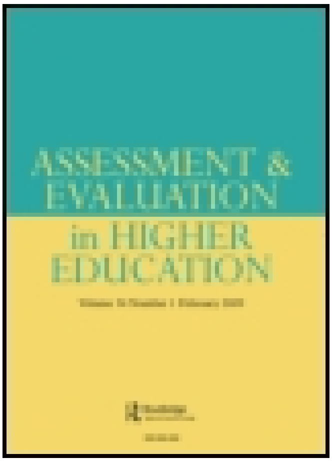

Assessment & Evaluation in Higher Education
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/caeh20

A new framework for grading

Kris H. Green [a] & Allen Emerson [a]
[a] St John Fisher College , USA
Published online: 19 Jun 2007.

Assessment & Evaluation in Higher Education
Vol. 32, No. 4, August 2007, pp. 495–511

A new framework for grading

Kris H. Green* and Allen Emerson
St John Fisher College, USA

Grading is one of the least liked, least understood and least considered aspects of teaching. After years of work, we have developed a grading system that is quite different from traditional and reformed approaches to grading and which meaningfully incorporates and integrates the collection of evidence, the evaluation of evidence, and the reporting of judgments about that evidence. This system satisfies the requirements of good grading system and answers many of the problems faced by more traditional methods by substantially changing the way in which grade information is aggregated, resulting in a final course grade that aligns qualitative evaluation with course learning objectives and carries direct qualitative meaning with respect to the course learning objectives.

Introduction

When we began requiring a significant amount of writing in our mathematics course for business students, we became increasingly dissatisfied with those traditional methods for grading student work with which we were so familiar, first as students and then as teachers. We had introduced writing as a way of bettering our students' understanding of how to apply mathematical concepts and procedures to real-life problem situations, all of which was a result of our having incorporated technology, in this case spreadsheets, into the course. Accordingly, we were able to provide our students with the opportunity of investigating more realistic, ill-defined problems than had been hitherto possible, since interpretation and contextualization of mathematical results could now replace memorization and manipulation of decontextualized mathematical procedures as the end-point of students' knowledge. In short, words replaced the familiar rows of mathematical symbols in student work and grading it became more a matter of evaluating processes rather than products. And the *sine non qua* of process, we came to believe, is revision.

Wiggins (1998, p. 7) pointed the way for us with his description that assessment should 'primarily educate and improve' rather than 'audit' student performance. This led us to the problem of how to provide effective feedback to the students to improve

*Corresponding author. St John Fisher College, Department of Mathematical and Computing Sciences, 3690 East Avenue, Rochester, NY 14618, USA. Email: kgreen@sjfc.edu

ISSN 0260-2938 (print)/ISSN 1469-297X (online)/07/040495–17

DOI: 10.1080/02602930600896571

both their thinking process and their use of mathematical tools for their revisions. The added complexity due to the more open-ended nature of the problems we had assigned meant that we could not just give students the 'correct' answers without undermining the legitimacy of the learning experience. After all, each student could potentially take a different route through the available procedures to develop his or her own individual solution path. Thus, we needed to incorporate feedback in way that would allow the students to re-engage the material in order to revise their approach both mathematically and analytically. But we were left with the question of how to provide this feedback in a way that maintained consistency from student to student but still addressed each student's work individually while, at the same time, not placing an undue burden on the instructor. Further, we wanted a system that could easily be adapted to give final course grades real meaning by rewarding student work in the areas of their best performance. In fact, Henderson *et al.* (2004, p. 165) point to a body of literature that suggests grading practices may have a larger influence on student behaviours than almost any other action an instructor can take. From our students' work we found we could not use a traditional grading method with its corrective solutions and clear distinctions between right and wrong answers. We needed a system that would allow us to evaluate work for process and understanding while at the same time reward more analytical and interpretive work. Unlike written work in some disciplines, student work on ill-defined problems is predicated upon correctly employing certain mathematical procedures, and the grading system must account for this as well as the structure of the argument and the style of its presentation. Our response to this was to develop a comprehensive grading system, which we call Categorical Objective Grading System (COGS), and which we believe can be adapted to any type of course material that involves students having to employ correct procedures within a problem context that requires argumentation, interpretation and persuasion.

In what follows, we begin by defining some basic grading and assessment terminology. Next, we present a model of what constitutes any good grading system. The majority of the paper then focuses on a description of COGS. We conclude by comparing COGS with one of the more popular grading systems in use today according to the grading model already developed.

Basic terms

Although many teachers use the terms *grading*, *evaluating*, and *assessing* interchangeably, each has a distinct meaning in educational practice. Following Wiggins (1998, esp. pp. 21–70, 241–288) we offer these definitions. By *assessing* we mean the collecting of student-generated data for the purpose of evaluating and grading it. Although much has been said in the literature regarding the nature of authentic assessments of student learning, we will not discuss this here. For our purposes, the result of assessing is a collection of student work samples in some form, whether tests, papers, projects, or presentations. *Evaluating* student work refers to the process of giving feedback in order to help the students improve their understanding of the course material. *Grading* is the process by which the work is assigned some code—usually A through F or a percentage—that represents the overall quality of the work. If your education was like ours, then you know that the distinction in the three terms is often blurred. Some teachers assume that the grade is a form of feedback itself. While it is possible for this coincidence to occur, it requires that the teacher carefully explain the grades in terms that relate to the objectives for the particular assessment. Furthermore, we suspect that once a grade is placed on student work and the work is returned to the student, the sometimes lengthy written feedback provided with the grade is often overlooked by the student, even assuming that such writing is legible. We will use the term *grading system* to refer to any systematic process that includes the three components above.

The term *aggregating* will be used to refer to the process by which a collection of grading symbols or evaluative feedback is turned into a single grading symbol. One of the most common methods for aggregating is to average the grades on individual papers. However, there is an additional hidden level of aggregation in most grading systems: the very process of putting a grade on a paper in the first place. For some systems this process of aggregation is the only grade given and the individual components from which it was derived are almost completely hidden, accessible only from any written comments that might have been provided on the work sample. We refer to such systems as *Level 1 systems*. A *Level 2 system* is one step beyond the Level 1 system in that it makes use of some sort of scoring sheet that (1) defines categories of achievement for demonstrating mastery of the material; (2) assigns scores to each category; and (3) aggregates these categorical scores into an overall score for the work sample. A *Level 3 system* is best exemplified by the use of rubrics for evaluating student work samples. Level 3 systems go one step beyond Level 2 systems by providing multilevel feedback in each scoring category, rather than only a summary grade for each category. The performance in all of the categories is then aggregated into a grade for the assignment.

All grading systems involve all three of the aforementioned components—assessing, evaluating and grading. Each, however, differs in the degree to which these components are transparent. The three grading systems presented above, which we will refer to collectively as *traditional systems*, are all similar in that each student work sample is marked with an aggregate grade. Later in the course, these separate grades are further aggregated to produce a course grade. In contrast, the system we are presenting here does not hide feedback behind an aggregate grade in the same way. In fact, no aggregation takes place in our system until specified points in the semester, usually at the midpoint and the end of the term. This change of timing for the aggregation helps to prevent a loss of information which can best be understood by tracking the way a traditional system produces a course grade. Traditionally, after students produce a work sample, the teacher has the opportunity, often unexercised, to provide an evaluation of the work through written comments or a rubric; this is qualitative feedback. The instructor then assigns a single grade, which is 'quantitative' information. (Even a letter grade represents quantitative information since it is a categorical ordinal variable.) At the end of the semester, all of these quantities, the grades, are aggregated in some way to produce a single, quantitative piece of information which is then translated into a letter grade that, supposedly, has qualitative meaning. For example, an 'A' is supposed to mean 'excellent work'. This mixing of data types leads to many of the frustrations commonly expressed by faculty.

The essential components of a good grading system

As an introduction to a model of a good grading system, consider the following group of students who have each earned an 80% average at the end of the semester. At most schools this percent translates into a B–, meaning slightly above average. Alfred completed most of the required course work nearly perfectly, only to botch a few assignments at the end. Bonnie did poorly at the beginning of the course then improved dramatically and carried nearly perfect work after the first exam of the course. Chandra handed in work that was so-so, not excelling in any of the course's objective areas, but not exactly failing any of them, either. At the end of the semester, each of these students needs to be assigned a grade. What would you really like to assign? To be sure, each of the students has, in an objective sense, an 80% average. Our concern is this: have all of these students earned a B–? Certainly we are not raising anything new here; indeed, Table 1 in Guskey (2002, p. 776) offers an even more extreme example of how traditional grading gives rise to this troubling issue.

We will define a good grading system to be one that satisfies the following eight criteria. It must be *reliable* across different graders. It must provide *consistency* in feedback to students. The grading marks and evaluation should be *transparent* so far as the course objectives are concerned. The system must allow for *differentiation* among different levels of achievement in order to avoid grade compression. It must be *stable* over time in order to promote validity and to discourage grade inflation. While all grading systems are subjective, a good system recognizes, acknowledges, and makes appropriate use of this *subjectivity*, rather than masking it or ignoring it. A good grading system must be *practical* in that it must not be too unwieldy or too difficult to implement. While all grading systems make use of aggregated data, a good grading system makes use of *meaningful aggregation* in order to prevent the loss of critical information about student performance. Wiggins (1998, pp. 251–252) and O'Connor (2002, chapter 6, pp. 140–157) argue that using the mean of several scores typically does not satisfy this criterion.

As we set forth these criteria for a good grading system, we must also address certain common problems. The number of layers of aggregation can lead to a *loss of information* in that the final grade has little to say qualitatively about student performance. Students are not measured by their achievement of certain standards and can, as a result, *survive by partial credit* as does Chandra in the above example. Moreover, the nature and rationale for grading are *misperceived* by the faculty, on the one hand, as a form of administrative monitoring of their teaching and by students, on the other, as being the focus of their learning. Grading systems suffer from *projection* issues whereby faculty view student submissions in the best possible light (see Henderson *et al.*, 2004, p. 167) by projecting their own teacherly understandings onto the student's work ('Ah, this must be what this student meant to say'). And most grading systems suffer from *misrepresentative aggregation* meaning they are unable to reward student progress during the semester, allowing poor work in the beginning to overshadow success at the end of a course (Bonnie in the above example) or allowing superior work at the beginning of the course to prop up poor performance at the end of the course (Alfred in the above example). We will argue that the grading system we are presenting fares better than more traditional systems when compared in terms of the essential components and the potential pitfalls.

Many teachers recognize these sorts of problems and modify one of the three systems described above in order to address such shortcomings. For example, one could drop grades, placing Alfred and Bonnie in a higher grade category, but not really affecting Chandra's grade. This method is relatively easy to implement and solves some of the problems above. It fails, however, to recognize that Bonnie has improved over the course as the material has (presumably) increased in difficulty. Alfred's performance has steadily declined, thus inviting the interpretation that he did well at the beginning because of his prior knowledge and not because of anything he learned during the course.

The essence of COGS

It is clear from current research that many discussions concerning the topic of improving grading practices focus on ways to shift from one level to another in the schema presented above, usually by making explicit the instructor's objectives on an assignment. While such improvements are useful, we feel that they still fall short of addressing the deficiencies. We believe what is needed is some way to systematically build grading out of the evaluation of student work so that it is transparently tied to the course objectives and actually reflects the students' capabilities. This process must be clearly communicated to the students, and they should be required to self-evaluate in order to take ownership of the system.

In developing COGS we have attempted to reflect recent research and understanding about how knowledge is organized and built upon in the mind. We believe that it reflects the learning objectives that are most important to the teacher, that it promotes consistency and allows for the differentiation of instruction. Most of all, we believe that it is adaptable to other courses that have different goals, different students, different instructors and different types of assessments.

COGS starts with the objectives of the course. These are then divided into three distinct categories that are evaluated in every assessment. Each of these categories is then differentiated into two levels of understanding/achievement. We typically refer to these levels as the *expected* level of knowledge and the *impressive* level of knowledge. We think of the expected level of knowledge as the basic level of understanding in a category that a student must achieve in order to pass the course satisfactorily and continue on to subsequent coursework. The impressive level of knowledge is the body of knowledge that helps differentiate between average and superior students. Table 1 shows examples of the three areas for quite different courses.

Table 1. Examples of three areas of expectations in different disciplines

Course	Grading area	Description
Mathematics service course for business majors	Mechanics and techniques	The basic mathematical definitions and computations of the course, as well as the computer techniques (Microsoft Excel, largely) needed for the course
	Analysis and reasoning	The planning of solutions to complex problems and the logical development of analyses for realistic business problems
	Communication and professionalism	The writing and the presentation of problem solutions, as well as the attitudes and behaviors of the students (attendance, working in groups, etc)
History course	Sources	The gathering, evaluation, and incorporation of source materials
	Analysis	The use of the evidence to present an argument that is logical
	Communication	The grammar, style and presentation of the argument
Physical science course for elementary education students	Engagement	Attendance and effort in the course as well as working in groups and completing required reading assignments
	Exploration	Student knowledge of and use of the process skills of science (control of variables, measuring, hypothesis testing, and reflection)
	Understanding	The conceptual knowledge of the material and the ability to apply this knowledge to new situations

Phase 1: evaluating a student work sample

Each assignment can then be scored using a matrix that is divided into the three objective categories for the course, each at two performance levels. The matrix consists of a 3-by-2 grid with a checklist of between, say, 3 and 10 items in each block of the grid. Each item in the checklist is a single criterion describing something that should be recognizable in the work sample. Thus, the instructor merely makes a binary choice for each item in each checklist. Each criterion is focused and specific to the assignment. Each is phrased in a positive way (e.g., 'Work shows understanding that the median is more stable than the mean') to promote student learning. The work of a particular student can then be compared to the matrix, and the criteria that the work meets can be checked off. Thus, the matrix provides both a grade and an effective means of feedback that students can then examine to determine exactly what they did or did not accomplish. After checking off criteria a student meets, the instructor can determine, using her professional judgment, if the work sample, overall, falls into the expected (E) or impressive (I) level of performance *with respect to each of the three categories*.

Phase 2: aggregating grades in each category

The instructor then tracks the levels of each student in all three areas on all assignments. At then end of the grading period a letter grade can be assigned to the student, based on aggregating the student's work in each category across all assignments first. If the student's work in a category is mostly at the expected level of knowledge, the aggregated grade is 'E'. If the work is mostly at the impressive level of knowledge, the aggregated grade is 'I'. This aggregation can be accomplished in a variety of ways, the simplest of which is to assign points to the 'E' and 'I' grades in a category and use a numerical method to average the grades.

Phase 3: determining course grade

These three aggregate grades are then combined to assign an overall grade in such a way as to allow an instructor to see, from a glance at one symbol, where the student falls in the continuum of learning. The key to assigning letter grades is to realize that this assignment is combinatorial in nature. Overall, a student will be in one of five combinations of aggregate grades:

1. At least one of the three areas is not even at the expected level.
2. All three areas are at the expected level, none are at the impressive level (EEE).
3. Two areas are at the expected level, one is at the impressive level (IEE).
4. One area is at the expected level, two are at the impressive level (IIE).
5. All three are at the impressive level (III).

The limited number of final results is the key to assigning and interpreting the grades. The III level is assigned a letter grade of A; the IIE level is assigned a grade of B+; the IEE level earns a B; and the EEE level is a C. Students failing to achieve expected level in all three areas receive a D or an F. One can also modify this scheme and assign the EEE a grade of B– or C or whatever. Simply set the bar so that each additional I grade raises the letter grade by a fixed number of steps on the grade ladder, regardless of whether the ladder is F, D, C, B, A or F, D–, D, D+, C–, etc. Notice that this scheme instantly allows an instructor to translate from a final letter grade back to a good picture of student performance. Unfortunately, as with any aggregation process, some information is lost; given a final letter grade, we can recover the number of areas in which a student achieved 'I' level work, but we cannot recover *in which* areas this was achieved. This is the reason that the three areas must be chosen carefully so that they apply to all work equally and they are all equally important for the course.

A brief example of COGS

To understand the system better, consider the following assignment (other examples of assignments and grading matrices are available online—see the web address given in Endnote 1).[1] In a basic statistics course, students are asked to evaluate the placement of two managers, based on (a) verbal descriptions of the managers and their motivations; and (b) the salary structure at the companies. Students are expected to produce statistical summaries of the four possible company placements and are expected to produce graphical representations as well. Their report is expected in the form of a memo.

For this assignment, one might easily break the grading matrix into the following major aspects:

1. *Mechanics and techniques (MT):* Does the report include correct statistical measures? Are the graphical representations correct? Are they properly embedded in the document? Are the outliers in the data identified? Is it clear from the report that the numbers and graphs have meaning to the report writer?
2. *Application and reasoning (AR):* From the statistical and graphical results, does the report draw conclusions? Are the conclusions based on multiple pieces of evidence? Are the conclusions drawn from the evidence valid?
3. *Communication and professionalism (CP):* Does the report begin with a short overview of the problem? Is the explanation clear and concise? Is the explanation and discussion integrated with the supporting evidence? Does the report have an executive summary and a clear statement of recommendations?

Once these broad expectations are determined, the grading matrix is constructed by separating these criteria into two levels: those you expect each and every student to get (or they will not be considered to have passed the assignment) and those that require a higher level of inferencing, more advanced writing techniques, or sophisticated mathematical tools/computer skills. The matrix is then a checklist of these items, separated into two columns (expected and impressive) and three rows (MT, AR, CP). Copies of the matrix are made for each student, and the instructor simply checks off the criteria that are present in the student's work. We use a marking system that indicates complete success with an 'X' and partial success with a slash. We do not typically make any comments directly on the student work. There is discussion in class about what each criterion means. Students must then analyze their own work for these expectations and modify their response in order to improve the work. If the checklist is modified to have two check boxes for each criteria, then it is easy to track student work on both the original submission and the revision.

At the end, students do not receive a grade on the assignment. These are meant to be formative feedback to help them improve, so we do give them an overall impression of their work in each of the three areas. This is indicated by the marks shown in Table 2.

At the end of the semester, after seeing the entire body of work from the student (including revisions of their assignments) we then aggregate the student grades across categories, producing an overall grade (from the table above). These aggregate grades then produce a letter grade for homework, as described below.

Table 2. Grading marks and descriptions for a sample assignment in basic statistics

Mark	Description
I	The student work is overall impressive in this category and has all criteria checked off
I–	The student work is relatively impressive, but is missing a few criteria
E+	The student has succeeded at all the expected criteria and has achieved a few of the impressive-level criteria in this category
E	The student has succeeded on all the expected criteria
E–	The student is missing a few of the expected-level criteria
0	The student is missing almost all of the expected-level criteria

Flexibility in the system

The description of the system above gives the basics. But for the system to qualify under the criterion of *practical* it must be adaptable to a variety of course content, teaching styles and assessment tools.

Some may feel that having only two possible levels of success in each of the three categories is simplistic in that it reduces the number of possible overall grades to five while, at the same time, requiring too much distinction in borderline cases. However, it is easy to adjust the checklists above to include partial success for almost making a level (see Table 2). At the end of the course these modified aggregate grades can be adjusted using a sliding scale. A base letter grade can be established from the main categories (E and I) as above, and then the pluses and minuses can shift the grade up or down the grade ladder. In practice, we have found that this works best if you carefully design the 'grade ladder' with duplicates of certain key grades in order to avoid too much overlap in the grades. Such a grade ladder is shown in Table 2, along with indications of the aggregate marks needed to achieve a particular letter grade. For example, if a student's work aggregates to 'E–', 'I–', 'E' in the three categories (regardless of which of the three got which mark), the base grade would be determined from EEI and translated to a B as a first step. The two minuses would then shift this grade down two steps on the scale. If the scale were to look like the one below, the resulting final grade would be C+.

COGS also includes an easy way to share responsibility for grading decisions with the students. Whenever grades are to be aggregated, an instructor can require that students do the work of aggregating their category grades by constructing an argument that tracks their growth. Students are then rewarded for reflecting on their work and can use this reflection to prepare for end-of-course assessments like final projects or exams. Many experts (see O'Connor, 2002, chapter 8, pp. 175–184, for example)

Table 3. Possible correspondence between traditional letter grades and COGS

F	D–	D	D+	C–	C	C	C	C+	B–	B	B	B+	B+	A–	A–	A	A
						EEE				IEE			IIE				III

advocate that students must be included in decisions about grades in order to promote fairness and transparency in grading systems.

COGS relies primarily on regular assessment through project-like activities and assignments. We also include tests and quizzes—traditionally considered to be more objective assessments—by assigning letter grades to these and then aggregating this data with the data from the grading system above. Thus, a student receives an aggregated letter grade on the assignments, a letter grade for quizzes, a letter grade on the midterm, and a letter grade on the final exam (which could also be scored according to our system, but this can become difficult in practice, since timed exams offer little opportunity for reflection and polish). These four grades can then be aggregated. In practice, we tend to average the assignments, midterm and final, and then use the quizzes to either move the grade up or down one step on the ladder, depending on performance. For example, a student with a 'B' average on the assignments and tests with an A average for the quizzes would received a course grade of B+; if the same student were to have a C quiz average, the final course grade would be a B–.

Additionally, the targeted nature of the feedback on the matrices provides students with enough information to modify their work for resubmission without the teacher having to solve the work or correct the vast part of it for them. This allows students to extend and deepen their knowledge through encouraging them to think about past material as new material is being dealt with. This works especially well with the three-tiered system for marking criteria in the grading matrix in that it allows an instructor to mark partial success for items where the student has not shown enough to prove that he or she understands or can do the task described in the criteria. This gives an instructor an easy way to avoid some of the problems with personal projection described in Henderson *et al.* (2004, p. 167). At the same time, it allows for more opportunity for student reflection, encouraging them to revise their work and improve their grades. Since assignments during the course are truly formative in nature, we designed the system to allow for growth opportunities. The system described in this paper naturally includes opportunities for revision through the grading criteria, which do not give away the answers to the assignment, but rather are designed to direct the students' attention as to how the work they have done does or does not meet the criteria.

The system is also inherently flexible from a student's perspective. The nature of the grading system and the specificity of the objectives for the assignments allow students to select how they want to approach the course. They can focus on assessment areas in which they are strong to establish a base grade and then use the system to help them target other assessment areas for improvement. But more than students developing a 'course strategy'—and our experience indicates that many do—is that a flexible system capable of handling significant revision helps students develop their own 'voice,' to take an independent path through a problem, one that is not quite like anyone else's, including the teacher's—all of which can be captured in the matrix either by checked off criteria in one of the three categories or even by a brief note as to what a student has done interesting that is not one of the criteria.

Expanded theoretical framework for COGS

While we have already placed this grading system into a theoretical framework above and before we compare it with other systems, we offer this expanded theoretical framework in order to help you understand the system through a variety of theories of learning. Our hope is that this will empower you to adapt COGS to your own needs and those of your students. The first question, of course, is why we should develop new grading systems at all, since one could argue that these are structural design questions about the way a course is graded overall, rather than direct changes in classroom teaching techniques. One answer to this is provided by research in critical thinking. Paul and Eilder (2000, section 4, p. 4) describe five 'structures and tactics' instructors can use to facilitate thinking. The last three of these are the:

- requirements students must meet;
- grading policies in the course; and
- performance profiles that correlate with grade levels.

In general, the guidebook advocates that rather than defining grade categories by percentage points from averages, grade categories should be defined verbally, describing the characteristics of a student whose work demonstrates 'A' or 'B' level thinking. Our system directly answers this charge, since each final letter grade can be reverse engineered to determine in how many of the objective areas the student performed well. While the system cannot recover all the information from the single letter grade, the loss of information about student performance is much lower in our grading system than in any other that we have seen. It also pushes the approach to grading much deeper than that advocated by the Foundation for Critical Thinking in that instructors can define three categories of objectives, rather than a single overall category. The single-category approach mixes many performance indicators and often runs the risk of becoming too subjective, since it is, in essence, applying a traditional-style rubric with one performance category to the entire course. This is poor practice, especially when compared with criteria for making rubrics useful, as defined by Wiggins (1998, p. 184): 'The best rubrics are those that … do not combine independent criteria in one rubric'.

Wiggins (1998, p. 12) also explains that assessment systems should 'use grades that stand for something clear, stable, and valid'. In the grading system presented in this paper, all assessment areas are equivalent; it is not a partial credit system. This means that 'I' level work in one assessment category will not 'bring up' insufficient work in another category. This forces students to improve an entire category in order to improve their grade, rather than nibbling at many different criteria. Moreover, this maintains the stability of the grades by making them clear and easy to interpret.

COGS is easily adapted to any set of objectives, and once these are determined and communicated, students know what is really important in the grading process. In a review of research literature regarding effective techniques to support learning, Marzano *et al*. (Marzano *et al*., 2001, pp. 92–102) discuss the importance of setting objectives and providing feedback to students regarding how their work aligns with the course objectives. In COGS, once the matrices are developed instructors can provide feedback more easily than with Level 1 or Level 2 grading systems and more consistently across assignments than with the Level 3 systems described above. In such systems, instructors have a great deal of flexibility in both the content and depth of the feedback they provide. While some may argue that this is good, we are all aware that students compare their work. If we want to send a consistent message about what constitutes successful problem solving or other work, the feedback must be consistent. That being said, COGS also allows for extra comments to be made in order to address any anomalous student work. However, these anomalies are not usually included in determining the grade. Rather, instructors can use them as feedback for the revision process and for revising the grading matrices since such anomalies point out criteria that are not defined clearly enough or are missing altogether.

In order to understand how to analyze course and assignment objectives to accommodate this system, we have found the revised version of Bloom's taxonomy (Anderson & Krathwohl, 2001, p. 107) most helpful. The general idea of the revised Bloom's taxonomy is to categorize learning objectives into two dimensions: the type of knowledge (factual, conceptual, procedural or metacognitive) and the cognitive process depth (remember, understand, apply, analyze, evaluate or create). Our system is similar in nature, having one dimension for the three disjoint types of knowledge in the course and one dimension for the two-level cognitive depth. In comparing the cognitive depths between the two approaches, a rough guideline is to group the lower cognitive dimensions of the revised taxonomy's cognitive domain as 'expected' knowledge and the upper levels as 'impressive' knowledge, as shown in Table 4.

This compression of the cognitive dimension into two categories is consistent with the revised taxonomy in that the specific types of knowledge are often associated with particular cognitive levels from the *remember*, *understand* and *apply* categories. For example, we often assess whether students can apply procedural knowledge, but we rarely assess whether students remember procedural knowledge. Thus, the lower three cognitive depth categories can be grouped together quite nicely.

We could also compare these performance levels in the assessment areas with the three levels of knowledge objectives described by Wiggins and McTighe (1998, pp. 9–10). 'Enduring understandings' refer to a few central ideas that frame the entire

Table 4. Correlation of COGS with Bloom's taxonomy

COGS level	Bloom's taxonomy cognitive levels
Expected level	Remember, understand, apply
Impressive level	Analyze, evaluate, create

course. The next largest category of knowledge is referred to as 'important to know'. These latter refer to are critical facts and skills, the lack of which would indicate failure to adequately address the course content. The largest category is referred to as 'worth being familiar with'. Our system is simpler, having only two levels. Our 'expected level' encompasses the bottom two levels of the Wiggins and McTighe model, while the 'impressive level' equates nicely with the top level.

We conclude this discussion of the theory underlying our system with some of the fine print. Comments on the matrix checklist must be positively stated in order to reward student work and to help them reflect on and recognize what they have done. The comments must be specific to the assignment, not generic to all assignments, in order to provide that they provide adequate feedback to support the revision process. To encourage reflection and revision of work, the comments should not 'spill the beans' but should give clues as to how best to proceed. Finally, the assignments we use this system with are not summative in nature. They are formative assessments, and we expect students to learn from doing the assignment rather than prove that they already know the material.

Comparisons between COGS and traditional rubric systems

Much has been written about how to address the shortcoming of Level 1 grading systems by modifying them toward Level 2 systems. Many have written of how to address the shortcomings of grading systems like the Level 1 systems by modifying them into Level 2 systems. Since Level 2 systems are easily modified into Level 3 systems, we will focus our comparison on the differences between COGS and Level 3 systems with their use of traditional rubrics. One of the primary differences is, of course, the timing of the aggregation. In Level 3 systems, the first aggregation occurs at the assignment level, while COGS delays the first aggregation until more evidence about student achievement has accumulated to allow for a more complete picture of a student's abilities. Thus, our system also does not conflate the natures of qualitative and quantitative data. Information in the matrices is conserved in its descriptive *categorical* form and not converted into an opaque numeric or letter grade (hence, the 'C' in COGS). Rather, all categories of a student's work are considered when a single letter grade is assigned to it. This provides gives the student with specific feedback as to areas of strength and weakness. By improving in these areas, the student can perform better on the objective measures obtained from quizzes and exams. All of these scores can then be combined to provide a more complete picture of a student's development. In doing so, COGS preserves the integrity and transparency of the course objectives throughout the entire course.

A second major difference between COGS this system and Level 3 systems is that the areas of assessment in Level 3 systems are typically different from assignment to assignment. With COGS, the same assessment areas are used throughout the course. This allows for students to continue working and improving in each area, since the requirements are consistent throughout the course and, moreover, consistent with course objectives. In our experience, attempts to enlist uniform categories from traditional rubrics across the progression of the course led to such generic feedback on revisions that it was difficult for students to know what specifically they could do to improve a particular assessment. COGS, on the other hand, does provide for specific feedback as well as for quality grading.

Notice also that by having only two levels of achievement, COGS is much simpler than a Level 3 system with its multiple levels of achievement that are often vaguely differentiated (does X most of the time, does X some of the time, etc). This vagueness in the differences between levels introduces unwanted subjectivity into Level 3 systems. The major sources of subjectivity in COGS come from making the decisions about the placement and content of each of the criteria in the checklist of the matrices and in determining a score for each category on an assessment. If these criteria are made specific enough, the first source of subjectivity can be further reduced, and, since each criterion is one of several that are used to determine success in an assessment category, the impact of an instructor making a mistake on one criterion is minimized. Further, since COGS encourages revision as well, students can take the opportunity to improve their handling of the material; thus, the formative aspect of the system works to our advantage when it is time for grading, and the projection problem is mitigated: an instructor can simply mark a criterion specially (by circling it, for example) in order to indicate 'I think you've got the idea, but you need to make it clearer'. Tables 5 and 6 give a complete comparison between COGS and Level 3 systems.

In conclusion, COGS provides answers to many of the problems associated with traditional grading systems. It exemplifies the notion that a grading system should do more than assign grades; it should embody the whole process of collecting evidence, providing feedback, and summarizing performance. While the initial development may take more time than with Level 1 or Level 2 systems, it is comparable to the time required to develop the rubrics for a Level 3 system, the kind many educational researchers support. The targeted nature of the feedback gives students the support they need to improve and succeed in specific ways for each assignment under easily identifiable categories that remain constant throughout all assignments. Further, COGS maintains a connection between grades and learning objectives that eases the burden on instructors seeking to find a way to improve overall assessment for accreditation purposes in an era of increasingly standards-based assessment. Standards-based assessment confronts us from many sides, but most notably from accrediting bodies such as NCATE and Middle States. These organizations are asking us to do something many of us are not used to doing: show that our students learn what we claim to be teaching them. A common theme among these accrediting bodies is that we cannot simply use grades as a measure of student learning. This is primarily due to the nature of determining those grades through inappropriate aggregation of student performance data. Such methods cannot show direct links between the learning objectives and student achievement. COGS, however, uses a method of aggregation that preserves information about student achievement relative to course objectives that is transparent and that can be convincingly documented. Furthermore, by targeting

Table 5. Comparison between COGS and Level 3 grading systems relative to the criteria for a good grading system

Criterion	Level 3 Leveled rubrics of performance indicators	COGS 3-tiered, 2-level matrices of criteria
Reliability	Performance indicators not specified clearly enough to ensure reliability	Each criterion is specific and binary, providing more reliability
Consistency	Performance indicators not specified clearly enough to ensure consistency	Each student receives identical treatment on identical criteria
Transparency	Because objective categories are not necessarily consistent from assignment to assignment, these may or may not directly relate to overall course objectives	Three components of each grading matrix are directly derived from the course objectives; levels of performance translate directly into course grades measured relative to these objectives
Differentiated	Multiple levels of performance indicators for each category ensure variety of possible outcomes	Standards are set up front; all assignments are evaluated relative to these standards
Stability	Vague differences between performance levels can translate into changes in interpretation over time	Reliability and consistency all but guarantee stability over time
Subjectivity	To the degree with which criteria are specified, subjectivity is reduced	Subjective decisions are made on a fine enough scale so as to average out
Practicality	Very easy to implement after rubrics are developed; many resources and examples are available.	Long development time for matrices, but easy to use once developed; these matrices are intended for use with complex assignments, the evaluation of which take more time in any grading system.
Aggregation	Occurs at the assignment level as well as across assignments	Occurs only at key points and is by category rather than by assignment

three main areas of objectives, instructors are encouraged to focus on their true learning objectives, rather than leaving such matters to take care of themselves. By linking qualitative meaning to course grades in various disciplines, it may be possible to integrate theory and practice into a more viable framework for measuring student achievement and teaching effectiveness for accreditation purposes than other more complex and less practical approaches.

Note

1. More information about the system as it relates to a mathematics for business course, including example assignments, sample work, and same grading matrices can be found online at the address http://keep2.sjfc.edu/faculty/kgreen/DataAM_Web/index.htm. Of particular interest to readers of this paper will be the materials under 'Instructor resources'.

Table 6. Comparison between COGS and Level 3 grading systems relative to the common pitfalls in grading systems

Problem	Level 3 Leveled rubrics of performance indicators	COGS 3-tiered, 2-level matrices of criteria
Loss of information	Aggregation hides all specific information about student achievement relative to objectives	Aggregated grade identifies levels of performance in all areas, but cannot show which areas
Partial credit	Because of assignment-level aggregation, superior work in one category tends to prop up weaker work in other categories	Students must perform consistently at a high level in a category across assignments in order to receive a higher grade
Perception	Vague nature of the feedback makes it difficult for students to see the value of the grading system *vis a vis* improving their performance; performance categories are tied to learning objectives, improving faculty perception of the value of the rubrics	Since grades are given only at specific points rather than on individual assignments, students tend not to focus on the grade but rather monitor their improvement; direct connections between objectives and grading ensure that faculty perceive the value of the grading system
Projection	Specific nature of criteria for success usually requires students to show what they know	Can only occur with individual items in the checklist, which can be made specific enough to eliminate projection
Misrepresentative aggregation	Aggregation across assignments suffers from all potential problems described in Guskey (2002)	Depends on the specific method of aggregation across categories

Notes on contributors

Kris Green has conducted research in relativity theory as well as the incorporation of writing in mathematics courses. He is the director of the Science Scholars Program at St John Fisher College and teaches courses in mathematics, physics, educational technology and freshman composition. While he holds a Ph.D. in applied mathematics from the University of Arizona, his wife, Brenda, and many others have awarded him an honorary doctoral degree in *Star Wars* studies.

Allen Emerson has been a war gamer, an econometrician, and a management consultant. He teaches in the Department of Mathematical and Computing Sciences at St John Fisher College and has been active in the development of the curriculum for the mathematical sciences and technology integration major and in uses of educational technology.

References

Anderson, L. & Krathwohl, D. (Eds) (2001) *A taxonomy for learning, teaching and assessing: a revision of Bloom's taxonomy of educational objectives* (New York, Longman).

Guskey, T. (2002) Computerized gradebooks and the myth of objectivity, *Phi Delta Kappan,* June, 775–780.

Henderson, C., Yerushalmi, E., Kuo, V. Heller, P. & Heller, K. (2004) Grading student problem solutions: the challenge of sending a consistent message, *American Journal of Physics,* 72(2), 164–169.

Marzano, R., Pickering, D. & Pollock, J. (2001) *Classroom instruction that works* (Alexandria, VA, Association for Supervision and Curriculum Development).

O'Connor, K. (2002) *How to grade for learning: linking standards to grades* (Glenview, IL, Skylight Professional Development).

Paul, R. & Eilder, L. (2000) *Critical thinking handbook: basic theory and instructional structures* (Dillon Beach, CA, Foundation for Critical Thinking).

Wiggins, G. (1998) *Educative assessment: designing assessments to inform and improve student performance* (San Francisco, Josey-Bass).

Wiggins, G. & McTighe, J. (1998) *Understanding by design* (Columbus, OH, Merrill-Prentice Hall).