ED 360 334 TM 020 131

AUTHOR	Nitko, Anthony J.; Niemierko, Boleslaw
TITLE	Qualitative Letter Grade Standards for Teacher-Made Summative Classroom Assessments.
PUB DATE	Apr 93
NOTE	33p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Academic Achievement; Cognitive Processes; College Faculty; College Students; Educational Assessment; Elementary Secondary Education; *Evaluation Methods; Grades (Scholastic); *Grading; Higher Education; Holistic Evaluation; Performance; Statistics; *Student Evaluation; Summative Evaluation; *Teacher Attitudes; Teacher Made Tests; *Thinking Skills
IDENTIFIERS	*Performance Based Evaluation

ABSTRACT

Qualitative methods for defining and assigning letter grades on classroom tests were studied. A hierarchical letter-grade scale is described that combines teachers' judgments of the importance of subject matter concepts and their classification of assessment tasks as reflecting cognitive processing skills identified from recent research. Using this grade assignment procedure shifts teachers' thinking so that grades on summative classroom assessments reflect quality levels of student thinking instead of simply the number of points students attain. The model incorporates a teacher's perception of the level of thinking that a student must use to perform a task and the value a teacher places on successful performance. Combining the thinking skills and importance factors can be done by crossing the factors in a two-way table, which can then be used to organize tasks into testlets or subtests. Four possible teacher-specific grading models can be derived from the ways teachers associate grades with thinking skills and subject content. The method was tested with 5 statistics instructors using an existing test and with 2 instructors whose 48 students took an examination designed for the grading method. This approach to grading provides an interesting rationale and merits further investigation. Four tables present study findings, and six figures illustrate the discussion. (SLD)

Qualitative Letter Grade Standards

for Teacher-Made Summative Classroom Assessments

by

Anthony J. Nitko
University of Pittsburgh

and

Boleslaw Niemierko
University of Gdansk

A paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, Georgia, April, 1993

Qualitative Letter Grade Standards for Teacher-Made Classroom Tests[1]

by

Anthony J. Nitko

Dept. Psychology in Education, School of Education, University of Pittsburgh

and

Boleslaw Niemierko

Dept. Education, University of Gdansk

Purpose

This paper investigates qualitative methods for defining and assigning letter grades on classroom tests. We describe a hierarchical letter grade quality scale that combines teachers' judgments of the importance of subject-matter concepts and their classification of assessment tasks as reflecting cognitive processing skills identified from recent research. The purpose of using this grade assignment procedure is to shift teachers' thinking so that grades on summative classroom assessments reflect quality levels of students' thinking instead of simply the number of points they attain. This paper is one of a planned series of studies that will explore general methods for using qualitative scales for assigning grades in various teacher-made tests.

Background and Perspective

Recently, Stiggins, Frisbie, and Griswold (1989) examined the recommendations of textbook authors in relation to the actual grading practices of teachers. Authors recommend that teachers base their grades strictly on achievement instead of on factors such as growth, interest, attitude, personality, or motivation level. Focusing on achievement encourages clear communication of student attainment through letter grade symbols. However, when one

[1] The authors would like to thank Mr. Fanmin Guo for completing some of the calculations reported in this paper.

considers recent studies in cognitive science, exactly what makes up achievement is either not well-defined or is open to question. Usually, teachers carry out assigning summative grades through quantitative means only, principally by summing points they have assigned to students' responses to assessment tasks. As a result, students and teachers may both see better achievement primarily as attaining a higher score on an assessment. Teachers assign higher grades to students who complete more assessment tasks correctly, usually ignoring the mix of thinking levels those correctly answered tasks represent.

Textbook authors do recommend planning classroom assessments using "thinking skills taxonomies." Authors often recommend constructing a two-way, subject-matter-topic by thinking skill grid: They recommend placing into the cells of the grid learning objectives or assessment tasks that require students to use particular thinking skills with particular content. The problem with this approach is not so much in the planning, but in the implementation. Teachers who follow the plan create tasks and assemble them into an assessment procedure, but they usually ignore the thinking skills framework when scoring and assigning grades.

Depending on the assessment, two students who have very different levels of cognition, could quite easily attain the same score. According to textbook recommendations, teachers should award the same grade to students attaining the same total score on a classroom assessr ιent. If that "same total score" reflects different levels of cognition, however, it is easy to see that its validity is open to question, even if teachers based letter grades strictly on achievement.

Assigning letter grades using total points only is a practice that fits closely with behaviorist principles such as using detailed inventories of behavioral objectives as teaching and assessment targets. As several researchers have recently pointed out, this behaviorist bias may inappropriately reflect how learning proceeds and ignores recent research from cognitive science (e.g., Frederiksen & Collins, 1989; Resnick & Resnick, 1991; Shephard, 1989, 1990). Often the behaviorist approach results in teachers drilling students on specific, isolated skills and postponing the teaching of higher-order thinking and reasoning until students learn the basic skills. Focusing the assignment of letter grades on students' demonstrated levels of thinking has the advantage of focusing teaching on how students think about the subject matter.

The approach we describe in this paper for establishing an ordinal letter grading scale is related to criterion-referencing to an ordered domain as described by Nitko (1980). It relies heavily on using the methodology of the multilevel criterion-referenced measurement system developed by Niemierko (1990) in Poland. Some of the methodology here is similar to a behavioral objectives approach which was used by Cox and Graham (1966) to develop a sequentially scaled series of testlets for placing students in a well-ordered arithmetic curriculum. Unlike the examples given in these sources, however, our approach focuses on "whole course" ideas and subject-matter "thinking skills", instead of on behavioral objectives in narrow skill domains. The approach here is more holistic (and perhaps less precise) than specifying behavioral objectives. It may be, however, more suitable for school settings than requiring specification of behavioral objectives.

The General Grading Model

The letter grade assignment model we investigated incorporates two interrelated factors. One factor is a teacher's perception of the level of thinking a student must use to perform a particular assessment task appropriately. The second is the value a teacher places on a student's successful performance of a particular assessment task. The model is used to derive a single letter grade scale (A = the highest grade) taking into account both factors. Before describing this process, however, we describe below each factor separately.

Thinking Skill Scale

The thinking skill factor was operatioɩ.alized by developing and applying a hierarchy-like thinking skill scale to a particular course of study (i.e., introductory statistical methods). The competencies and skill descriptions were adapted from Dimensions of Thinking: A Framework for Curriculum and Instruction (Marzano, Brant, Hughes, Jones, Presseisen, Rankin, and Suhor, 1988) and influenced by Guilford's (1950) comments about the particular subject matter. The authors of the Dimensions do not consider the core thinking skills to be hierarchically ordered, however. We used the Dimensions of Thinking because it is a summary of recent findings from cognitive science written for teachers and curriculum developers, not for researchers. The Dimensions suggest lessons and materials to enhance students' learning of "core thinking skills." Thus, if a grading scale could be linked to methods of teaching thinking skills within a subject-matter, assessment and thinking-skills instruction could be more closely aligned. Figure 1 is the thinking skill scale we used in this study.

Insert Figure 1 here

Although the thinking skills categories in Figure 1 are ordered, it should not be assumed that this is a teaching order or a letter grade order. As far as the assignment of letter grades is concerned, teachers may express preferences for one of at least two orders. The first is to assign the highest letter grades to students who demonstrate the higher-order thinking and the lowest grades to students demonstrating mainly knowledge of specific facts (concrete, declarative knowledge). A second possibility is to assign the highest grades to students who demonstrate the lower-order thinking skills, while assigning lower grades to those manifesting the higher-order skills. The second option may at first seem absurd. However, some teachers, especially in certain humanities (such as philosophy, fine arts, and humanistic pedagogy) tend to emphasize very general, strategic thinking as a kind of minimum competency requirement which even the lowest level student should attain. Specific, concrete knowledge may be seen as a further stage in a students' progress. In some settings, at least in Europe, broad erudition is sometimes viewed as the basic platform of future attainment, and therefore a minimum to be attained by all students in the course. At any rate, we should not be too quick to say that only the higher-order thinking skills should receive A-level letter grades without further investigation and asking teachers their preferences.

The Value of Content Scale

The second factor is the value of the content assessed. This factor was operationalized by developing an ordered scale of the perceived importance of successfully performing on a specific assessment task. The importance of successful task performance was rated in terms of its structural centrality to the subject-matter at a given level of instruction, and its relevance for further study of the subject and as a life skill. If an assessment task is rated as absolutely essential, successfully performing it is a prerequisite to a student's progress in understanding the discipline and to his/her real-world success. When teachers rate the importance of tasks which they use to assess their own students, their attention is focused on what they value about the content of the instructional targets for which they will hold students accountable. Figure 2 shows the four-point scale we used.

Insert Figure 2 here

Our experience is that many teachers think of virtually all assessed instructional content as "important by definition." That is, the subject-matter content of any task which they include in an assessment procedure is important to know, otherwise it would not be there. If nearly everything on a teacher's assessment instrument is seen as "important," then in order to discriminate among the tasks, it is necessary to add adverbs to an adjective-anchored scale creating anchors such as, "very important" and "very, very important." In some respects, this is similar to the relevance judgments made in Ebel's (1972) recommendation for standard-setting.

As with the thinking skills scale, it should not be assumed that teachers express preference for assigning the highest letter grades to the most highly rated content. Some teachers may express a preference for requiring knowledge of the most important (valued) content as a basic or minimum outcome which all students should attain. If so, they would require its knowledge for awarding of lower grades. From this perspective, higher grades would be based on knowing somewhat less central content. A second approach might be to require knowledge of less important content for lower letter grades and knowledge of more important content for higher grades.

We would agree that only important content should appear in an assessment instrument, especially if it is to be used for summative student evaluation. Thus, when a teacher associates less important content with lower grades, we will not assume that trivial task performance or subject-matter knowledge is required for the various letter grades. Rather, we shall assume that teachers are distinguishing among various degrees of relevant and valued content and task performance.

Combining the Two Factors

The model we investigated requires combining the thinking skills and importance factors into a single letter grade scale. This can be done by crossing the factors in a two-way table. Each assessment task receives two ratings (one for each factor) and thus can be doubly classified in the table. Figure 3 shows the two-way table into which the assessment tasks would be classified.

Insert Figure 3 here

Once all of the tasks have been classified in the table, the table can be used to identify those tasks which are especially suited to assess "A", "B", "C", etc. students. The general procedure is to use information obtained from the table to organize the tasks into subtests or testlets (cf., Wainer and Kiely, 1987; Wainer and Lewis, 1990). Each testlet assesses one level of the letter grade scale. Thus, a teacher establishes separate testlets for A-level, B-level, and so on. Within each testlet, we set a passing score derived from students' performance and on the need to maintain an ordered scale. Once a passing score is set, we scored each testlet 1 if the student passed and 0 of the student did not pass. This results in a pattern of 0s and 1s for each student that we could reference to a qualitative letter grading scale. For example, if there were four testlets, the pattern 1111 means passing all four testlets and a teacher would award the letter grade A; the pattern 1110 means passing the lower level testlets but not the highest level and the teacher would award B, and so on. The pattern 0000 is an "F".

Teacher-Specific Grading Models

In the preceding discussion, we pointed out that teachers may associate higher letter grades (i.e., A and B) either with (a) higher-level or (b) lower-level thinking skills. They may also associate higher letter grades either with (c) more-valued or (d) less-valued task content. Combining these four possibilities with the two-way classification shown in Figure 3, leads to four possible teacher-specific grading models. We label these alpha, beta, gamma, and delta and depict them in Figure 4.

Insert Figure 4 here

In the figure, each panel shows a two-way table for classifying assessment tasks. The arrowhead shows the direction of the higher letter grades. The four models are distinguished according to which combination of thinking skills and content importance teachers associate with higher letter grades. It is important to recognize that in these approaches, knowledge and skills are cumulatively reflected in the higher letter grades.

The alpha model for grading is one in which a teacher believes that in order for a student to attain the minimum satisfactory grade (i.e., a "D"), that student should understand the most important content at a concrete declarative level, perhaps with command of a few simple procedures. An "A" student, on the other hand, goes well beyond this to attain, not only knowledge of the most important content and basic procedures of the discipline, but also has command of some of the lesser important content while exhibiting systemic and innovative thinking.

In the beta model for grading a teacher believes that to attain the minimum satisfactory grade, a student should understand the less important content at a concrete declarative level, perhaps with a command of a few simple procedures. An "A" student has command of the less important content and the basic procedures, but also demonstrates knowledge of the most important content and can exhibit high levels of systemic and innovative thinking.

Teachers using the gamma model for grading believe that to attain the minimum satisfactory grade, a student must understand the most important content and use it at high levels of systemic and innovative thinking. "A" students also have command of these higher level thinking skills and content, but they also have command of a body of lesser important knowledge and basic procedures for using it in declarative and concrete ways. This model may prevail in school subjects where strategic skills are fundamental and, therefore, are required for the lowest grade, while detailed knowledge is considered a less important goal for everyone to attain.

The delta model for grading implies a teacher believes that students deserve the minimum satisfactory grade if they have knowledge of the less important content but can use it in higher levels of thinking. The "A" students go beyond this to also know the most important content and in a least declarative and concrete ways. This model may prevail in school subjects where strategic skills, self-knowledge, and minimal knowledge of content are required of all.

Two points should be made about these models at this juncture. The first is that the models stand as hypotheses and require research to verify that they are used and applicable to various subjects and teachers. The second is that these are models for summative letter grade assignment, not teaching or learning strategies. A teacher may easily adopt any of the approaches implied by the models for introducing content or teaching thinking skills. A teacher may start a lesson with important facts (alpha), with certain superficial observations (beta), with an important and complex problem (gamma), or with a quasi-important but complex problem (delta). Different teaching approaches for different lessons and subject-matters making teaching more diversified and flexible. The major question here, however, is whether the four models exist or can exist in classroom grading practice.

Incorporating a Skill vs Content Emphasis Into the Models

The four models in Figure 4 do not show how teachers may assign grades when they value thinking skills more highly than content knowledge or vice versa. The discussion to this point has implied that the two dimensions are cf equal value to all teachers. This is clearly not the case, however. Some teachers put more value and emphasis on one or the other dimension in their letter grade assignment. Thus, students' demonstrations of thinking skills rather than content knowledge may influence the grade assigned by a particular teacher.

Figure 5 shows how tasks classified in our 4 x 4 scheme might be organized into grade-level testlets when different preferences are expressed by teachers for the two dimensions. The heavy line shows the boundaries for the classified tasks that would constitute each grade-assigning testlet. These configurations are only illustrative: Others are also possible depending on a teacher's idiosyncratic values for the content and skills exhibited by the assessment tasks.

Insert Figure 5 here

Empirical Tryout

Subject-Matter of Application

The subject-matter for which the grading models were studied was introductory statistics taught as a beginning graduate course in a school of education. The course was a standard pre-calculus course, requiring only elementary algebra as a prerequisite. The course covered descriptive statistics and an elementary introduction to inferential statistics. The type of summative assessment that was the focus of our applications was the midterm examination.

Study I: Post Hoc Composite Judgments

Data The first study was a post hoc analysis of part of an existing statistics course's midterm examination. This was a paper and pencil test containing 48 response-choice items, each item being scored dichotomously. The test was not constructed with the grading models or the two-way classification of Figure 3 in mind. An existing open-ended part of the examination was not used in Study I. The test was a secure test that one instructor administered in the course during one semester in each of three consecutive years. Over the three administrations, item data were available for 120 students.

Method Five statistics instructors (not including the course instructor) independently classified the test's items according to the rating scales shown in Figures 1 and 2. For each item, the frequency distributions of the thinking skills and importance ratings were obtained. The median rating for each item on each dimension was calculated. (There was, of course, much variation between raters. Coefficient alpha was .42 and .43, respectively, for the two scales.) These median ratings were then used to classify each item in a two-way table as per Figure 3. The results are shown in Panel A of Table 1.

Insert Table 1 here

As can be seen in Table 1, none of the 48 items had median ratings that classified them into the "not important" or the "systemic, innovative" levels. This means that testlets could not be assembled from the existing collection of items to distinguish "A" from "B" students. Therefore, we combined the SI and HP rows and the SI and NI columns of the table. The results are shown in Panel B.

The marginal patterns of the item difficulties (p-values) tend to support the hierarchical nature of the thinking skills (but the p-values do not prove it), with items assessing lower-order skills being slightly easier than those assessing higher-order skills. Further, items assessing the least important content were easier than items assessing most importantly rated content. This pattern suggests that an alpha type of grading model might underlie this set of items, and that a slight preference for thinking skills over content knowledge might be operating among the instructors as well. The heavy line in the body of Table 1 separates the cells that contain the items for each grade-level testlet. Thus, the cells with AB in them contain the items that constitute the AB-level testlet, those with C in them the C-level testlet, and D the D-level testlet. There were 11 D-level items, 26 C-level items and 11 AB-level items. The items were rearranged into these testlets, and each student received a score on each testlet.

The next step was to set a passing score for each testlet. Passing scores may be set in many ways (see Jaeger (1989) for a review). However, here we decided to set it arbitrarily at 75% since teachers typically do not use anything but an arbitrarily set passing score for classroom assessments. Each student's testlet score was compared to the passing score. A testlet score of 1 was assigned if the student's score was equal to or larger than the passing score, and 0 was assigned otherwise.

Results Table 2 shows the testlet passing pattern frequency distribution. The reproducability coefficient was .80. (Reproducability is in reference to reproducing the testlet passing pattern from the letter grade.) This coefficient is probably too low to justify using the particular testlets as an ordered scale for grading. There are too many "irregular" or nonscalable patterns. Reproducability is affected by the marginal distributions of examinee scores and the testlet passing score. For example, if we set the passing score at 80% for each testlet, the reproducability coefficient drops to .68. Perhaps we could improve reproducability by setting different passing scores for each testlet, but this is not likely to be an acceptable strategy for teachers to use. We should note, however, that a logic somewhat different than "completely scalable" could be adopted. This logic, which we shall describe later in the paper, accepts "unscalable" student performance patterns, and still uses the thinking skill focus of grades.

Insert Table 2 Here

Study II: Specially Designed Assessment Instruments

Study I used an existing test that was not designed to fit the scheme we proposed. As a result, there were no test items that could be reliably classified into the higher levels of the thinking scale. (Some instructors did, however, classify items into the highest category but the median classification was used in Study I.) Next we explored whether we could build midterm assessment tasks especially to fit the various categories of our scheme and which were acceptable to individual instructors teaching the course. We explored the grading model in this context.

Data In the second study we solicited the cooperation of two instructors teaching different sections of the same course. The instructors taught in quite different styles and the sequence of content was different so that by the time of the midterm examination only about 60% of the content topics were taught to all students in both sections. (Nevertheless, by the *end* of the course all students were taught all of the topics.) We shall distinguish the two sections by referring to them as 1 and 2, and the instructors as 1 and 2, respectively.

Section 1 had 22 students who completed the midterm examination. Section 2 had 34 students, but 26 of them completed the examination. The remainder of the students were excused from the midterm examination to attend a conference and were given a substitute examination outside of the study. In both sections students were told they would use the instructors' normal testing and grading schemes.

Method Using the instructors' syllabi, their past examinations, and our own item pool, we drafted separate midterm examinations tailored to each section's content coverage. Both response-choice and short-answer tasks were drafted with care being taken to include sufficient items of both formats at the "systemic, innovative" levels. The items for each examination were reviewed by the respective instructors for content accuracy, wording, and relevance to the course. After this review they were revised. Each instructor then independently rated the items for thinking skill level and importance, following the same procedure as was used in Study I.

During the course of this rating several items were revised or eliminated because the instructors viewed them as either too difficult or requiring a much higher level of student thinking than that for which they were willing to hold students accountable. Both instructors expressed concern that their own student ratings would drop if the examinations were too difficult or extended students' efforts beyond the material on which they lectured directly. The result was that there were far fewer higher level items than we intended.

Results Table 3 shows the results of the classification of tasks for each instructor. As can be seen, instructors were reluctant to classify any items as "not important", further supporting our contention that teachers perceive the content of virtually all tasks on their assessments as important. Instructor 2, however, classified no items as "very, very important", whereas Instructor 1, classified 42% of Section 1's midterm items as "very, very important".

Insert Table 3 Here

We already mentioned the instructors' reluctance to put challenging tasks on their midterm examinations. This is further verified by their classifications: Instructor 1 classified only 7 (out of 52) tasks as assessing "systemic, innovative" thinking, while Instructor 2 saw none of the items on the Section 2 midterm as assessing the highest thinking skill level.

Conversations with the instructors during the test development phase indicated they conceptualized different grading models[2]. We deduced that Instructor 1 preferred an alpha model (see Figure 4) with a skill preference. We deduced that Instructor 2 preferred a beta model with equal weightings of skill and content. As a result, testlets were formed to reflect these models. The heavy lines in Table 3 show the boundaries for the tasks that comprise each testlet. Since Instructor 2 classified the examination items into very few categories, it was not possible to develop a testlet that distinguished A and B letter grades. Thus, three testlets, instead of four, were created for this instructor. (In practice, we would work with the instructor to produce acceptable tasks so testlets to distinguish "A" and "B" could be produced.)

Following the same student scoring procedure as in Study 1, we graded the students' performance on each testlet, determined passing scores, and assigned grades. In this study, we

[2] Figures 4 and 5 are outcomes of our investigation, and they were not available at the time we developed the two examinations. Thus, the instructors did not see them.

used a passing standard of 70% (again, arbitrarily set) for each testlet in each section. The results are shown in Table 4.

Insert Table 4 Here

There is some improvement in reproducability over the situation in Study I. However, reproducability is not exceptionally high. As we mentioned, reproducability is a function of the testlet passing scores and the marginal distribution of students' performance on the tests. We calculated the reproducability coefficients for each section's examination for several passing scores. These are shown in Figure 6. As can be seen, reproducability can be as high as 95% if passing standards for the Section 1 examination are lowered to 60% for each testlet. A comparable gain for the Section 2 examination was not obtained. This lack of gain in reproducability appears not to be simply a function of the number of grade levels (testlets) used. If we collapse the four testlets from Section 1 into three, we could improve reproducability by lowering the passing score. Perhaps, it is the number and thinking levels of the items that permit more discrimination among students, and hence result in testlets which have more scalable patterns.

Insert Figure 6 Here

In the discussion below, we argue that scalability in the reproducability sense is not the sole criterion for the usefulness of this grading approach. Other logic may be compelling to continue the approach, but scalability is one measure of the quality and interpretability of the resulting grades.

Discussion and Conclusions

Our investigation indicates that is feasible to design a grading scale that is linked closely to a teacher's qualitative emphasis on content importance and thinking skills. The four models of grade assignment we describe provide a framework to identify the value system teachers implicitly use when conceptualizing the meaning of grades. A classroom assessment instrument can be designed to be consistent with that framework. The process we described is the following:

1. Using an ordered thinking skills by importance grid, rate and classify assessment tasks on both dimensions.
2. Decide on a grading model and content vs. skill preference value structure which is consistent with one's educational values and the course objectives.
3. Use the results from Steps 1 and 2 to construct testlets to reflect letter grades of A, B, C, and so on.
4. Set a performance standard (passing score) for each testlet.
5. Assign letter grades to students based on their pattern of passing the testlets.

Requirement of Challenging Assessment Tasks

Our studies showed us that it was difficult for instructors to set tasks that challenge students to think beyond what they were explicitly taught in the classroom. It was equally difficult for instructors to see the content of their assessment tasks as unimportant. However, to assess higher order thinking skills requires that students be presented with somewhat novel situations which may not have straightforward solutions. Unless instructional time is spent on rehearsing strategies for solving unfamiliar problems, there may be little hope that assessment instruments and grades will focus on higher order thinking skills.

Is Scalability a Requirement?

The grade assignment approach we explored was designed to produce testlets for each letter grade category. This implies a scalable pattern of student performance: An "A" student will pass the A, B, C, and D testlets; a "B" student will pass the B, C, and D testlets; and so on. Full scalability in the sense of no students obtaining an irregular pattern is an ideal but perhaps an unnecessary criterion to attain.

If a teacher produces A, B, C, and D level testlets which are qualitatively different by following the procedure we outline in this paper, the most important goal has already been obtained. That is, the goal of the proposed procedure is to focus attention on valued content and qualitatively different levels of student thinking. With some instruction and practice, teachers should be able to construct qualitatively different testlets and set passing scores for each.

A teacher could use a modified grade assignment rule. Instead of saying, "You must pass the D, C, and B level testlets in order to get a 'B'", the teacher's rule could be, "You must pass three out of the four tests to get a 'B'." Then, the teacher should make clear to students the testlets' qualitative differences. This strategy has several advantages:

(1) It maintains the major focus of the grading procedure on qualitatively differences in students' thinking.

(2) It maintains the students' focus on learning and using different kinds of thinking skills and reinforces the idea that not all content is equally important.

(3) It relieves the problems that might otherwise fall to the teacher who still must grade "irregular" student testlet patterns fairly.

Scalability or reproducability is certainly an ideal goal. Our assessment construction courses could focus on how to create assessment tasks and testlets to improve reproducability and, thereby, improve the meaningfulness of grades. Our traditional treatment of item properties could easily be reinterpreted to focus on how assessment task construction and their operating properties can enhance the meaningful scalability of students' assessment performance.

Making Grading Models Explicit

In Figures 4 and 5 we presented a framework that describes several possible grade assignment models that teachers may have internalized, but not made explicit. If teachers use such a framework to rationalize the way they assign grades, this may provide them with a coherent way to articulate their grading schemes. The AFT, NCME, and NEA (1990) Standards for Teacher Competence in Educational Assessment of Students states that:

> Teachers will understand and be able to articulate why the grades they assign are rational, justified, and fair, acknowledging that such grades reflect their preferences and judgements: (p. 4) (emphasis added)

Using one of the grading models (alpha, beta, gamma, delta) in these figures and being able to explain it is certainly consistent with the intent of the Standards. If a grading model is followed in a consistent way, then instructional emphasis, assessment instruments, and letter grade evaluations will be part of the seamless fabric of instruction, learning, and assessment.

Conclusions

We cannot recommend wholesale adoption of this approach to letter grade assignment without further exploration and research. The approach provides an interesting rationale, but beyond that it raises the issue of whether we need a theory of letter grade assignment. It seems to us that grading is currently done on a rather ad hoc basis that may not be consistent with instructional goals and educational values. A coherent theory of grading may be necessary. Our investigation raises several methodological issues also For example, what item analysis, task analysis, and cognitive analysis techniques are important for teachers to use if they seek to implement a coherent grading system that reflects thinking skills and their own values and preferences. An expansion of this investigation to other subject matters and levels of instruction will be required before we have solid answers to such questions.

References

American Federation of Teachers, National Council on Measurement in Education, and National Education Association (1990). Standards for teacher competence in educational assessment of students. Washington, DC: National Council on Measurement in Education.

Cox, R. C., and Graham, G. T. (1966). The development of a sequentially scaled achievement test. Journal of Educational Measurement, 3, 147-150.

Ebel, R. L. (1972). Essentials of educational measurement (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Frederiksen, J. R., and Collins, A. (1989). A systems approach to educational testing. Educational Researcher, 18, 27-32.

Guilford, J. P. (1950). Fundamental statistics in psychology and education (2nd ed.). New York: McGraw-Hill.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), Educational measurement (3rd ed.). New York: Macmillan, 485-514.

Marzano, R. J., Brant, R. S., Hughes, C. S., Jones, B. F., Presseisen, B. Z., Ranking, S. C., and Suhor, C. (1988). Dimensions of thinking: A framework for curriculum and instruction. Alexandria, VA: Association for Supervision and Curriculum Development.

Niemierko, B. (1990). Pomiar sprawdzajacy w dydaktyce: Teoria i zastosowania. (Criterion-referenced measurement in education). Warszawa: Panstwowe Wydawnictwo Nankowe.

Nitko, A. J. (1980). Distinguishing the many varieties of criterion-referenced tests. Review of Educational Research, 50, 461-485.

Resnick, L. B., and Resnick, D. P. (1991). Assessing the thinking curriculum: New tools for curriculum reform. In B. R. Gifford and M. C. O'Connor (Eds.), Future assessments:

Changing views of aptitude, achievement, and instruction. Boston: Kluwer Academic Publishers.

Shephard, L. A. (1989). Why we need better assessments. *Educational Leadership, 46*(7), 4-9.

Stiggins, R. J., Frisbie, D. A., and Griswold, P. A. (1989). Inside high school grading practices: Building a research agenda. *Educational Measurement: Issues and Practice, 8*(2), 5-14.

Wainer, H., and Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 195-201.

Wainer, H., and Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27,* 1-14.

Table 1. Composite classification of 48 midterm examination items from the thinking skills and content importance ratings of five instructors. (Study I)

A. Classification before combining rows and columns

		Importance rating					
		VV	V	SI	NI	k	p
Think-	SI	0	0	0	0	0	--
ing	HP	4	4	0	0	8	.69
Skills	LP	8	17	3	0	28	.74
Rating	CD	3	8	1	0	12	.79
	k	15	29	4	0	48	.74
	p	.84	.69	.77	--		

B. Classification after combining showing testlet boundaries

		Importance rating				
		VV	V	(SI,NI)	k	p
Thinking	(SI,HP)	4	4	0	8	.69
Skills	LP	8	17	3	28	.74
Rating	CD	3	8	1	12	.79
	k	15	29	4	48	.74
	p	.84	.69	.77		

C. Testlet lengths from Panel B

"A,B" testlet = 11 items, "C" testlet = 26 items, "D" testlet = 11 items

Table 2. Frequency distribution of the "regular" and "irregular" testlet passing patterns for the midterm examination. (Study I, passing standard = 75%

	Passing Patterns	Letter grade	Number of students	Cumulative frequency
Regular patterns	1 1 1	(A,B)	50	50
	1 1 0	C	16	66
	1 0 0	D	22	88
	0 0 0	F	8	96
Irregular patterns	1 0 1	C?	11	107
	0 1 1	C?	9	116
	0 0 1	D?	1	117
	0 1 0	D?	3	120

Reproducibility of regular patterns = .80

Table 3. Midterm examination results and instructors' classifications of their test items. (Study II)

A. Total score statistics for the two examinations

	Section 1 examination	Section 2 examination
Number of items: open-ended	33	21
five-choice	19	17
Number of students	22	26
mean	49.1	26.8
standard deviation	7.7	5.8
coefficient alpha	0.87	0.77

B. Instructors' classifications of items showing testlet boundaries

Instructor 1

	VV	V	SI	NI	k	p
SI	2	3	2	0	7	.60
HP	11	8	0	0	19	.73
LP	2	10	2	0	14	.75
CD	7	4	1	0	12	.88
k	22	25	5	0	52	.75
p	.75	.73	.84	--		

"A" testlet = 7 items, "B" testlet = 21 items, "C" testlet = 13 items, "D" testlet = 11 items

Instructor 2

	VV	V	SI	NI	k	p
SI	0	0	0	0	0	--
HP	0	7	5	0	12	.61
LP	0	2	16	1	19	.67
CD	0	0	7	0	7	.80
k	0	9	28	1	38	.68
p	--	.61	.67	.71		

"A,B" testlet = 7 items, "C" testlet = 7 items, "D" testlet - 24 items

Table 4. Frequency distribution of the "regular" and "irregular" testlet passing patterns for the midterm examination of two course sections. (Study II, passing standard = 70%)

	Passing pattern				Letter grade	Section 1 (alpha model)		Section 2 (beta model)	
						Testlet len.	Num. stud.	Testlet len.	Num. stud.
Regular patterns	1	1	1	1	A	7	6	--	--
	1	1	1		(A,B)	--	--	8	6
	1	1	1	0	B	21	4	--	--
	1	1	0	0	C	13	3	6	1
	1	0	0	0	D	11	4	24	6
	0	0	0	0	F	--	2	--	6
Irregular patterns							3		7
Totals						52	22	38	26
Reproducability of regular patterns							.86		.73

CD LEVEL. CONCRETE, DECLARATIVE LEVEL

The student has gained intuitive insight into basic statistical phenomena and is able to use this insight for recognizing simple graphical and tabular representation of phenomena. The student also can reconstruct verbal definitions of basic statistical concepts and follow the definitions in solving the simplest word problems.

The core thinking skills prevailing at this level of competence are (a) *focusing* on statistical phenomena and (b) *remembering* statistical terms and concepts.

LP LEVEL. LOWER PROCEDURAL LEVEL

The student can draw immediate implications of definitions of the main statistical concepts and apply the implications to simple numerical and word problems. The student has attained an ability to deal with statistical data whenever the operations remain closely tied to basic concepts and procedures demonstrated by the instructor or in typical textbook examples.

Besides focusing and remembering, the core thinking skills prevailing at this level of competence are (a) *information gathering* (i.e., selecting and bringing to consciousness the relevant formulae and data) and (b) *organizing* statistical information (i.e., rearranging it, classifying, grouping, and labeling) so as to easily and effectively use it in problem-solving.

HP LEVEL. HIGHER PROCEDURAL LEVEL

The student has conceived logical connections between basic statistical concepts. The student can use them to reorganize problems so they resemble a teacher or textbook model or to control the processing of data. The student is able to draw inferences using statistical knowledge that permits the student to solve more complex statistical problems.

Besides focusing, remembering, information gathering, and organizing, the core thinking skills prevailing at this level of competence are (a) *integrating* selected elements of statistical knowledge and skills (i.e., combining them and restructuring them in subsystems) and (b) *generating* new information using integrated knowledge and skills for problem solving and interpretation.

SI LEVEL. SYSTEMIC, INNOVATIVE LEVEL

The student is able to use abstract operations on statistical concepts. The student understands statistics in a mathematically correct way, with adequate flexibility and restraint, at a level appropriate for an introductory course. The student is able to solve novel or unfamiliar applied statistical problems for which a teacher or textbook did not present a model, but that the student may appropriately solve using the introductory statistical concepts the teacher taught.

Besides focusing, remembering, information-gathering, organizing, integrating, and generating, the core thinking skills prevailing at this level of competence are (a) using appropriate theoretical *analyses* of statistical formulae and phenomena (i.e., identifying their components, structures, attributes, and interrelations) and (b) *evaluating* statistical concepts and problem solutions for their reasonableness, accuracy, efficiency, and correctness.

Figure 1. Thinking skills scale for an introductory statistics course.

Directions Rate each assessment task's importance on the following scale:

4 -- absolutely essential, very very important
3 -- very important
2 -- somewhat important
1 -- not important

Use the following definition of importance:

> The importance of a task concerns its centrality to the structure of the subject-matter at a given level of instruction as well as its relevance for further study of the subject and as a life skill. Understanding important content and being able to use that understanding to do what the task requires, makes the student's performance on the task an indispensable prerequisite to his/her progress in understanding the discipline and to later success.

Figure 2. Scale for rating the importance of assessment tasks

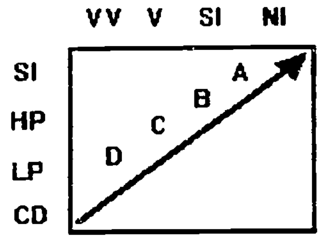
			Instructional importance of the task's content			
			Most			Least
			VV	V	SI	NI
Level of thinking in the task	Highest	SI				
		HP				
		LP				
	Lowest	CD				

Notes:

(1) SI = systemic, innovative level; HP = higher procedural level; LP = lower procedural level; CD = concrete, declarative level.

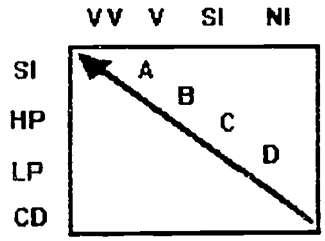
(2) VV = absolutely essential, very very important; V = very important; SI = somewhat important; NI = not important.

Figure 3. Two-way grid for classifying teachers' judgments of assessment tasks.



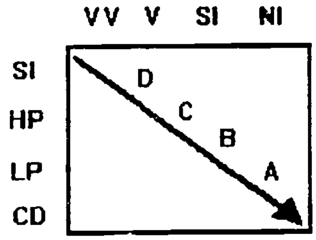
Alpha Model of Grade Assignment

The least achieving students understand absolutely important content at a concrete, declarative level and have command of a few simple procedures. The highest achieving students go well beyond this to attain, not only understanding of the most important content and basic procedures, but also command some of the lesser important content while exhibiting high levels of systemic and innovative thinking.



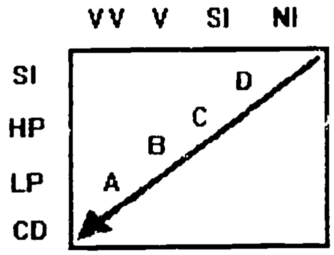
Beta Model of Grade Assignment

The least achieving students understand somewhat important content at a concrete, declarative level and have command of a few simple procedures. The highest achieving students go well beyond this to attain, not only understanding of the lesser important content and basic procedures, but also command some of the absolutely important content while exhibiting high levels of systemic and innovative thinking.



Gamma Model of Grade Assignment

The least achieving students understand absolutely important content at a very high procedural level and can use that content innovatively and creatively. The highest achieving students also have command of these higher level thinking skills and content, but they have in addition, command of a body of lesser important knowledge and the basic procedures for using it in declarative and concrete ways



Delta Model of Grade Assignment

The least achieving students understand content of lesser importance at a very high procedural level and can use that content innovatively and creatively. The highest achieving students also have command of these higher level thinking skills and less important content, but they have in addition, command of a body of the most important content and the basic procedures for using it in declarative and concrete ways

Figure 4. Four possible models that a teacher may use as a basis for assigning grades when considering thinking skills and content importance.

			Instructional importance of the task's content			
			Most			Least
			VV	V	SI	NI
Level of	Highest	SI	A	A	A	A
thinking		HP	B	B	B	A
in the		LP	C	C	C	B
task	Lowest	CD	D	D	D	C

A. Skill preference orientation to alpha grade assignment

			Instructional importance of the task's content			
			Most			Least
			VV	V	SI	NI
Level of	Highest	SI	B	B	A	A
thinking		HP	C	B	B	A
in the		LP	D	C	B	B
task	Lowest	CD	D	D	C	C

B. More or less balanced skill and content orientation to alpha grade assignment

			Instructional importance of the task's content			
			Most			Least
			VV	V	SI	NI
Level of	Highest	SI	A	A	A	A
thinking		HP	A	B	B	B
in the		LP	B	C	C	C
task	Lowest	CD	D	D	D	C

C. Content preference orientation to alpha grade assignment

Figure 5. Three different alpha grade assignment models which teachers might use depending on the values (weights) they place on the thinking skill vs. content dimensions of the assessment tasks.

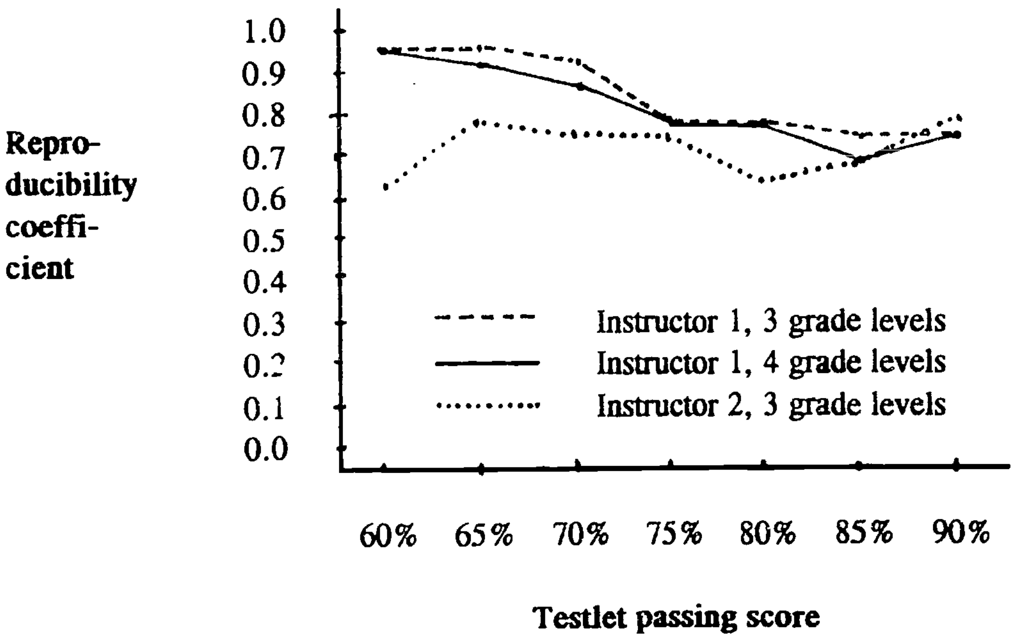


Figure 6. Reproducibility coefficients for the assessments in Study II when various testlet passing scores are set.