TEACHERS' EVALUATIONS OF STUDENT WORK

James V. Mead[1]

Nine-year-old Susan waits patiently in her seat. The teacher walks around the room giving out the subtraction worksheets. The teacher stops at Susie's desk, hands over the paper with a slight smile, and then continues distributing them. Eagerly Susan searches for the grade. Her heart sinks as she makes out a hastily formed letter D. She scans the paper looking for clues to tell her what went wrong. Ms. Smith is at the front of the class signaling the start of the lesson. Susan puts her paper in the folder deciding this is one work sample that will not go home. The thought that Ms. Smith "hates me" crosses her mind while she struggles to concentrate on what her teacher is saying.

Evans's study (cited in Simon and Ballanca, 1976) found a consistent stream of survey research, dating from Starch and Elliot's National Education Survey in 1912, that claimed teachers had arbitrary grading procedures. Therefore Susan's teacher, Ms. Smith, could have a strange reason why Susan deserved a D. Where would Ms. Smith learn about grading? The literature gives her little guidance. Ms. Smith's teacher education program probably paid little attention to grading. Her colleagues, principal, and district policies either may suggest that assigning grades to individual papers is a scaled-down version of assigning course grades, like her own high school or college, or may advise against giving grades. It is important that we distinguish our interest in assigning grades to one piece of work from global grades given to students for courses. While they are related, this paper focuses on the task of grading individual work samples.

Stiggins, Frisbie, and Griswold (1989) report no empirical research on grading worthy of inclusion in the 1986 edition of *What Works* from the Department of Education. Nitho's (1989) editorial to a special issue of *Educational Measurement* devoted to grading supports the opinion of Stiggins et al. and describes the research base as "thin." Our review of recent empirical work show how thin. There are two empirical studies. Our findings coincide with Stiggins et al., who report 12 out of 15 veteran high school teachers favored grading that rewards student effort. Agnew (1985) in a survey study of assigning general course grades notes that student effort and improvement are significant considerations. Agnew also reports that teachers of low-ability secondary students emphasize effort more than teachers of high-ability groups.

This state of affairs exists because teacher grading falls down a crack between two well-established research literatures. Stiggins, Conklin, and Bridgeford (1986) reviewed the two bodies of literature on testing and teacher decision-making. Their review ascribes importance to teacher

---

[1]James V. Mead, doctoral candidate in educational administration at Michigan State University, is the archivist for the National Canter for Research on Teacher Learning. The author recognizes the special contribution of Mary Kennedy in preparing this paper and also wishes to expresses his appreciation to Deborah Ball and Bill McDiarmid for their helpful comments and assistance in preparing this paper.

assessment but also suggests why teacher grading might be neglected. The first body of literature—the decision-making literature—builds no explicit connection between assigning grades and other pedagogical decisions which do get discussed. The subcategories of teacher decisions that interest researchers focus first on teacher planning. A second category looks at teacher thinking while engaged in teaching. The third category deals with teacher theories and beliefs (Borko, Cone, Atwood Russo, & Shavelson, 1979; Clark & Peterson, 1986; Clark & Yinger, 1979). Clark and Peterson observe that researcher ideas drive teacher thinking research and not empirically derived categorizations of teacher practice.

The second body—testing literature—describes a teacher-initiated evaluation as "spontaneous performance assessment" (Stiggins & Bridgeford, 1985, p. 273). This jargon conveys the value placed by that community on assessment that "arises spontaneously from the naturally occurring classroom"; they are not "systematically planned and designed" (p. 273), as are structured performance tests. The educational measurement community fixes its sights on documenting student achievement with batteries of tests aimed at public accountability (Stiggins et al., 1986).

We found a striking parallel between a historical theme that recurs in the literature and constant references by the interviewees about their distaste and discomfort at being asked to grade. Several works (for example, Simon & Ballanca, 1976; Terwilligar, 1971, 1989) express a dissatisfaction with teacher grading procedures. The reasons for that discomfort vary. Authors about to advocate their own opinion (read favored grading system) cite some lack of objectivity on the teachers' part. Other authors, mostly from the early sixties onward, see grading work as inconsistent with progressive or scientific teaching visions, methods, and objectives. Terwilligar (1989) speaks for many informants in our study when he states, "Assigning grades is undoubtedly one of the most distasteful aspects of teaching." He also supplies the attitude many elementary teachers silently endorse—"It is a necessary evil that has little to do with the task of teaching" (p. 15).

This paper examines the criteria teachers describe when assigning grades to individual pieces of mathematics work. We see grading as a visible mark of a teacher's evaluation of student work. It penetrates a small part of what Weinshank (1980) describes as a professional mystique in how teachers evaluate students. We supply what Stiggins, Frisbie, and Griswold (1989) describe as missing—"an analysis of the underlying assumptions and philosophies teachers use in the grading process" (p. 6). By presenting such an analysis we hope that teacher educators consider providing a sustained treatment of grading practices and their rationale.

Source of Data

The data come from the Teacher Education and Learning to Teach Study of the National Center for Research on Teacher Education (NCRTE).[2] We chose one question: "What grade would you give this paper and why?" The question subset formed part of a larger structured exercise (Section C). This exercise investigated responses to a series of connected tasks on the topics of subtraction and slopes. We analyzed 226 responses from 90 informants to the grading question from all 11 investigated by the NCRTE. Those sites ranged from preservice teacher candidates, novice teachers in schools but still training, and experienced teachers.

In Section C, elementary informants were shown a subtraction worksheet completed by Susan[3] (see Figure 1). The work sample showed the answers clearly written but with some obvious errors. Secondary informants were given two sheets of graph paper with the problems and solutions, again neatly drawn, but were told this was Lynn's incomplete homework (see Figure 2). Using the word "homework" did produce an important variation between elementary and secondary teacher responses. Our interest lies in why teachers give grades and this was not affected by the variation in the task. Both groups of teachers were then asked the following questions:

1. What do you think is going on here with Susan (or Lynn)? What do you think she understands? Why? What do you think she doesn't understand?

2. Okay, imagine that Susan (or Lynn) is a pupil of yours. How would you respond to this paper? After the baseline interviews, informants were asked how they would respond if Susan (or Lynn) were a lower or higher achieving student.

3. What grade would you give this paper? Why? If the person resists the idea of grades: Would you mark this paper in any way? Then give the statement of school policy that requires teachers to assign grades as a final probe. After the baseline interviews, informants were asked directly how they decided the student had tried if they offered some opinion about effort.

While Susan's and Lynn's papers represent different topics, the work samples have important similarities. Both papers show the students in the early phases of grappling with important new topics in school mathematics. In the subtraction problems (Figure 1), Susan consistently takes the difference between the two numbers. However, sometimes she takes the top number from the bottom, sometimes the bottom number from the top. Sometimes she uses a "regrouping" or "borrowing"

[2]The precursor to the National Center for Research on Teacher Learning.

[3]Pseudonyms are used for students and teachers in the study.

procedure, yet in other sums she fails to. It is not transparent, based simply on the visual evidence, why she changes strategy.

The informants are free to provide a range of possible explanations. Four of the five calculations Susan gets right show her using a regrouping technique successfully. Did Susan simply not notice she should regroup in the first two sums, for example? Were the sums Susan got right the ones she copied from a friend? What does an explanation to justify an evaluative grade on this work look like?

Lynn (Figure 2) uses one form of the slope equation in questions 7, 8, 9, 10, and 11, namely:

$$y = mx + b$$

Lynn unfailingly constructs the above equation correctly with the supplied values when she writes it out. This equation sometimes helps her draw the graphs asked for in the first part of the exercise. She encounters difficulty when the second part of the exercise asks her to express the equation in a different form:

$$Ax + By = C$$

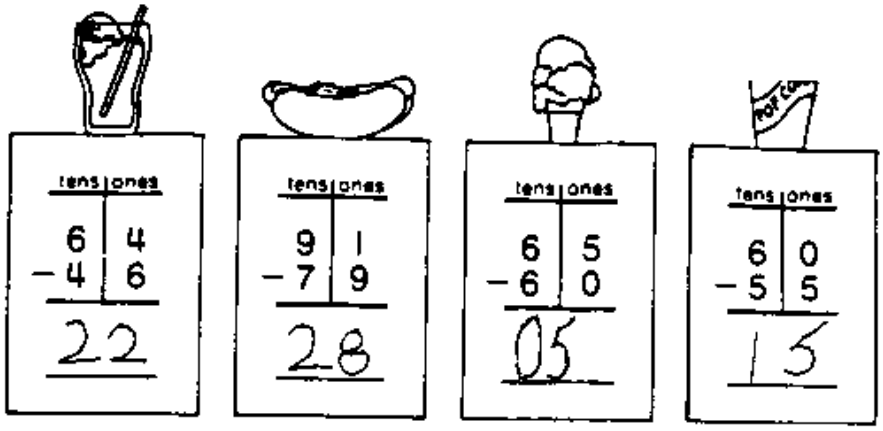
Lynn manages to manipulate the equation in Questions 7 and 8 but then fails in all other questions. The graph on Question 7 even has what looks suggestively like a better attempt erased. In drawing the graphs, Lynn's best attempts are in Questions 9 and 11. Question 12 is missing from the paper. Question 13 is attempted but incomplete and 14 has a question mark by it. As they did with Susan's work, the teachers are invited to speculate on what sense they can make out of this work sample.

Discussion and Presentation of Data

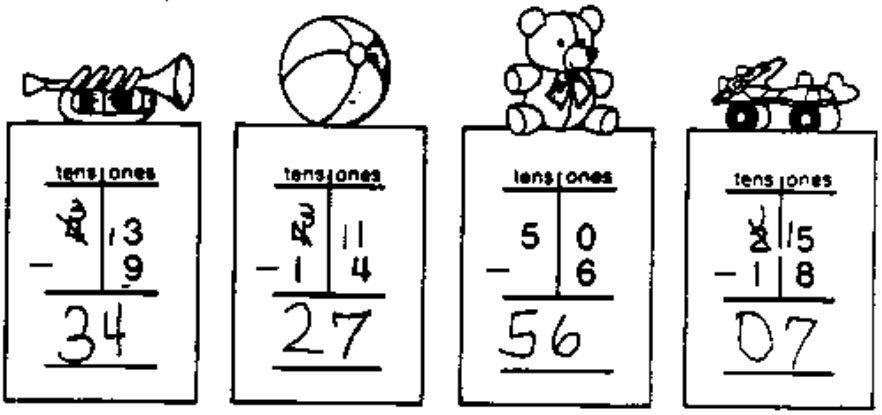
Table 1 shows the grades assigned to these two work samples at three different points in time: at the beginning of the program (BL), at the end of the program (EP), and a year after program during independent teaching (IT). The number choosing a grade in the total column suggests a normal distribution of the elementary subtraction work or the secondary graph exercise. Most teachers, with some notable exceptions, judge the work samples to be C, D, or F grade work. The near normal distribution could suggest that letter grades are distributed randomly. However, the central tendency also gives an appearance of a consensus about the appropriate grade.

name Susan

Downer Street School had a fair.
How much food was not sold?



How many prizes were left?



How much money does each child have left?

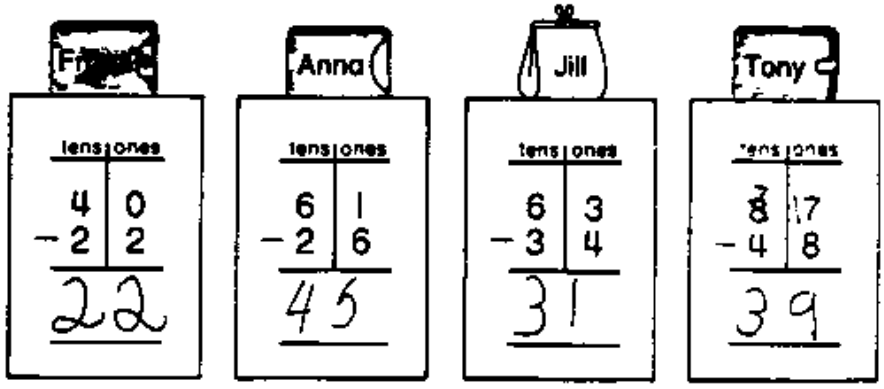


Figure 1. Susan's completed subtraction worksheet.

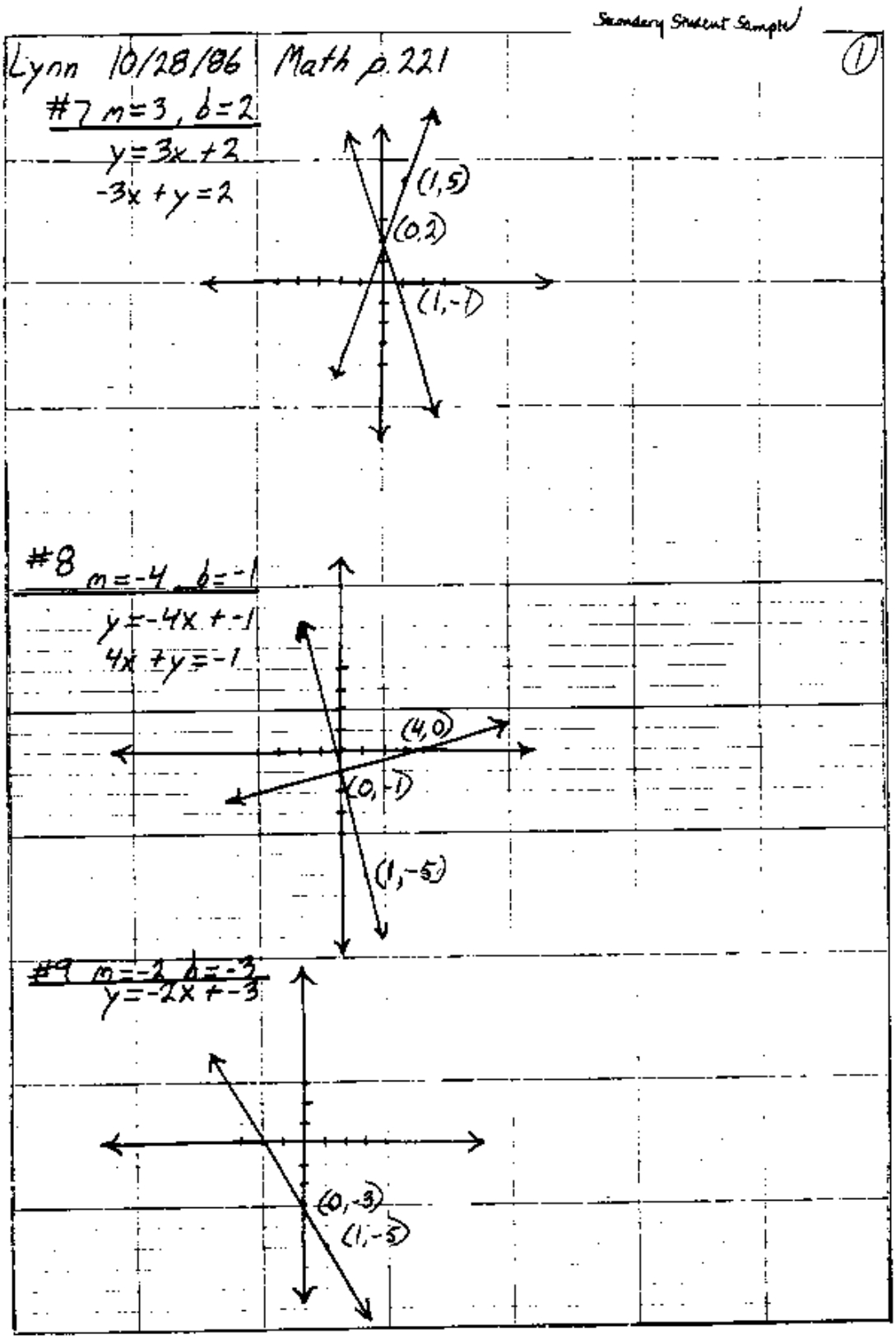


Figure 2. Lynn's incomplete homework.

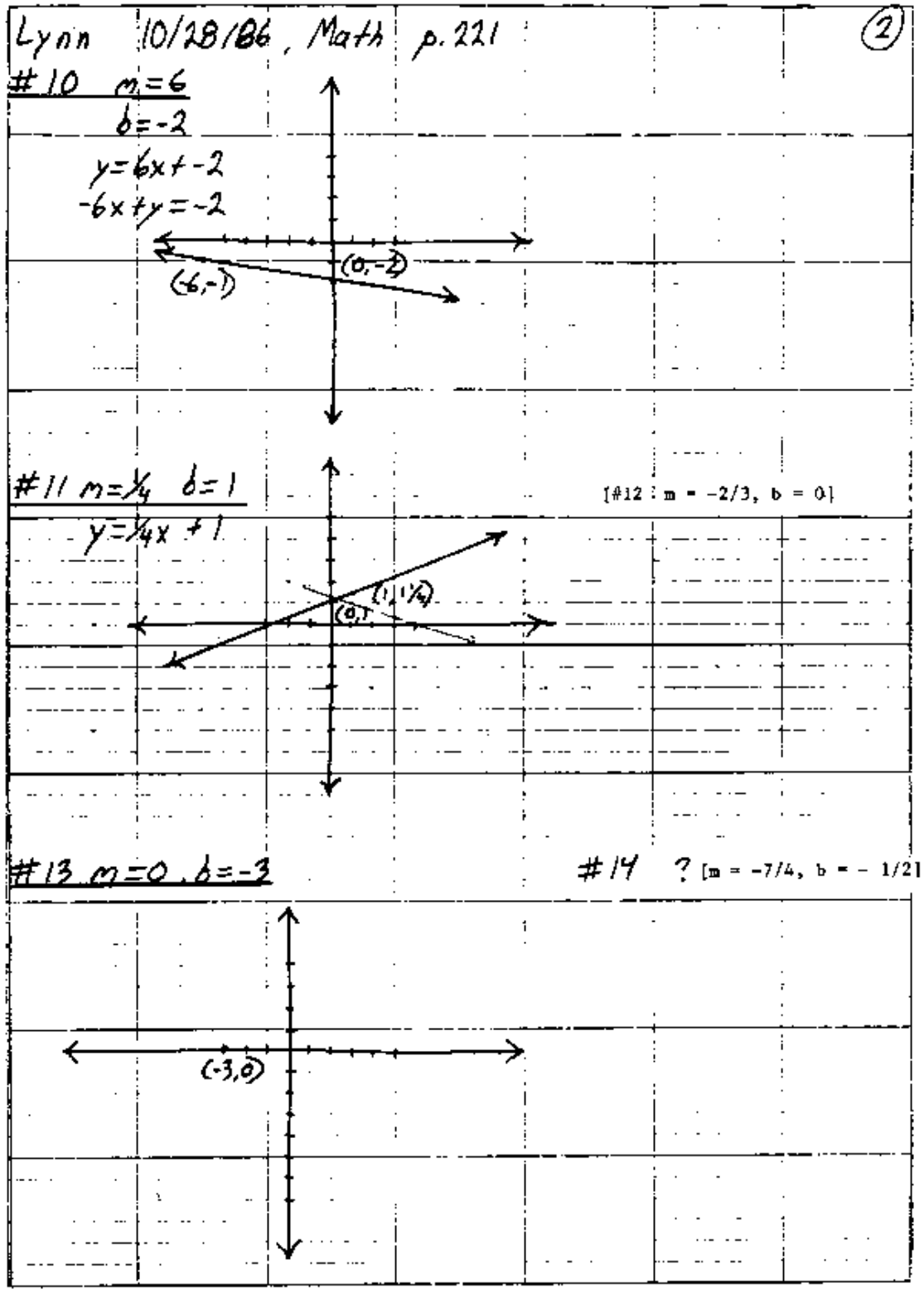


Figure 2. (continued).

Table 1

Distribution Grades in C10 Data

	Elementary				Secondary Math			
	BL[a]	EP[b]	IT[c]	Total	BL	EP	IT	Total
A[d]	2	1	0	3	0	0	0	0
B	3	0	1	4	5	2	0	7
C	8	6	1	15	5	7	1	13
D	13	5	4	22	6	4	1	11
F	9	3	7	19	2	0	0	2

Note. The following responses were excluded in the table:

38 middle-of-program instances
25 instances that are not elementary or secondary-math teachers.

Other Marking schemes:	Elem.	Sec.
Pass/Fail type	2	0
Satifactory/Unsatisfactory	3	0
Check system	9	2
Refused to grade	4	4

[a]BL = Beginning of program

[b]EP = End of program

[c]IT = A year after the program instruction during independent teaching.

[d]Each grade in the table represents multiple variations grouped together, for example A = A+, A, A-, A/B, B/A, A-/B+, or B+/A-.

Table 2

Distribution of Grades and Summary of Criteria for One Preservice Elementary Site

Name	Baseline	Middle of Program	End of Program	Independent Teaching
Letty	C	C		
	right procedure does not understand borrowing	for effort		
Lara	C-	F	F	
	type of work practice or test	got over half wrong	school district policy	
Lucille	F	D		E
	most wrong	most wrong		other student's performance
Leslie	Check minus		F	C
	depends on timing of work		based on student understanding	timing of work and grade on a curve
Lori	E	E		F
	missed 50% of the problems	missed over 50% of the problems		conditional: is it test or practice?
Lee	B	B		F
	partial understanding	most appear right		she does not understand
Louise	D	C	60-69%	
	to encourage she got some right	understands subtracting but not borrowing	other student's performance	
Lisa	U	Between U & S	50%	
	conditional on if this were a first attempt	not wanting to discourage	based on the number wrong	

Table 3

Objective Versus Qualitative (Inferential) Grading Styles

	Elem. *N* = 143	Sec. Math *N* = 58
Objective count only	46	16
Qualitative grading and no counting	32	25
Mixture of both counting and inference	9	9
TOTALS	87	50

The center of the distribution, or modal grade for elementary teachers, was a D grade. A skewed distribution pattern of the elementary teachers may be a function of the task the student was assigned. Simple subtraction calculations are amenable to a quick overview and evaluation. Secondary teachers faced the visually more demanding task of manipulating equations and graphing. Secondary teachers were less likely to assign an F to the work than their elementary colleagues. The center of the distribution for the secondary teachers clustered around the C grade.

The note for Table 1 shows the number of times alternative grading systems, other than letter grades, were proposed. This table suggests a strong preference for letter grades and not another evaluation scheme. Even those that refuse to grade or use another scheme assume that assigning a grade means putting the A through D and F symbols on the paper.

Many interviewees reluctantly graded the samples but only four elementary and four secondary teachers refused to grade the work. Four elementary and one secondary teacher judged the level of understanding and required the student to do the work again. The other refusals were based on a need for more information, a view that grading was destructive to learning, and a mathematics version of the writing process approach. For example, Frank, an induction program teacher, claimed math is a continuous process with work like Susan's counting as a "draft." Susan would be allowed to "edit" this work several times before she "published" it. Here, editing meant having to correct and to work on further examples, with publishing being the submission of the work for teacher evaluation. Concern for what grading represents within their role as educators disquieted many elementary teachers.

Elementary teachers showed more variation in the marking schemes, as reflected by the note for Table 1. The nonstandard check system allowed elementary teachers especially to avoid assigning letter grades. Assigning a check, check-plus, or check-minus counted as less "valuative" to them and simply a recognition of that work's completion. Daphne, from the New Jersey Elementary Alternate

Route Program, anticipated the question in the third and final interview.

> I wouldn't give it a grade, but if I had to give it a grade . . . I would give it a check. That means that she attempted to do it. . . . You don't want to devastate this kid when they are learning a new problem. They are trying. You don't really want to give them a grade.

Many elementary teachers deemphasized or postponed assigning a grade and suggested strategies and marking schemes to do so. These include allowing students to make corrections several times until it is all correct and then assign a grade. Some teachers wanted to talk to the student and not make the marks on the paper. "See me" is a substitute for grading. Other teachers wanted to circle correct or incorrect calculations. A couple of elementary teachers substitute a smiley or sad face to communicate messages to their students.

It proved impossible from these data to see any clear pattern suggesting that elementary or secondary math novice teachers change their grading because of program participation. The grade distributions across responses appear random at each time point with roughly comparable midpoints across time. It also was impossible to predict whether one informant applied the same criteria consistently to the grade assigned over the different time points. Table 2 gives an example of the grades assigned by one group of preservice elementary teachers, showing the variation both among informants and in one informant, over time.

The wide variation in grades assigned and the lack of apparent change over time raises some interesting questions. What is the basis on which teachers make grading decisions? For instance, are some dimensions of student work more important to them than others? How do they hope to influence students with these grades? The remainder of this paper examines these questions.

Table 3 summarizes three main strategies respondents used as a basis for assigning grades. An "objective count" referred to in Table 3 means the informants graded by counting the number of subtraction sums or parts of the graphing exercises that were correct. For example, Michelle, a secondary math teacher candidate in a preservice program, claimed that "it was straightforward, so many out of so many. Say that each problem was worth two points, one for the equation and one for the graph; then it'd be 2, 4, 6, 7, 8 . . . [pause] maybe 9, out of 14?"

Qualitative grading meant the informant inferred something about student effort or understanding from the work. Linda, an elementary teacher candidate, suggests several qualitative inferences she can make from the work in front of her.

> Because I can see that she knows what she's doing . . . it's just a matter of her doing it to all of them. She understands regrouping, but maybe somewhere along the

line . . . she had a bad day that day and only felt like doing four [laughs]. . . .

Table 3 shows that most teachers favor either objective counting criteria or qualitative criteria and not a mixture of both methods.

More interesting than the mechanics and distribution of the grades were the wide variety of considerations that lie behind a single grade. The following quotes from elementary informants illustrate the diversity of explanations for the modal grade of D. These explanations could serve as interesting stimuli for discussion and analysis in a math methods class.

Ginger, a preservice elementary teacher at the end of the program, wants to follow the district grading policy at first. Freed by the interviewer (Int.) from policy constraints, Ginger chose a general class performance criterion to decide the D grade. She used the rest of the class as a measuring stick to which she can hold Susan's performance. This measuring-stick strategy is used by other teachers. Ginger illustrates a concern teachers have with this strategy regarding negative or positive feedback to Susan. If Susan does poorly, they worry whether giving negative feedback like this to Susan may not be a pedagogically smart thing to do. If Susan is like everybody else or better in the class at this point then Ginger would like to give this positive feedback to Susan. What is the status of negative feedback and how should teachers handle these situations?

Ginger: Depends on the system. When the school system gives me a set percentage and says, "This is an A, this is a B, C, D, E, F," I would have to do it that way. I'm not going to deviate from it because it's a district policy or a school policy; it's not my decision to change that. Um . . . I don't know. If it had to be given a letter grade, which would be really hard. If they just say it has to be given a grade, a "satisfactory" or something would be easier.

Int.: Let's assume, just for the example here, that you had to give it a letter grade, and you didn't have specific guidelines from the district to follow.

Ginger: What I'd probably do is compare my class's papers and then I'd try to base a grade on that. I would have a very difficult time failing a child based on this because it shows some understanding, I think. I'm hoping that she did it herself, those few she got right. I wouldn't want to fail her, and then again, I can't give her an A or B on it. I'd have a real hard time. . . . If I had my way, I hope I could give her a C, but I'd probably compare it to others. I would see how many others in the class got [them] right. If the class as a whole only got 4 or 5 right, then she did pretty well in comparison to the rest of the class. If the whole class got 12 while she only got 3, then she's not doing very well. I'd have to grade it perhaps give the work a D, as much as I'd hate to.

Fay, an induction program elementary teacher, used effort to decide the grade. Here is a variation on the negative feedback puzzle. Fay observed that Susan has many sums "wrong," yet feels Susan's effort should not go unrecognized. Her decision, faced with this dilemma, is to give a low but passing grade:

Fay: I would give her credit for attempting the problems.

Int.: So, what grade would you give the work?

Fay: I think I would give her a D. I cannot fail a student for trying.

Int.: Okay, so you would give her a D for making the effort.

Fay: [Yes,] for making the effort.

Fiona, from the same program, used multiple lines of reasoning to give a C or perhaps a D or no grade at all. Here the measuring stick was Susan's past performance, not the performance of her classmates. At the end Fiona suggests another way to deal with negative feedback, and that is to give Susan more work:

> If this was the very first subtraction sheet she did, maybe I'd just give her a C. If this was the 15th paper she's done on subtraction, then probably a D or so. It depends where we're at in learning subtraction. Or I just might throw it away and meet with her a couple of more times and have her do a new one [worksheet] and grade that new one.

Beatrice, from an inservice mathematics program, gives a warning about assuming that all teachers work in schools where D is a pass and F is a failing grade. She also lays out a contingency based on the tasks in the worksheet. Notice the strong emphasis in Beatrice's discussion on getting answers right:

Beatrice: Seven right—I would be forced to give her a D.

Int.: So what does it take to get an F?

Beatrice: There is no F.

Int.: So D is the lowest, just four grades, and there is no F, right?

Beatrice: Also it would depend if it is a practice paper. If this were a first practice paper they would not get a letter grade on.

Int.: Would you mark them wrong or something?

Beatrice: Yeah, you have to show they were wrong.

Beverly, another teacher from the same inservice program, suggests the reason for a D stems from what the grade will be used for. Beyond the question of whether Beverly means that a worksheet grade would go directly on a report card is another interesting discussion. If grades on individual pieces of work contribute to a report card or course grade, then how is that achieved? How many worksheet grades or test grades or other evaluations should contribute to a grade on a report card? Do we assign different weights to tasks? All of this presupposes agreement that report cards should have grades. It also assumes that some form of grade accumulation is reasonable. Beverly said: "If it were a report card grade, if it was on a report card, I wouldn't give her an A or a B or a C. I'd probably end up giving her a D."

Mavis, a preservice elementary teacher candidate, invokes lack of understanding as a basis to assign the D grade. Mavis noticed that Susan failed to regroup, and we can only speculate that Mavis acknowledges that Susan understands subtraction has something to do with finding the difference between two numbers. Notice how Mavis assumes the purpose of the worksheet was not to test subtraction generally but to check Susan's conception of regrouping:

Mavis: Less than half are correct; the concept is not understood, or it's not shown to me that the idea is understood. And that's a serious deficiency.

Int.: What do you mean by the concept?

Mavis: It's the concept of regrouping. Regrouping is not understood.

The secondary modal grade was a C. Again as with the elementary informants, the secondary interviewees offered a wide range of criteria for their grades. Gerald, a secondary-math teacher from a preservice program in his first year of teaching, gave a C grade. Gerald believes the student understands most of the ideas but also thinks effort is important. His ideas about effort link to his ideas of how students learn.

Gerald: She has attempted them, except the last two, I guess. I give the kids' credit just for having done the assignment. If you do it wrong, then let me show you that

you've done it wrong, you're learning. You're learning just as much for doing it wrong as you are from doing it right. She has most of the ideas down. She just got one concept a little mixed up I think, which is slope. Judging from this work I would have to give her at least a pass her for it.

Int.: It sounds like you think it's important to consider whether the student is trying?

Gerald: Right.

Int.: Why do you think it's important to include effort as a factor in deciding the grade?

Gerald: Because I think you learn just as much from doing something wrong as you do from doing it right. If you can learn from doing something wrong, then you get credit for your effort.

Gilbert, like many secondary teachers, makes a distinction based on the task. Here the grading criteria depend on whether the assignment is homework or a quiz. Secondary teachers often discounted homework as legitimate work for grading. They argued against grading homework for a variety of reasons. Sometimes teachers wanted students to practice or reduce work pressures on themselves or because students got "help" with work outside the class. In this homework case, Gilbert invoked the fairly common criterion that the appearance of completing the work decided the grade:

> I guess it's hard for me to say, because my grading policy—I think as far as homework at the high-school level, it seems that it's necessary to, just to get 'em to do the homework. Period. And then to worry about whether it's right or not. Um, I had a teacher in high school who graded each paper twice. That is, the first time she checked to see whether there was work done, whether there was an answer, whether it was right or not. After the students had a chance to correct the work, we turned it in for an actual grade. I liked that policy. I would do the same thing. If this were a quiz, I would evaluate it objectively.

When asked, "What would that mean?" he said, "A certain number of points for each problem. And maybe two points for the problem. That they get the equation `right,' that they get the graph of the line `right.'"

Joan, another preservice mathematics teacher, sees the equation and the graphing as carrying different levels of significance:

> I probably wouldn't give her a very good one because she didn't understand it. Getting

> the equation right wasn't that hard. Most of the points would probably be in the graphing of the line because that's what shows that she really understands it.

It is more difficult to understand Cain's claim that the work sample shows a partial lack of understanding. What is typical is his considering the possible psychological or motivation impact of the grade: "Well, she had some concepts down, but not well. I mean she is at least doing something. She is on her way. I do not want to discourage her too much." Thus what looks from a distance like a single thread of commonality—a C grade or D in the elementary cases—consists of separate distinct strands that vary greatly. The individual strands represent a loose weave of reasoning that describe why teachers assign the grade that they do. As the reasons for the grades differed so much, we focused on the rationale these teachers and teacher candidates used to justify their grades. Tables 4 and 5 break out teacher strategies for inferring student effort or understanding respectively.

Table 4 shows two main sources of evidence and a concomittant low level of criteria on which the informants base inferences about the effort a student expended. The internal source for an inference meant that the informants cited evidence from the work sample in front of them. In contrast, an external source means the informants' inferences used evidence beyond what is physically present in the work sample before them. It is not our intention to argue that either source is better or beyond what an informant's verbal description of the source said for the criteria.

Table 4 shows that twice as many inferences came from the work itself. The most popular internal inference source referred to the physical evidence in front of them. Many claimed the inference was easy, that "it is obvious" that the student tried. Further questions revealed the main evidence in support of the judgment was that the student attempted most or all of the problems on the page. The "obviousness" of effort lie in the physical presence of marks on the page. Looking at internal sources of evidence may suggest a desire to appear objective and not let extraneous factors color the judgment.

Geoffrey, from a preservice program, is succinct in his evidence for effort. When asked, "How can you tell that she's gone to some effort?" he responded, "Well, she's attempted to do all of them." Sharon, from another preservice program, is equally convinced. Asked how she knows the student made a great effort, she says, "She worked all the problems, even if they weren't right."

The second internal source for inferences about effort was the neatness of the work sample. Six people used the evidence of neatness in presentation as their criterion for effort. Both the work samples had clearly written features. The respondents inferred different things from this level of presentation about effort on Susan's or Lynn's part. Carol said,

Table 4

Teacher Inference About Student Effort

Internal Source for Inference	# of Instances	External Source for Inference	# of Instances
Tried all problems on page	19	Consider student's overall class performance	6[a]
Neatness of presentation	6	General student's behavior patterns	5
Student not tried based on lack of work shown, or some vague feature	3	Handing in work complies with a teacher policy	2
Evidence of no cheating based on mistakes left	1	Evidence of cheating[b] based on odd/even problem numbers	1
Only slight errors on every problem	1	Careless work if student had a lot of prior practice	1[a]
Internal source	30	External source	16

[a]Teachers who made an explicit association between a student's effort and his or her academic performance.

[b]The evidence for cheating comes from the structure of mathematics textbooks that give only odd or even answers in the back. If a student gets only those answers for the odd or even numbered problems right, then this is good evidence of cheating. It may tie to a more general practice teachers have of assigning only the odd or even numbered problems for homework (whichever is not supplied in the textbook) to stop cheating. It speaks volumes about the teacher's sense of the sanctity of getting the right answer unassisted!

> If she was assigned 7-14, let's say it was five points, I would give her four out of five points, because she tried almost all of it. It looks like she started 13, but didn't really put much of an effort into it. At least she could've drawn in a line, even if she thought it was wrong. And 14 she put no effort into. So, it's almost all done, she spent some time on it, so. . . . And her lines look straight, it looks like she didn't, just do it sloppily or carelessly.

Catherine, a secondary mathematics teacher like Carol in an induction program, first felt Lynn made some effort. There are few clues to how Catherine decided Lynn made an effort. However, Catherine sees the effort as less than Lynn's best and so represents the third category of variation of internal source inferences. Unlike the top category, Catherine thinks Lynn is "hiding" her work and so does not deserve the top grade. At the end she explains how students do not always put down all the thoughts they had to get the answer. Catherine, it seems, has some pattern of "working" she would like to see. It is vague, as the third category states, because she never verbally described what she expected to see. It could be based on how she personally works these problems out or some worked example she showed on the board and wants replicated in a student's work. The expectation of what Catherine expects to see may lie in her vision of what she implies is going on in Lynn's head and not shown on the paper:

> A check means, first that it's not all complete. I see that she made a real effort. A check-plus means that you did all the problems and that you tried really hard, that you really made an honest effort. To me, to get a check-plus you need to show your work and you need to make an effort. You can get everything wrong, but show all your work and I'll give you a check-plus. What I mean by show your work: At this point there are many things that she did in her head to get these graphs. I wouldn't have to ask her how did you come up with this line. If this was down on her paper I could find out right off the bat by looking at the points. So I would give it a check.

The last two internal sources may represent idiosyncratic pieces of reasoning and yet they are interesting. They contrast with the simple reasoning of other internal sources. One secondary teacher inferred by going through Lynn's paper in detail that the mistakes Lynn left showed she made a genuine effort. The lack of obvious erasing was the basis of this inference. An elementary teacher looking at Susan's work implied that the errors did not show understanding as one might expect but showed the student had made an effort.

The right-hand column of Table 4 summarizes inferences about effort based on criteria that were not part of the paper as such. The most popular external source for the inference was student performance. This student performance measuring-stick strategy turns out to be a complex method of

verbal reasoning that covers a variety of related criteria. This group talked of the importance of considering how the time at which this work sample was assigned related to other work that the student or the whole class had completed, and about whether Susan's and Lynn's work was from the beginning, the middle, or at the end of a sequence of learning. The performance measure could also be compared to other work by that student or to other students in the class. Considering general patterns of performance allows the teachers to adjust the grade upward when an academically weak student acts above the expected level. Or it allows a teacher to penalize a student who shows a lack of progress.

Five teachers wanted to judge effort based on a student's general behavior in the classroom. Looking at general behavior enabled the teacher to decide if the work represented an honest effort on the student's part. This allows for cases like the lazy-but-intelligent student to get penalized for laziness. Jessica, a preservice teacher candidate, is very explicit about the level of student compliance required and offers some explanation of a link with effort and teachability:

> Number one, does she pay attention in class? Number two, does she apply herself to her work? Does she finish it without looking at other people's papers? When I get to work with her individually, what is her attitude? Is she closed?—Well even if she's trying, she could be defensive or she could be open to suggestions. I just think by the way she pays attention and by the way she applies herself and by the way she reacts if she finds out that she's only gotten two correct, if she didn't try she wouldn't be real upset. Many people wouldn't.

The other three external source categories contain some variations on the judgment about general academic work patterns or student personality traits. An interesting rationale, used by two teachers, was that, because Susan or Lynn had complied to the teacher's request and handed in the work, this showed effort.

Overall, the table reveals the importance many teachers attach to students showing some form of compliance with instructions. Neat presentation, making a visible attempt at all or most of the problems, and behavior in class highlights the importance teachers give to students doing what they are told to do. Only 7 teachers (see footnote a in Table 4) out of 46 made an explicit association between a student's effort and his or her academic performance. For the rest, student effort is a separate, stand-alone criterion, which teachers need to recognize and reward when grading.

Table 5 shows inferences about students' understanding. This table shows a sharp difference between elementary and secondary-math teachers. Over one third of the elementary teachers and teacher candidates made an inference about student understanding. In contrast, 57 of the 58 secondary teachers and teacher candidates described some inference about their student's understanding. In addition, more than half the 58 secondary-math teachers provided a detailed verbal breakdown of the

Table 5

Inference Categories in Student Understanding

	Inference Category	Elem. *N* = 143	Sec. Math *N* = 58
Holistic evaluation	Partial understanding	15	15
	Partial understanding based on general lack of manipulation	N/A	8
	Ideas not understood	9	4
	Ideas are understood	3	1
	Mistakes show overall lack of comfort with ideas	2	0
Evaluation based on counting	No understanding based on count	3	0
	Understanding based on count	2	0
Evaluation based on failure to be consistent	Lack of consistency in application	8	0
	Easily confused by certain features of the task when applying idea	3	1
Evaluation split into small steps or parts	Understands subtracting not borrowing	5	N/A
	Idea of slope understood	N/A	2
	Intercept understood	N/A	10
	Major equation quoted correctly	N/A	8
	Negative numbers not understood	N/A	8
	TOTAL INFERENCES	50	57

Note. As in Table 4, one informant may make several inferences about the work sample.

Table 6

The Distribution of Inferences About Past, Present, and Future Performance

Past	Present	Future
Grade rewards effort by student (summative evaluation)	Grade will cause poor self-image (impact on learner)	Grade affects motivation to try later (impact on learner)
	Grade reflects present level of understanding (summative evaluation)	
Secondary only 56 instances	Secondary and elementary 89 instances	Elementary only 59 instances

particular idea the student seemed to understand.

The top row of Table 5 includes interviewees who mentioned student understanding as an issue but did not say how they learned about that understanding. The student "partially understood" or understood some ideas. Equal numbers of secondary and elementary teachers offered these vague responses. However, with the smaller number of secondary informants, those 15 inferences represent a significant proportion of the secondary teachers. Even when pressed, none of the teachers were explicit about the criteria on which to base that judgment. Many responses in the holistic category, which vaguely referred to some sort of understanding with no clear criteria, may also be based on a count. In response to the probe, it was often said that understanding or lack of understanding was obvious. The quality of obviousness could come from some silent count that the informant failed to verbalize.

Generally these and other holistic assessments suggest grading was not a practice that the teachers have thought about. This seems the most likely explanation for the lack of articulation over both secondary and elementary teachers. It is difficult to tell, for instance, from Jay's reply (a secondary mathematics teacher candidate) what the basis is for a judgment on the student's understanding. Imagine the effect on Lynn if this were the feedback she received about her work. Jay hints at some standard but might benefit from some discussion on how to express that judgment with students:

> Jay: C-minus maybe. Because the understanding is almost right, it seems like it just from looking at her paper.

Int.: How can you tell that?

Jay: Because she is not way off. It's not like she doesn't understand at all. She has some idea of what she is supposed to be doing. At least from a purely procedural standpoint she knows which part she needs to manipulate. She just misses with the manipulation I suppose.

Ginger (an elementary teacher candidate) mentions understanding being in the "ball park." What makes it in the park is not clear, at least to us on this evidence; neither is the "something" she sees as "consistent." Whatever that unstated consistency is apparently qualifies the student for a better grade and a different assessment of their understanding. Ginger said,

> To be honest, that's a part of teaching I'm dreading most, having to assign a letter to a child. I think I'd probably assign a C. She did understand the concept in part. I think that effort was behind it. She is not coming up with the answers totally out of the ball park. I do see something consistent but I couldn't justify giving her anything higher either.

Elementary teachers often commented that Susan had partial or even complete understanding but that in calculating the individual problems they failed to apply the ideas consistently. Mindy, a preservice elementary teacher, is looking for consistency from Susan. The work sample showed Susan applying the difference rule consistently. Mindy may see inconsistency in some "borrowing" procedure. So to draw the conclusion that the teacher sees mathematics as simple rote application of a rule may not be the right one. Mindy in the next example suggests Susan knows how to take the difference—that is what she "knows how to do"—but does not use a "borrowing" strategy yet. Consistency may assume a meaning closer to *successful use in the right circumstance* as distinct from *mindless application of a rule*: "I guess I would say 50. Fifty percent. Even though she didn't get 50 percent of the problems right, she can do it, she knows how to do it, but she just didn't do them all that way."

Many secondary teachers, but only five elementary teachers, talked about a checklist of features for the task that identifies significant points of student understanding. In the explanations the teachers articulated specific criteria for their grading, as shown by the subdivisions in this category. They would explicitly describe Lynn's failure to understand the use of negative numbers. In contrast, few elementary teachers mention explicitly Susan's lack of "borrowing" as a reason for the grade.

In Lynn's graph work they focused, for example, on small failures to manipulate the basic equation and then inferred some lack of understanding on Lynn's part. At first glance this comes closest to evidence of teachers' detailed diagnosis of Susan's and Lynn's understanding. Unfortunately,

like many instances of checklist procedures, it is not clear what a check in the list means. It could mean that the teacher needs to work with Lynn on negative numbers or it might simply express the summative judgment that Lynn failed to understand negative numbers.

We are confronted with a fascinating paradox. Elementary teachers, at least based on these data, give vague verbal evidence of their evaluation. However vague those descriptions are, elementary teachers willingly use that information for further teaching on their part or extra student practice or other action. Their secondary colleagues, in contrast, describe detailed evidence but then do little except pass judgment on the students' understanding.

Secondary teachers provide a more detailed description of the point of failure but then feel compelled simply to grade that understanding level. There is a temptation to conclude that their elementary colleagues articulate and maybe even understand the subject matter less clearly. However, it might be that elementary and secondary teachers view student responsibility for understanding differently. The secondary teachers may see the students as intellectually and morally more blameworthy for their failures.

We describe this paradox in more detail as a final issue. Here we wanted to know what pedagogical purposes these grades served? Table 6 summarizes the distribution of inferences about the students' past, present, and future performance made by elementary and secondary teachers. Elementary teachers prefer to use the grading task to promote a future teaching or learning effort, using grades in a way that affects the future work of their students. According to Mindy,

> I would just write, "Try again." I wouldn't want to write "poor" on it or give her zero or a flunk or a fail, because she didn't fail in my opinion, she needs to be encouraged to try again and do it the right way.

Judith, another elementary preservice teacher, is about to give Susan an F after being reminded about the school policy. In this quotation, she shows resistance to the idea of grading followed by reteaching through discussion and finally the perceived effect of grades on future student motivation:

> I have a real hard time with grades. If this was my class and I could do whatever I wanted, I would talk to her about this, we'd go over it again, I'd have her do it all over. I wouldn't give her a grade; I think that is so discouraging; I think it is so counterproductive.

Secondary-math teachers, on the other hand, like to use grades to reward past effort.

> Carson: If they did all the homework, or tried to do it all, even they didn't get the right

answer, they get the credit. Uh, they get only partial credit if it's incomplete.

Cecil: She didn't write anything down. And she should have tried. I mean in my classes I would always make that a point. I'm going to tell my students to at least try, write something down. Write anything down, even if it's wrong.

Gilbert, just at the end of his preservice program, has a two-step scheme based on his own school experience, which rewards both effort and level of understanding:

> I guess it's hard for me to say, because my grading policy—as far as homework is concerned—at the high-school level, is that it's necessary to just to get them to do the homework. Period. Then to worry about whether it's right or not. I had a teacher in high school who graded each paper twice. The first time she checked to see whether there was work done, whether there was an answer or not. And, and after having a chance to correct the work, we turned it in for an actual grade. I like that policy enough so I would do the same thing.

Both elementary and secondary-math teachers are prepared to make inferences about the student's present self-image or level of understanding and express that as a grade. Jessica wants to preserve Susan's self-image through two indirect strategies to point out errors. First, she will mark only correct answers and then reteach the whole class so Susan will not feel isolated. Finally, she expresses her evaluation of Susan's level of understanding:

> I would mark the ones that were correct and then she'd wonder why the other ones weren't, maybe; but as a rule, I'd probably would go over it with the whole class, and if she's smart at all she'll know hers weren't correct. . . . I couldn't give her a lot for her grasp of the conceptual skills, probably unsatisfactory, maybe satisfactory, plus based on effort or at least satisfactory, but S-minus or unsatisfactory in the way she performed this.

Elementary teachers see different significance in grading compared to their secondary colleagues. Recall from Table 1 that they were far more likely than their secondary counterparts to express their reluctance or dislike for giving a grade even as they did so. Elementary teachers show a reluctance to grade or talk of grading as a final step and often graded with equivocations and constraints. Many elementary educators see the grade, especially assigning a letter, as a significant and even a distasteful task. Elementary teachers look for a pedagogical criterion when grading. There is a sense of "never too late" and a flexible evaluation policy that postpones any final judgment about the ability and potential of children.

The elementary teachers who assign F grades invariably mentioned the need to allow students another chance. Leslie, a preservice elementary teacher, was typical, claiming the student lacked some understanding, and if allowed to repeat the work, after teachers' help, the student could improve the grade. Table 6 shows that most elementary teachers are concerned about the effect of grades on student motivation. Whenever there was a likelihood of a poor grade, many add a need for further teaching or giving the student another chance. Leslie believes that

> grading should be a way of encouraging students to continue with the kind of behavior that you want. Continuing with the skill you're trying to teach them. Since she only has four that actually have the correct answer. If you did it on a strict counting basis and just said, "This is your score," it would be a bad score. It wouldn't encourage her getting these right. It wouldn't give her any sort of reward for doing that. But by going back and showing her how she could continue, where she was doing the sums right, and [therefore she could then] get a good score, then you would be encouraging her.

Elementary educators look forward to future student performance or to further teaching. Student motivation is part of that concern for future student performance. Their secondary colleagues are more likely to see student work as representing past or present accomplishment that needs rewarding.

Secondary teachers function in an institutional environment that reinforces their tendency to dispense reward to their students. Secondary teachers become gatekeepers on whose largesse students move to the next level in the track. Summative evaluation is needed to form part of the permanent record for a student. Providing just reward for visible effort and compliance is an appealing and powerful position. Secondary teachers feel less institutional press to avoid failing grades and assume students bear responsibility for their own failure. School administrators often audit secondary teachers' gradebooks to ensure there are not too many high grades in any given class. The organizational press is toward limiting reward. Mathematics courses in many high-school environments can influence which classes students take next year and the college they eventually apply to. The math department enjoys an elevated and privileged position in many high schools.

Teachers rewarding effort come closest to the generally assumed summative purposes the public often gives to grades: The teacher acts as a judge with no explicit pedagogic purpose, such as providing feedback to aid future student performance. We noted mostly secondary teachers using a seemingly detailed checklist. Taken with Table 6 it looks as though summative evaluation with no pedagogic purpose is the more likely candidate for the meaning of a check on the list. A check is not a flag for further teaching or learning opportunities to be provided for the student.

Conclusion

School mathematics at face value is a ripe subject for objective evaluation. It is deeply embedded in the public's belief that sums are either right or wrong. While there is much counting of right and wrong answers in the data, many teachers feel compelled to infer a variety of things from that count. Teachers here struggled to make personalized connections that are congruent with their role as a teacher.

These data show several things very clearly. First, it is unlikely that there is or ever will be a single foundation for all grading practices discovered by a teacher (or researcher). The measurement community's marked lack of visible enthusiasm to write or seriously consider teacher grading supports this conclusion. Second, the data suggest some interesting and some alarming criteria on which we could base the assignment of grades.

Third, beyond historical inertia no informant offered a compelling rationale that argued we should use the letters A through F to show our evaluation. The reverse impression seems closer to the point. Giving a letter grade hid a multiplicity of evaluations based on a variety of rationale. Fourth, the data suggests how a common teaching practice has a life and substance of its own. In doing so it has become self-sustaining while simultaneously loathed, ignored, or marginalized.[4] Fifth, the data suggest two broad sets of purpose to which a grade could be put. A grade can form either a judgment on past effort or form part of a strategy to influence present or future student performance. There is no exclusionary choice here, but such a complex set of potential outcomes cries out for discussion and reflection.

For teacher educators, grading represents a novel point of entry to consider instructional behavior and learning opportunities the prospective teacher might provide for students. Specifically, if the student teacher wants to design a sequence of hands-on experiences to teach subtraction, what part, if any, would a worksheet evaluation, such as Susan's work sample, play in that sequence? More generally, what is the appropriate and most informative way to evaluate student performance and progress? Given the different ways teacher education now sees teaching and learning, What role does the teacher's evaluation of a student's performance play in these new methods of teaching and learning?

Thinking and talking about the grading question might lead to some spirited discussion of what evidence of student understanding is. Beyond that, discussing the question of how to grade might get us clearer on what the "it" is that we want the students to learn. Susan's worksheet might appropriately measure her applying the "borrowing" rule of subtraction. Lynn's graphs might be graded to evaluate her manipulation of the standard slope equation. Our informants do not show an overwhelming

[4]It is interesting historical footnote that among the NCRTE researchers there was some debate as to why we should include a grading question in the structured exercises.

consensus that this is the purpose of grading they had in mind when they comment on these two student work samples.

One troublesome topic in the data is how we deal with negative or poor evaluations of students. It is a dilemma faced by physicians, politicians, and others who sometimes have to bring bad, not good, news. The student work samples were designed to present this dilemma. Some teachers constructed rationales that allowed them to give credit beyond the "normal" response. Others give a poor grade but then talk about revising the grade and giving second chances. Is it defensible to exclude all negative evaluations from students and give only positive evaluations? Is it the very young who should not be told bad news? Who really feels worse giving a negative evaluation, the student or the teacher? These and a multitude of other ethical questions need serious consideration.

Grading appears as a neglected practice from the researchers' perspective. Following their lead, teacher educators have not paid much attention to it. Not all schools and teachers enjoy the luxury of ignoring the visible evaluation of their students. Even if we decide that no grade should ever appear on a student's work, it seems the education and research community need to argue that case in a public forum.

References

Agnew, J. E. (1985, March). *The grading policies and practices of high school teachers*. Paper presented at the meeting of the American Educational Research Association, Chicago.

Borko, H., Cone, R., Atwood Russo, N., & Shavelson, R. J. (1979). Teachers' decision making. In P. L. Peterson & H. J. Walberg (Eds.), *Research on teaching: Concepts findings and implications* (pp. 136-160). Berkeley: McCutchan.

Clark, C. M., & Peterson, P. L. (1986). Teachers' thought processes. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 255-296). New York: Macmillan.

Clark, C. M., & Yinger, R. J. (1979). Teachers' thinking. In P. L. Peterson & H. J. Walberg (Eds.), *Research on teaching: Concepts findings and implications* (pp. 231-263). Berkeley: McCutchan.

Nitho, A. J. (1989). Editorial. *Educational Measurement: Issues and Practice, 8*(2), 4.

Simon, S. B., & Ballanca, J. A. (Eds.). (1976). *Degrading the grading myths: Primer of alternatives to grades and marks*. Washington, DC: Association for Supervision and Curriculum Development.

Stiggins, R. J., & Bridgeford, N. J. (1985). The ecology of classroom assessment. *Journal of Educational Measurement, 22*, 271-286.

Stiggins, R. J., Conklin, N. F., & Bridgeford, N. J. (1986). Classroom assessment: A key to effective education. *Educational Measurement: Issues and Practice, 6*(2), 5-17.

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

Terwilligar, J. S. (1971). *Assigning grades to students*. Glenview, IL: Scott, Foresman.

Terwilligar, J. S. (1989). Classroom standard setting and grading practices. *Educational Measurement: Issues and Practice, 8*(2), 15-19.

Weinshank, A. B. (1980). *An observational study of the relationship between diagnosis and remediation in reading* (Research Report No. 72). East Lansing: Michigan State University, Institute for Research on Teaching.