students procrastinated too much. This led to two changes for the class of '77. Students in this class had to take tests on assigned test dates, and were required to demonstrate mastery in the most basic objective, Knowledge of Terminology, in order to pass the course. This greater exposure to test taking and the requirement for passing probably played crucial roles in the higher mean scores that this class achieved.

It seems significant to note that the three classes did better than it was possible to predict from their relatively low SAT scores. The mean for these groups for the SAT scores is at about the 24th percentile. However, on the Anderson-Fisk Chemistry Test, their mean achievement score fell at approximately the 63rd percentile. It would seem that such evidence would support mastery strategists in their contention that achievement has little relationship to aptitude.

CHAPTER V

CONCLUSIONS AND RECOMMENDATIONS

Conclusions

Letter grading plays a significant role in the lives of people. It is the basis of selection processes which importantly affect careers and future life styles. It even seems to play havoc with individual perceptions, self-concepts, attitudes and behaviors.

There exists abundant research evidence supporting the claims of its critics that letter grading is neither a highly valid nor reliable measure of educational outcomes. There exists significant evidence that letter grading can seriously affect a student's aspiration level, self-image, behaviors, mental health, attitudes, values and motivation. The fact that letter grading has been in use for almost eighty years in spite of such evidence suggests that if letter grading is to be replaced, it will be replaced only by a system that is as administratively functional as letter grading.

That an alternative evaluative system is sorely needed is evident. Alternative systems used were reviewed and were found to be either not administratively functional or

lacking the potential for being more highly valid and/or reliable than letter grades, and/or more constructive in terms of affective consequences.

There seem to be three sources of variance responsible for both the relatively low validity and reliability of letter grading found through research studies (reliability coefficients are generally of the order of 0.60). The differences in teachers' standards are pin-pointed as the chief cause of variations in letter grading. These differences are due to different foci given units of content by different teachers in accordance with their individual teaching philosophies, as well as differences in criterion measures.

Another source of variation seems to come from attributes that students bring to the test performance that have been shown to confound achievement; one affecting the other and vice versa. These variables relate to such factors as aspiration level, self-image, anxiety and prior success experiences.

The third crucial source responsible for variations in grading comes from those variables that uniquely interact between teacher and pupil to prejudice the very grading process itself. Interacting variables such as personality, sex and attitudes on the part of both student and teacher; fatigue, personal pressures and prior recall of the teacher; and even penmanship of the students have been shown to play significant roles in influencing this process.

The technique found that seemed to have the greatest

potential for bringing organization to the profusion of standards is the use of the kind of clearly delineated objectives that define a domain of measurable performance. This device is widely perceived as crucial for the development of a highly valid evaluative process that would clarify present differences about performance requirements. Such objectives are not widely used, however, for the kinds of objectives that have been touted are generally content-oriented.

Criticism of these kinds of objectives falls into three major categories: 1) content objectives can be sufficient for defining a specific domain of measureable achievement, making them useful and valid in any given classroom at any given time for measuring student progress, however when used in this manner they tend to place restrictions on the teaching process in that particular classroom; 2) they lack the kind of generality that would make any one set of objectives useful in all classrooms; and 3 when specific content objectives are defined they end up being fairly large in number, making them not very functional administratively.

This study found that the first three of the six major cognitive objectives developed by Bloom et al were feasible alternatives to content objectives. Bloom et al point out that content can be learned at differing cognitive levels, representing a hierarchy of cognitive skills and abilities. Sheer memorization represents the first level - Knowledge; the ability to translate, interpret and extrapolate

knowledge represents the second level - Comprehension; and the ability to use one's knowledge and comprehension represents the third level - Application. Thus the same content can elicit different cognitive behaviors, simply by varying the phraseology of a question (see Appendix C). And the same kinds of cognitive behaviors can be elicited by differing bodies of content. This quality gives these objectives the kind of generality needed to cut across classrooms. Thus by defining a domain of measurable learning performance with objectives that can be used in any educational setting, the promise of a standardized measure looms on the horizon.

The cognitive objectives have three more desired characteristics. They do not place undesirable restrictions on what or in what manner content is to be used in any given course, they are are only six in number making them administratively functional, and test items seem to be readily criterion-referenced to these objectives.

This study's findings support the feasibility of expecting independent judges to criterion-reference test items to these cognitive objectives with significant validity. This study's findings support the findings of Kropp and Stoker and the claims of Bloom et al that these objectives have the needed generality to cut across classrooms, and that they have a hierarchical nature that discriminates cognitive skills and abilities. (This latter quality could, moreover, prove to be of considerable aid in improving guidance processes.)

In summary, this study supports the defining of domains of measurable performance for the purpose of evaluating educational outcomes but suggests that the domains so defined be the six major cognitive objectives developed by Bloom et al. For only such objectives can provide the kind of standardization needed to develop a highly valid and reliable measure of educational outcomes that could be useful in any educational setting and with any educational philosphy in an administratively functional manner.

The other aspect of letter grading that concerned critics was evidence that the kind of normative system in which letter grades are used affects negatively such personal attributes of students as values, attitudes, behaviors, self-image, aspiration levels and even achievement.

Research evidence indicates that there are several techniques in use that can alleviate some of these negative side-effects. One is the use of an absolute measure which evaluates the extent of a student's progress rather than his/her standing in a given group, and thus aids in the reduction of the debilitative effects of chronic failure. Another is the use of mastery strategies with an agreed-upon criterion measure in the 85% correct answer range. (Block demonstrated that this percent mastery seems to maximize attitudes without lowering achievement appreciably). Significant research evidence suggests that the use of mastery strategies in teaching tends to raise achievement levels of students, even in structured classrooms using

traditional grading practices. For the evidence points to positive relationships between achievement and anxiety, achievement and self-image, achievement and aspiration levels. While other variables such as sex, class, boredom, personal pressures, attitudes, penmanship, and motivation have been shown to crucially confound not only achievement but the grading process, too. Perhaps it is because such strategies in which repeated test-taking is allowed tend to reduce the effect of those interacting variables which students bring to the testing process as well as those variable which interact with teacher and student to influence the very grading process itself.

This study investigated the use of a mastery strategy with a criterion measure of 80% correct answers. The investigation was carried out in traditional classrooms traditionally graded by letter grades. The findings supported the claims of the mastery strategy proponents that mastery strategies tend to move students towards maximizing achievement. However, empirical observation suggests that such strategies would have more potential in educational settings that are structured according to a student's own learning rhythms and modes. It appears that the structured classroom with its cut-off marking periods seems to militate against the mastering of learning materials. The mere knowledge that the student must be evaluated in a limited period of time has the effect of placing control of the evaluative process in the hands of the students and not the teacher or educational institution. For in such systems,

and particularly those in which normative marking is employed, research evidence supports the fact that teachers tend to teach to, and grade on, the basis of the class "average", regardless of the actual achievement level of any given class.

The kind of structure that suggests itself, and that would be in keeping with the dictates of the Carroll postulate, is a continuum of learning made up of learning centers where subject matter is taught at varying levels of content and cognitive difficulty, and through which a student could move according to his/her own learning patterns. Students thus would move horizontally across content and cognitive level, but they would move vertically to a higher cognitive or content level only when they had demonstrated mastery at a lower level. For such a system, educators could develop a criterion measure that would be used by all to define mastery.

In such a mastery system, not time bound, the date that the student was able to demonstrate the desired mastery could be simply recorded on a mastery cognitive profile. The area and level of content and the cognitive objectives to which the course had been geared could be specified on the profile, thus providing a visual picture to administrators of the particular content area and level in which a specified cognitive objective had been mastered (see Appendix A).

Such a profile would have to be filled out only for those students who had mastered the material of a given

course. Teachers thus would be relieved of the burden of averaging test grades, and the questionable ethics of designating a sizeable portion of their student body to a failure category. Students would be relieved of the debilitative burden of being labeled failures. Perhaps most importantly for all students, the evaluative process would aid in returning meaning to the awarding of a degree or a diploma. It would be a relatively simple process to establish minimum requirements in terms of mastered cognitive objectives in any given area and level of content for the awarding of those degrees, and they would represent under such conditions a body of mastered material, unlike present degrees which often represent material half learned or, as is all too frequently the case, not learned at all.

Whether learning is accomplished in a traditional learning structure or by a mastery strategy in a traditional classroom setting, or whether one follows the logical dictates of the Carroll postulate that learning is in large measure a function of time and then sets up a continuum of learning through which students can move according to their own learning patterns, an evaluation system based on cognitive objectives could contribute considerably toward improving the validity and reliability of any process measuring learning progress, and its administration need not be cumbersome. In traditional systems a percent score could be recorded in place of date of mastery.

Introducing a mastery learning strategy into the teaching process simply provides an added measure for

assuring that the high validity and reliability that would be possible with the use of cognitive objectives would be supported by the teaching process. For mastery strategies seem to reduce the negative effects of intervening variables on student behavior as well as their subsequent effect on achievement. Placing such strategies within the framework of a continuum of learning would carry the evaluative process one step further towards the maintenance of high validity and further improvement of students' behaviors and achievement.

This study supports the feasibility of expecting teachers to be readily able to develop tests criterion-referenced to the first three major cognitive objectives. However, one cannot ignore the knowledge that the interaction of the classroom teacher and student can produce its own source of interacting variables that have been shown to color a teacher's judgment re summative evaluation. One can only conclude that even with the greater potential for increased validity and reliability for the evaluative process with the use of cognitive objectives and mastery strategies within a continuum of learning, there would still remain intervening and interacting variables that could keep the validity and reliabilty of such measurement from achieving its greatest potential.

Perhaps the time has come, therefore, to concern ourselves with the suggestions made over the years by both students and researchers that the evaluative process be removed from the purview of the classroom teacher. Creating

district testing centers manned by test development experts could create a climate where it would be more possible than in the classroom to develop tests that are reflective of highly valid and highly reliable samples of content and cognitive objectives. Centers which involve the testor only in the summative process and which keep the interactions between testor and testee to a minimum would reduce the effect of the kinds of interacting variables that have been shown can influence final evaluations.

At such centers regular intervals could be set aside for students to demonstrate mastery of cognitive objectives in subjects required for degrees or diplomas. Removing the summative evaluative process from the classroom could provide added benefits. It would allow the teacher more time to develop better rapport with the students and provide the kind of climate that would permit the teacher to concentrate on working as a guide through the learning process, on providing for more effective formative evaluation, and on dealing more constructively with those affective behaviors that seem to be related to achievement.

Recommendations-

The statistical analyses of and the empirical observations gleaned from this study suggest the need for further investigation of the following:

1. A replication of the study with a more normative sample than the skewed sample used herein.

2. A replication of the study in a nongraded school in which a mastery strategy can be employed which is not time bound by six-week marking periods.
3. A replication of the study on a college level to include the remaining three objectives that were not used in this study.
4. A study which attempted to determine whether the very requirement of mastery itself is the most crucial factor in motivating students to demonstrate mastery.
5. More extensive research on the whole subject of motivation.
6. More in-depth studies of the Carroll postulate that achievement is a function of time. For if findings continue to support the postulate and if mastery strategies continue to demonstrate that they do maximize achievement, then the logical conclusion from such findings would be to restructure schools so that they are learning centers, where students can learn according to their own learning modes and rhythms of learning.

If the learning process is to be evaluated in as fair a manner as other substances of material value that contribute to life styles are weighed and measured, then we must have a

learning system that: (1) provides the opportunity to all students to demonstrate that they can master a body of knowledge by giving them sufficient time and learning aids for the process, (2) evaluates students in terms of what they have mastered, (3) measures educational outcomes with a standardized measuring tool of the kind that can be derived from the six major cognitive objectives developed by Bloom et al, and (4) evaluates students' summative performance outside the purview of the classroom.

By moving in these directions, a marking and reporting system should result which could be acceptable to both critics and proponents of present letter-grading systems. For in such a system intervening and interacting variables that effectively reduce the validity and reliability of present marking and reporting systems based on normative procedures would be removed from the summative evaluative process, negative side-effects could be reduced, and a relatively constructive communication system would be available. The use of cognitive objectives would clearly define the domain of learning being measured and a fixed criterion measure, accepted by educators in general as one which maximizes achievement and attitudes, would effectively standardize achievement outcomes without restricting the teaching process in the classroom. The use of an absolute measure should also alleviate the effects of chronic failure, undue competitiveness, widespread cheating and distorted educational patterns that make the appearance of learning more important than the actual learning process.

Hopefully intrinsic motivation would return because of the feed-back nature of mastery strategies and the knowledge of precisely what educational outcomes are being measured. The administrative aspects of marking and reporting would be removed from the classroom, and a functional report could be provided.

APPENDIX A: MASTERY COGNITIVE PROFILE

MASTERY COGNITIVE PROFILE

INORGANIC CHEMISTRY - LEVEL I

NAME _____ STARTING DATE_____

COGNITIVE DOMAIN OBJECTIVES:[1] MASTERY DATE_____

1.10 - KNOWLEDGE OF TERMINOLOGY _____
Knowing meaning of words or terms.

1.11 - KNOWLEDGE OF SPECIFIC FACTS, TRENDS,
1.32 CLASSIFICATIONS, PRINCIPLES, THEORIES. _____

2.00 - COMPREHENSION. _____
Ability to translate, interpret,
and extrapolate.

3.00 - APPLICATION. _____
Ability to predict, solve problesm, and
use abstractions in concrete situations.

---

[1]Numbering system follows that in the Taxonomy of Educational Objectives: Cognitive Domain.

APPENDIX B: FORMS USED TO VALIDATE TESTS

23049 Nottingham Drive
Birmingham, Michigan 48010
September 10, 1976

Dear Colleague:

The enclosed test is part of a research project on the cognitive domain. You may keep and use the test, if you so choose. Needed for the research project are the independent judgments of experts. Therefore, the giving of a half-hour of your time to classify the questions into the four categories listed would be sincerely appreciated. While you will find that there is frequently overlap in terms of the objectives, please choose the objective you perceive is most represented by the question. Simply make a check mark in the appropriate column.

A prompt response will be gratefully received. Enclosed, also, is a stamped self-addressed envelope for the return of the accompanying form on which the hoped-for categories will be recorded.

Sincerely yours,

Lillian Rosenberg Hurwitz
(Mrs. Jacob I. Hurwitz)
Chemistry Teacher

COGNITIVE DOMAIN OBJECTIVES

QUESTION NUMBER	KNOWLEDGE TERMINOLOGY	KNOWLEDGE SPECIFICS	COMPREHENSION	APPLICATION
1.				
2.				
3.				
4.				
5.				
6.				
7.				
8.				
9.				
10.				
11.				
12.				
13.				
14.				
15.				
16.				
17.				
18.				
19.				
20.				
21.				
22.				
23.				
24.				
25.				
26.				

Cognitive Domain Objectives:

1.00 Knowledge of Terminology
knowing meaning of word or term.

1.XX Knowledge of Trends, Theories, Facts, Etc.

2.00 Comprehension
ability to extrapolate, translate and interpret theory.

3.00 Application
ability to predict, use abstractions in concrete situations.

RANDOMLY SELECTED TEST ITEMS FROM A POPULATION OF 1,280

ITEMS CRITERION-REFERENCED TO COGNITIVE OBJECTIVES

OBJ-

(2.00) 1. The electrical force between two particles will decrease if

a. the distance between the particles decreases.
b. the distance between the particles increases.
c. only if the charges on the particles increase.
d. none of these.

(1.20) 2. The expression which best expresses the relationship between the speed of light, its frequency and wave length is

a. c = f b. c = f
c. f/ = c d. = cf

(1.20) 3. All of the following statements concerning the e/m ratio of an electron are true except

a. Millikan is given the credit for its determination.
b. It is expressed in coulombs/gram.
c. Its determination was accomplished by the use of a magnetic field.
d. none of these.

(3.00) 4. Write the balanced equation for the following reaction:
Phosphorous pentoxide and water producing phosphoric acid.

(2.00) 5. When an aerosol bomb is used to spray an insectcide, the can becomes cold because

a. heat is being absorbed. b. heat is being given off.
c. heat is being adsorbed. d. ice forms inside the can.

(1.00) 6. The word that best fits the amount of space occupied in three dimensions.

_____

(1.20) 7. The shape of the orbital in which the electron moves is indicated by

a. the quantum number. b. orbital quantum number.
c. magnetic quantum number. d. spin quantum number.

(3.00) 8. If the energy difference between two energy levels about an atom is 46.12 kcal/mole, what frequency of light might one expect emitted when an electron spinning about that atom drops from a higher to a lower state?

_____

What would one expect the color of the light to be?

_____

(1.20) 9. Radioactive substances have all the following properties EXCEPT

a. their nuclei disintegrate spontaneously.
b. they may give off electrons from their nuclei spontaneously.
c. their atoms can decay into new kinds of atoms.
d. they give off cosmic rays.

(1.20) 10. The color of light in the frequency range of 10^{12}

a. red b. green
c. blue d. colorless

(3.00) 11. CO absorbs light at frequencies near 1.2×10^{11}, 6.4×10^{13}, and 1.5×10^{15} vibrations/sec. Name the spectral regions in which it absorbs and the color of the CO as derived from this information.

_____

(1.20) 12. Which of the following is a compound?

a. water b. oxygen
c. carbon d. iron

(1.00) 13. The component of a solution that is present in the smaller amount is called the

_____

(1.00) 14. The word that best fits the phrase, capacity to do work

_____

(1.00) 15. Which of the following statements concerning the model of the atom is not true?

a. Its mass is concentrated in its nucleus.
b. All the many different atoms have the same number of positive charges in their nuclei.
c. The number of protons in the nuclei of atoms is related to their atomic numbers.
d. Atoms of the same kind always have the same number of protons in their nuclei.

(2.00) 16. The average velocity of neon molecules at a given temperature and pressure should be the (the same as, more than, less than) the average velocity of an equal volume of helium molecules. CROSS OUT WRONG WORD.

(2.00) 17. The total number of atoms/molecule in the compound $Al_2(SO_4)_3$ is

_____

(3.00) 18. What chemical compound would you use to separate the fluoride ions from the bromide ions that you suspected were in an aqueous solution?

_____

Write the net ionic equation.

_____

(1.20) 19. Solutions are composed of which of the following two parts?

a. solutes b. salients
c. solvents d. electrodes

(3.00) 20. Calculate R from the following data:
No. of moles = 0.5 Pressure = 750 mm Hg
Volume = 12.2 liters Temperature = 298° C.

_____

(1.00) 21. Compounds formed with negative halide ions are called

_____

(1.00) 22. The amount of heat needed to change from the solid to the liquid phase is called

_____

(1.20) 23. The number of orbitals that make up the sublevel s is

a. 1 b. 2 c. 6 d. 3

(1.20) 24. Which of the statements concerning the isotopes of oxygen is not true?

a. They all have the same number of neutrons.
b. They all have the same number of protons.
c. They all have different mass numbers.
d. They all have the same number of electrons in the neutral state.

(3.00) 25. What volume would 1.02×10^2 moles of any gas occupy at room temperature and one atmosphere pressure?

_____.

(1.00) 26. Equal volumes of gases at the same temperature and pressure contain the same number of molecules. To whom is this Law credited?

_____.

APPENDIX C: WRITING QUESTIONS CRITERION-REFERENCED TO COGNITVE OBJECTIVES

Examples of Differences in Phraseology of Test Items Related to the Same Content but Designed to Elicit Specific Cognitive Responses

1.00 Knowledge-

1. The volume of a mole of any gas at STP is generally about

   a. 22.4 liters

   b. 1 liter

   c. depends on the weight of the gas.

   d. depends on the size of the molecules.

2.00 Comprehension-

2. Which of the following comparisons between the volume of a mole of carbon dioxide and the volume of a mole of oxygen at STP would one generally expect to be true?

   a. The volume of carbon dioxide is one and one half times greater than the volume of oxygen.

   b. The volume of oxygen is twelve liters less than the volume of carbon dioxide.

   c. The volume of carbon dioxide is about three times greater than the volume of oxygen.

   d. There is no difference in volume between the carbon dioxide and the oxygen.

3.00 Application-

3. You are planning to collect five moles of nitrogen at STP. Determine the number of liters you could anticipate so that you can prepare for the proper size container.

APPENDIX D: MARKING PROCEDURE

Step-by-Step Procedure for the Development of a Marking and Reporting System Consonant with the Findings of This Study

1. Divide content of course into units of study.
2. Select test items and criterion-reference them to relevant cognitive objectives.
3. Establish a criterion measure for mastery in the 80-85% range.
4. Develop a mastery strategy that will fit the time limits of your course.
5. Write a profile reflecting the area and level of content and cognitive objectives mastered by your students in your course.

BIBLIOGRAPHY

Asimov, Isaac. Realm of Measure. Boston, Houghton Mifflin Co., 1960. 182p.

Bayley, Nancy. "On The Growth of Intelligence." American Psychologist. 10, 805-818. 1955.

Cattel, Raymond B. Ed. Handbook of Experimental Psychology. Chicago, Rand McNally and Co. 1966. 959p.

Dressel, Paul L. and Lewis Mayhew. General Education: Explorations in Evaluations. Wash, D.C. American Council on Ed. 1954.

Dayton, Mitchell C. The Design of Educational Experiments. New York, McGraw-Hill Book Co., 1970. 441p.

Getzels J.W. and Jackson, P.W. Creativity and Intelligence: Exploration with Gifted Students. New York. John Wiley and Sons. 1962.

Guilford, J.P. The Nature of Intelligence. New York. McGraw-Hill Book Co., 1967. 65p.

_____ "The Structure of Intellect". Psych. Bull. 53:267-293 July, 1956.

Hertzberg, Alvin. and Edward Stone. Schools Are For Children. New York. Schocken Books. 1971.

Hoyt, Donald. "The Relation Between College Grades and Adult Achievement." ACT. Research Reports. Sept., 1965.

Hunt, McVicker J. Ed. Human Intelligence. New York, transaction books, 1972. 280p.

Leonard II, Wilbert M. Basic Social Statistics. New York, West Publishing Co., 1976. 468p.

Runyon, Richard P. and Audrey Haber. Fundamentals of Statistics. Mass., Addison-Wesley Publishing CO. 1972. 351P.

AUTOBIOGRAPHICAL STATEMENT

The author was born in Boston, Massachusetts in 1920, and attended Boston Public Schools, graduating in 1937. She then attended Simmons College, also in Massachusetts, where she majored in Chemistry. She graduated in 1944 after taking off three years to work at her chosen profession. In 1947 she married Jacob I, Hurwitz who later obtained his doctorate at the University of Michigan. They subsequently had three sons and when the youngest was twelve years old, the author returned to school to obtain a Masters Degree in Science Teaching They were living in New York at the time and the author attended Columbia University, receiving her degree in 1968, when she was elected to Kappa Delta Pi and Pi Lambda Theta.

The author had moved to Detroit in 1965 when her husband had accepted a position at Wayne State University's School of Social Work. Shortly after her arrival in Detroit the author was offered a position with the Hamtramck School System. She has been teaching Chemistry and assorted other Sciences there since accepting the position almost fourteen years ago. She has also been Chairperson of the Science Department almost as long. In 1970 a National Science Foundation Scholarship in Physics brought her to Wayne, which led her to pursue doctoral studies there.