Calibrated Peer Review for Interpreting Linear Regression Parameters: Results from a Graduate Course

Felicity B. Enders
Sarah Jenkins
Verna Hoverman
Mayo Clinic

Journal of Statistics Education Volume 18, Number 2 (2010),
www.amstat.org/publications/jse/v18n2/enders.pdf

Key Words: Statistics education; Writing assignment; Interpreting regression coefficients.

Abstract

Biostatistics is traditionally a difficult subject for students to learn. While the mathematical aspects are challenging, it can also be demanding for students to learn the exact language to use to correctly interpret statistical results. In particular, correctly interpreting the parameters from linear regression is both a vital tool and a potentially taxing topic. We have developed a Calibrated Peer Review (CPR) module to aid in learning the intricacies of correct interpretation for continuous, binary, and categorical predictors. Student results in interpreting regression parameters for a continuous predictor on midterm exams were compared between students who had used CPR and historical controls from the prior course offering. The risk of mistakenly interpreting a regression parameter was 6.2 times greater before the introduction of the CPR module ($p=0.04$). We also assessed when learning took place for a specific item for three students of differing capabilities at the start of the assignment. All three demonstrated achievement of the goal of this assignment; that they learn to correctly evaluate their written work to identify mistakes, though one did so without understanding the concept. For each student, we were able to qualitatively identify a time during their CPR assignment in which they demonstrated this understanding.

1. Background

Biostatistics can be very challenging for students of other disciplines to learn (Berwick et al., 1981; delMas et al., 2006; delMas et al., 2007; Weiss and Samet, 1980; Windish et al., 2007; Wulff et al., 1987). Not only are the mathematical aspects demanding, learning to appropriately interpret statistical results can be a very intensive process. This aspect of biostatistics may be especially challenging because statistical interpretations, while seeming to use language rather than math, are actually quite restrictive with respect to both structure and wording. The goal of this project was to design an assignment centering on peer feedback to help train students to give more accurate interpretations.

Van de Ridder et al. (2008) define feedback in the clinical education setting as "specific information about the comparison between a trainee's observed performance and a standard, given with the intent to improve the trainee's performance." In the educational field as a whole, feedback is an invaluable tool for increasing knowledge and skills (Moreno, 2004; Pridemore and Klein, 1995) and motivating further learning (Lepper and Chabay, 1985). Lou et al. (2003) define conditions for feedback to be effective in promoting learning: "appropriate corrective messages need to be sent to learners; the messages themselves need to be interpretable by learners; and learners need to possess prerequisite prior knowledge, motivation, and strategies to respond effectively to the feedback they receive." Peer feedback can be one tool for providing a learning experience to students. The goal of peer feedback is to give each student three learning opportunities; in the original assignment, as a peer reviewer, and in receiving peer reviews.

Many question the accuracy of peer reviews. Falchikov and Goldfinch (2000) reviewed 48 studies in a meta-analysis to compare reviews by instructors to reviews by peers. They found that peer assessments correlated with instructor assessments ($r=0.69$). However, McCarty et al. (2005) found peer reviewers had a tendency to provide higher scores than faculty by an average of three points on a 10-point scale ($p<0.0001$). Fortunately, the goal of CPR is to enhance learning by providing students an opportunity to act as graders. Exact replication of a review by an expert statistician is not required. In this way, CPR makes use of a finding from many studies that peer feedback can help students develop critical reviewing skills and judgment through providing a peer review to others (Boud, 1990; Davies, 2006; Orsemond et al., 2004; Pope, 2005; Reese-Durham, 2005; Topping and Ehly, 1998; Trautmann et al., 2003; Van de Ridder et al., 2008; Van den Berg et al., 2006). Although based upon numbers, such critical reviewing skills are still vital to correctly interpreting regression parameters.

One reason statistics is so challenging is that many words have particular meanings in statistics. This lexical ambiguity means that students have to learn the precise meaning of words within this context before they can appropriately interpret statistical results (Kaplan et al, 2008). For instance, the word "mean" has a different meaning in statistics than in typical English. Even within statistics, the word "mean" appears to some students to have a different meaning for a t-test, where the actual mean is used, and for linear regression, where the theoretical mean at a particular value of X is used. For physicians and many other regular consumers of statistics, correct interpretation of regression parameters is one of the most critical statistics education outcomes.

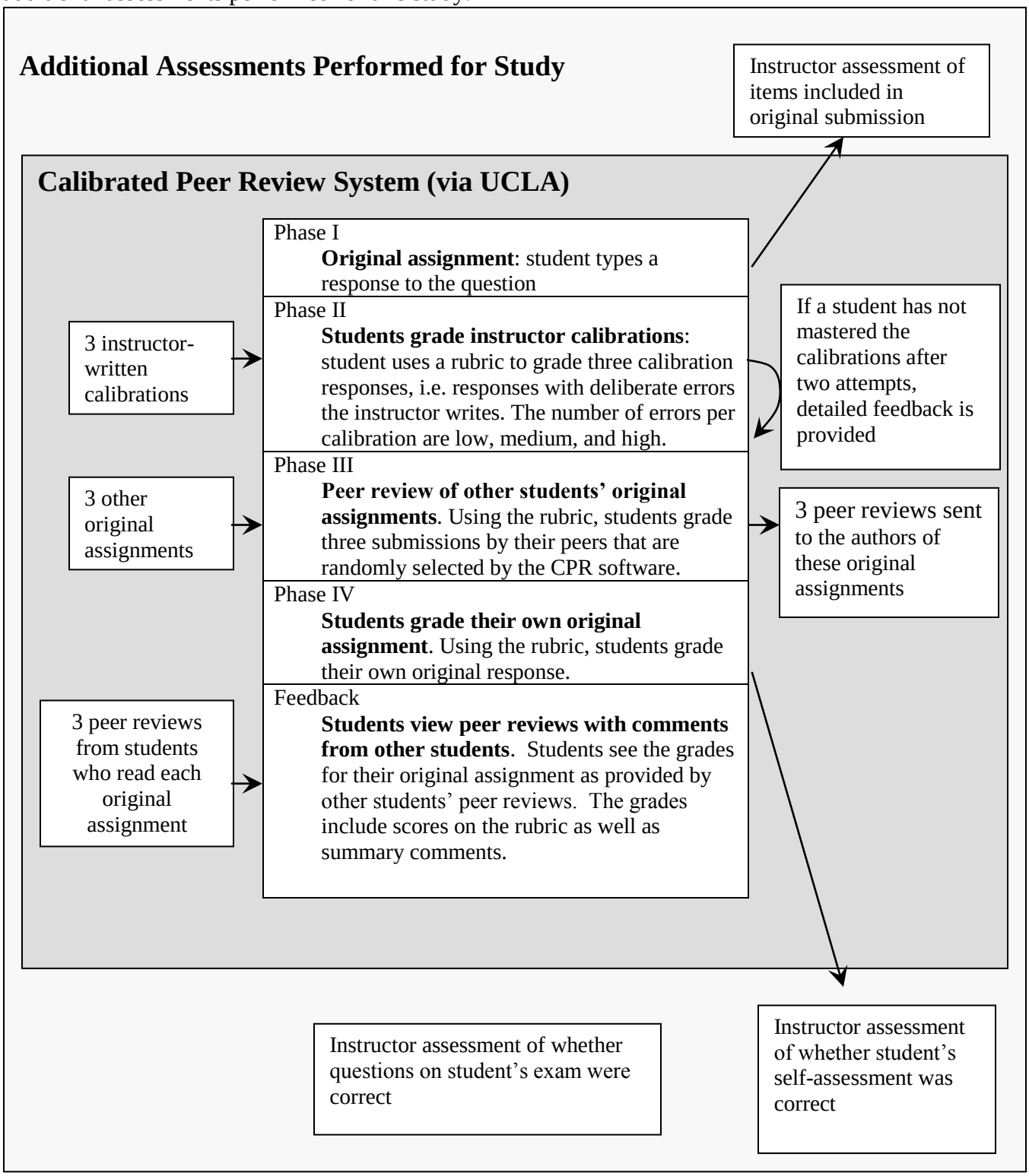
Within the context of linear regression, another challenge to correct interpretation is the necessary difference in parameter interpretation based on the type of predictor variable used. For continuous predictors, the "slope" is an actual estimated slope of the outcome variable per unit difference of the predictor. For binary variables, the "slope" is an estimated difference in means between the groups for which the predictor is set at one and zero. For categorical variables coded with dummy variables, the "slope" is similar to a binary variable, except the reference category is defined by the full set of dummy variables. These differences may not become clear until the student masters the corresponding interpretation of linear regression coefficients.

Former students in our course in 2004 and 2005 had been repeatedly exposed to interpretation during lectures and in the textbook and then had to write interpretations in computer laboratory sessions and on graded homework assignments. However, by the time of the course examinations many students (22% in 2005) were still making mistakes on basic interpretations of estimated linear regression coefficients. This led us to develop the Calibrated Peer Review (CPR) module discussed in this paper.

2. Calibrated Peer Review

Calibrated Peer Review is an online peer review system available at cpr.molsci.ucla.edu. Every CPR module has four phases which are described in Figure 1. Each student first writes an original assignment. After grading calibration assignments written by the instructor and intended to train students to provide appropriate reviews, each student then provides peer reviews for three other students' original assignments. Finally, the student provides a self-assessment of his or her own original response and then is provided access to the peer reviews other students have provided evaluating his or her own original assignment as a final learning opportunity. All portions of the CPR system are blinded. The assignment, calibrations, and grading rubric used in the CPR module may be found in the appendix.

Figure 1: The phases of a CPR assignment, including interaction between students and additional assessments performed for this study.



It should be noted that while the questions in this CPR module may seem to require statistical reasoning, the module is primarily intended simply to increase statistical literacy for this advanced statistics topic (delMas, 2002; Garfield, 2002; Rumsey, 2002). Specifically, the module is intended to ensure that 1) students know and choose the correct words and components of regression interpretation and 2) students understand the difference in interpretation required by different types of predictor variables.

The author of a CPR module writes the text entry question including any attachments or links needed, the three calibrations, the grading rubric, and the feedback for mistakes made in grading each calibration. The CPR module can then be published either within the author's institution or for the general public. There are currently only two other CPR modules with statistical titles in the CPR assignment library ("Linear Functions and Slope" and "Linear Regression" both of which are written at an introductory level). When an instructor chooses to use a particular assignment, he or she controls the evaluation points associated with each phase and the timing of the assignment. In this assignment, 15% of the score was devoted to the student's original submission, 30% to the calibrations, 30% to the reviews, and 25% to the self-assessment. However, the CPR score was not used in the course; the CPR assignment was an ungraded exercise. The CPR system is set up so that students have a certain period of time to complete phase I, and then a second period of time for phases II - IV, thus ensuring that all students receive peer reviews and have the opportunity to provide peer reviews and no student sees the grading rubric before completing phase I.

CPR is widely used; the assignments have been used by over 600 institutions. Nevertheless, CPR has both advantages and disadvantages. On the plus side, it provides a way to include essay questions in a large course in which grading such questions might otherwise be arduous or impossible (Carlson and Berry, 2003; Carlson and Berry, 2005; Hartberg et al., 2008; Heise et al., 2002; Kim et al., 2005; Lin et al., 2001; Prichard, 2005; Wise and Kim, 2004). CPR has been cited as a way to help students improve their writing by bringing greater depth to their critical thinking skills (Hartberg et al., 2008; Kim et al., 2005; Prichard, 2005). CPR models the peer review process for publications, thus providing a tool that students may need for future success (MacFarlane et al., 2005). Learning to review one's work may help students to better reflect upon their own writing as well as that of others. The review process also encourages students to write precisely, a skill needed to communicate in the increasingly collaborative scientific community. Gunersel and Simpson (2009) found that initially low-performing students improved in writing and in reviewer competence over the course of three CPR assignments, though their analysis used internal CPR scores rather than external assessments of accuracy and competence. CPR also provides students the opportunity to become familiar with a grading rubric, analogous to the maxim that one doesn't really learn a subject until one teaches it. Finally, it turns a writing assignment into an active learning exercise. One downside of CPR is that it can seem repetitive to students, especially if the grading rubric (used at least seven times during a CPR module) is long and involved. Also, the final CPR "score" includes peer reviews, so many instructors may not want to include it as a component of the course grade. Finally, CPR modules rely on students having prior exposure to critical content; while background material can be included within the assignment, the focus of the assignment is on the text entry and using the grading rubric. Overall, CPR modules can be an excellent addition to a statistics course to

reinforce topics that can be evaluated with free text questions. Since our focus was solidifying interpretation of linear regression, this seemed like a perfect fit.

3. Methods

During this project we developed a CPR module to reinforce interpretation of estimated linear regression coefficients for continuous, binary, and categorical predictor variables after these methods had been introduced in lecture. The module was pilot tested, after which rubric questions were refined to better match the query asked of students. The module as described in this study was then used within a second course in biostatistics in 2006 targeted to physicians and other graduate students. The course was formatted so that linear regression with one predictor variable of any type was introduced prior to the midterm examination. Use of a continuous predictor was introduced in week 1, followed by time spent on understanding the mechanics and tools used in linear regression. Binary and categorical predictors were introduced in week 3 immediately followed by the CPR assignment, and the midterm exam was held in week 4. After the midterm, topics covered multiple linear regression; the final exam was given in week 8. The class met for three one-hour lectures per week plus one two-hour guided computer laboratory session with the instructor and teaching assistants. There was approximately one homework assignment per week. Prior to each examination, students were provided with practice exams, exams from prior years' course offerings, to help them study. The course was tiered, so that material became increasingly difficult in labs, homework, and examinations. Examinations included questions in which the student was asked to perform a problem analogous to one seen before (such as interpreting an estimated regression coefficient) and questions in which the student was required to integrate concepts from different sections of the course and engage in statistical reasoning.

In addition to the standard CPR processes, the instructor (F. Enders) graded both the students' CPR text entry and analogous questions on the students' course examinations using the CPR grading rubric. Both evaluations were blinded to the student's identity. The Pearson correlation was used to compare the self-assessment to the instructor's score. The first evaluation of the CPR module was a paired comparison of the score received in phase I of the CPR module vs. the score received on the analogous exam questions; statistical significance was assessed with the Wilcoxon signed rank test. The exam questions for the continuous and binary predictors were given on the midterm; the interpretation of a categorical predictor was included in the final exam. No exam question assessed interpretation of the reference category of a binary predictor, so the first three rubric questions were not used for the exam. A graph was also used to compare students' scores on their self-assessment from phase IV with the instructor's evaluation of their text entry from phase I. Cronbach's alpha was used to assess internal reliability of the grading rubric across responses to all three calibrations.

Individual items from the CPR grading rubric were compared with respect to the proportion of positive responses within the student's original assignment as graded by the instructor, the student's self-assessment, whether the self-assessment matched the instructor-assessment, and within the student's examination responses. A paired assessment was performed with McNemar's exact test to compare 1) the student's self-assessment to the instructor's assessment, 2) The items correctly graded in the self-assessment to correct exam items, and 3) the responses

on the original assignment vs. the exam responses, both as graded by the instructor. We chose to assess each item separately with 5% significance to identify the specific areas in which the CPR tool helped students; the overall assessment is provided by the Wilcoxon signed rank test described above.

Blinded grading of course examinations by the instructor is standard practice in this course, so we were also able to compare the percentage of students who produced perfect interpretations for the relevant questions on the midterm examination from the year of the study and the prior year; statistical significance for this analysis was assessed with Fisher's exact test. The two course offerings were very similar except for the introduction of CPR. They shared the same instructor, teaching assistants, topics, timing, and the majority of laboratory and homework questions. However, the examinations were entirely different between the two course offerings. Unfortunately, the only relevant exam question with analogous wording between the two years covered interpreting linear regression for a continuous predictor variable; interpretations for other types of predictors are excluded from this portion of the analysis. For the 2006 class, this comparison used the portion of the grading rubric pertaining to use of a continuous predictor. For the 2005 class, exams had been returned to students so we instead used scores for individual questions. Since the grading differed by year, we restricted this analysis to only perfect vs. imperfect scores. The CPR grading rubric is similar to exam grading for these questions but provides somewhat more detail, so the results should have been biased towards fewer completely correct responses in 2006.

The final evaluation assessed students' perception of the CPR module at the time they completed the course. Students were anonymously asked "How did you like the CPR system" with response options on a linear analog scale from 0 "I hated every minute" to 10 "I loved every minute" and "How helpful was it to grade your own submission with response options from 0 "Not helpful" to 10 "Very helpful." Students' response to the two questions was compared with the Wilcoxon signed rank test. Statistical analyses were performed with the Stata statistical software package (v. 8.2, College Station, TX). Two-sided 5% type I error was used to determine statistical significance for all analyses.

In order to qualitatively explore the learning process within the CPR assignment, we chose students with the minimum, median, and maximum text entry scores as graded by the instructor. We followed students' learning process through the assignment for one question each; the question was chosen to be a typical mistake for students at that level.

4. Results

4.1 Pre-Post Evaluation of Efficacy

Of the 29 students enrolled in 2006, 28 (97%) participated in the calibrated peer review activity. Only those completing phase I of the CPR module were included in the first analysis. The instructor-graded score of the text entry and analogous course exam questions used the same grading rubric and had a theoretical range between 0 and 19. There was a significant improvement from the text entry to the course examination as graded by the instructor ($p<0.0001$; see Table 1). We also assessed the correlation between students' own rating of their

text entry during the phase IV self-assessment and the instructor's scoring of their text response: the correlation was high (r=0.95, 95% CI: 0.89, 0.98).

Table 1: Item-level and overall comparisons of self-assessment and instructor-graded CPR scores with exam scores (questions are shown in appendix)

	Self-Assessment of Original Assignment	**Instructor-Assessment of Original Assignment**	**Instructor-Assessment**	**Exam**	
Question	**No. (%) Yes**	**No. (%) Yes**	**No. (%) Correct in Self-Assessment**	**No. (%) Correct**	**P***
1	25(96)	25(96)	26(100)		
2	24(92)	24(92)	26(100)		
3	25(96)	25(96)	26(100)		
4	24(92)	24(92)	26(100)	24(92)	1.0
5	25(96)	24(92)	25(96)	23(88)	1.0
6	**16(62)**	**14(54)**	**22(85)**	**26(100)**	**0.003**
7	24(92)	25(96)	25(96)	26(100)	1.0
8	24(92)	25(96)	25(96)	26(100)	1.0
9	23(88)	24(92)	25(96)	25(96)	1.0
10	23(88)	23(88)	26(100)	26(100)	0.25
11	24(92)	24(92)	24(92)	26(100)	0.5
12	20(77)	21(81)	25(96)	25(96)	0.22
13	25(96)	25(96)	26(100)	26(100)	1.0
14	**17(65)**	**17(65)**	**26(100)**	**24(92)**	**0.016**
15	25(96)	24(92)	25(96)	25(96)	1.0
16	22(85)	22(85)	26(100)	26(100)	0.13
17	**17(65)**	**16(62)**	**25(96)**	**24(92)**	**0.008**
18	24(92)	25(96)	25(96)	26(100)	1.0
19	**14(54)**	**16(62)**	**18(69)**	**24(92)**	**0.0078**
Median (Min, Max) for score	**16.5 (8, 19)**	**17 (7, 19)**	**18 (14, 19)**	**16 (12, 16)**	**<0.0001****

* comparing the paired response of whether the item was present in the instructor-assessment of the original assignment vs. whether the item was present in the exam. The p-values were calculated with McNemar's exact test.

** comparing the number correct from the instructor-assessment of the original assignment (excluding the first three questions) vs. the number correct on the exam. The p-value was calculated with the Wilcoxon signed rank test.

We compared the number (%) responding yes to each question from the self-assessment and the instructor-assessment of the original assignment, but did not find any statistically significant differences. Similarly, the number (%) correct for each question in the instructor-assessment of the student's self-assessment and in the exam yielded no statistically significant differences. Only by assessing the most extreme difference visible in this table (the number (%) correct from the instructor-assessment of the original assignment and the exam) did we observe statistical significance.

4.2 Evaluation of Efficacy Using Historical Controls

45 students enrolled in the course in 2005. Only the 28 students completing the CPR assignment in 2006 were included in the inter-year comparison; the remaining student correctly interpreted the relevant midterm exam questions. In 2005, 10 (22%) of students imperfectly interpreted the estimated coefficients for a linear regression model with a continuous predictor. In 2006, only 1 (3.6%) student provided an imperfect interpretation after completing the CPR module. The difference was statistically significant ($p=0.042$).

4.3 Student Evaluations

The qualitative evaluation at the end of the 2006 course offering showed that students were ambivalent about the CPR system. For the question "How did you like the CPR system" the median response on a scale of 0 to 10 was 5 with the middle 50% of responses falling between 2 and 7. However, when asked the question "How helpful was it to grade your own submission," the median response was 7, and 50% of the responses were between 6 and 8. When each student's response to these two questions was compared, students tended to respond more positively to grading their own text entry submission than to the CPR system as a whole. The median difference was 1 with the middle 50% of responses lying between 1 and 2 points better for grading one's own submission ($p<0.0001$). The CPR module was evaluated for internal reliability with Cronbach's alpha, resulting in a reliability coefficient of 0.70.

4.4 Qualitative Exploration of the Learning Process

In order to assess the learning process within the assignment, students with the minimum, median, and maximum text entry scores were selected in conjunction with a single answer they provided. While the students were chosen objectively, the answers they gave that were selected for evaluation were typical of students at their level within the class. There were several students with the maximum text entry score as graded by the instructor. Eight students received a perfect score from the instructor. The majority of them also gave themselves a perfect score on their self-assessment. One who did not, student A, established a depth of learning not anticipated for this assignment.

Student A

In the tables showing the results for the three selected students (A, B, and C), the following notation has been used, described here for Table 1 only. Three students wrote text entry responses (shown in Table 2) which student A was asked to review both by assessing whether question 19 (amongst others) was fully addressed (yes or no) and by providing overall feedback for the entire assignment (not shown in Table 2). Throughout this section, text written by students is presented exactly as originally written. All the grades shown in Table 1 are provided by student A with the exception of the one for the original text entry, which reflects the instructor assessment.

Table 2. Student A: Assessment of Question 19

Was the direction of the difference in birth weight correctly-specified (babies with >40 weeks gestation have greater birth weight than babies with 38-40 weeks gestation)?

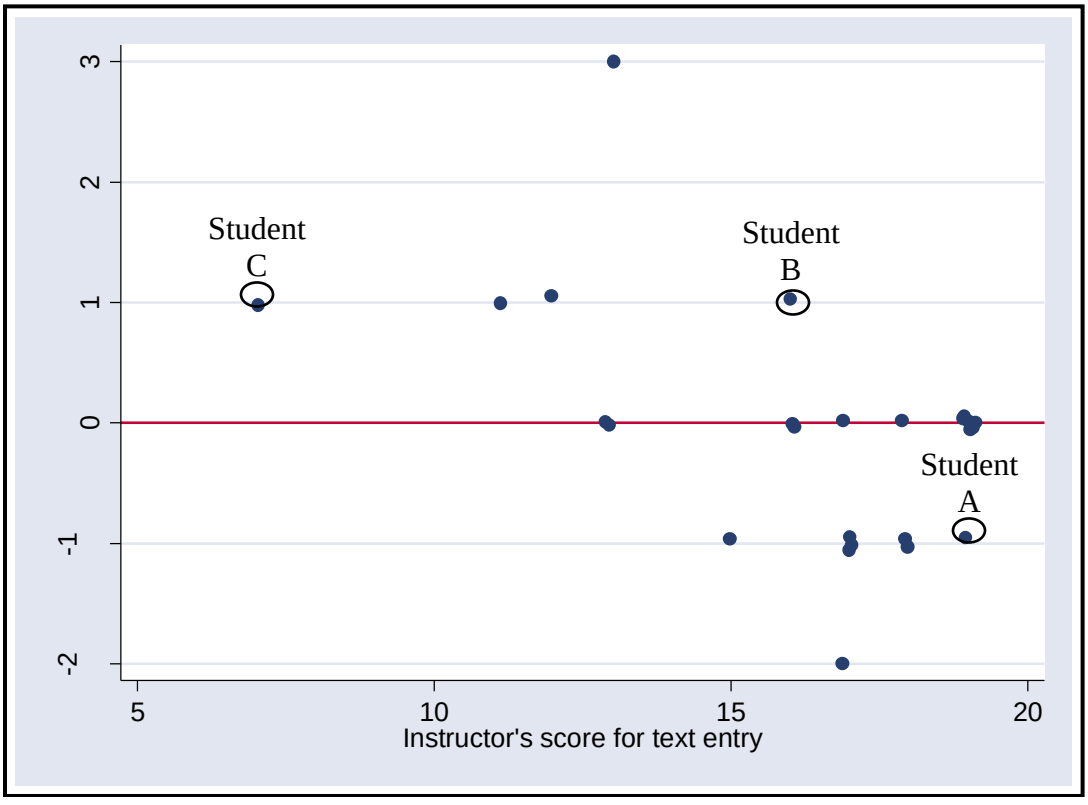
Stage	Text	Grade
Phase I		
Text Entry	The difference in average birthweight between newborns who are older than 40 weeks of gestation and newborns between 38 and 40 wks is 321.56 grams.	*Yes**
Phase II		
Low Calibration	Babies born more than 40 weeks weigh 321.56 more than babies who weigh less than or equal to 40 weeks.	No
Mid Calibration	The estimated average of birth weight for babies with the longest gestation age is 321.56 grams more.	No
High Calibration	The estimated mean birth weight for babies with more than 40 weeks of gestational age is 321.56 grams more than babies with 38 to 40 weeks of gestational age.	Yes
Phase III		
Review of Text Entry 1	b2, 321.56 gm is the difference in estimated mean birth weight between an individual whose gestational age is >40 weeks and one whose gestational age is <38 weeks. We expect the individual with the gestational age >40 weeks to be an average of 321.56 gm heavier than an individual with a gestational age <38 weeks.	No
Review of Text Entry 2	The difference in mean birthweight between children born at greater than 40 weeks gestation and those born between 38 and 40 weeks of gestation is 321.56 grams.	No
Review of Text Entry 3	The average birthweight is 321.56g higher for a child at >40 weeks gestational age compared with a child born at 38-40 weeks gestational age.	Yes
Phase IV		
Self-Assessment	Same as Phase I	No

* Instructor's grade for student A in question 19.

While the instructor graded Q19 as correct because the order of groups implied the direction of association, student A clearly decided this was insufficient during the assignment. This conclusion may have happened during the calibration stage (phase II). All calibrations (low, mid, and high) specified the direction of the association explicitly, though other mistakes were made. Student A's response for the first peer review in phase III is inconclusive with regard to direction, because they incorrectly specified the reference group and these issues are combined in the compound question. However, the response for the second peer review demonstrates that student A had already identified her concern regarding explicitly specifying the direction. The comment student A provided to this student further illuminates her thinking: "Great explanation, only missed the direction in part C B2." In contrast, the student writing the final text entry reviewed by student A explicitly specified the direction by using the word "higher" and was graded as "correct" by student A for this question.

The instructor grade was provided after the assignment ended and was never seen by the students. The CPR assignment also ensures that the feedback from other students is never visible until after the self-assessment (feedback not shown in Table 2). While the first time student A definitively demonstrated improved understanding on this point was in the reply to the second peer review, it is possible that the learning process began as soon as student A saw the grading rubric in phase II. Student A is typical of several stronger students who scored themselves more poorly during the self-assessment than the instructor's score of their text entry (Figure 2).

Figure 2: Plot of the difference between self-assessment scores and instructor-evaluated text entry scores as compared to instructor-evaluated text entry scores



Student A's concern about delineating the direction of the association was itemized three times during the grading rubric (questions 6, 12, and 19). These questions were grouped in a revised grading rubric created based upon the results of this evaluation. The revised grading rubric may be found in the appendix. Additional minor changes were made to the question and calibrations (changes not included).

Student B

Nine (32%) of the students made a mistake on question 14. Student B was one of them (see Table 3). Question 14 strikes at the essence of the difficulty students typically have transferring their knowledge of dummy variables into correct interpretation of regression parameters for dummy variables. However, Question 14 does not discriminate between understanding dummy variables and misinterpreting them as compared to incorrectly grasping parameterization with indicator variables to specify a categorical predictor variable.

Table 3. Student B: Assessment of Question 14

Was b0 interpreted in terms of the correctly-specified reference group (babies with a gestational age between 38 and 40 weeks of gestation)?

Stage	Text	Grade
Phase I		
Text Entry	b0 (3056.07) is the average birth weight in grams for those children whose mothers are in the first category (reference category of gestation week <38 weeks).	*No**
Phase II		
Low Calibration	The overall mean birth weight is 3056.07 grams, on average.	No
Mid Calibration	The estimated mean birth weight for babies with a gestational age between 38 and 40 weeks is estimated to be 3056.07.	Yes
High Calibration	The birth weight for babies with 38 to 40 weeks of gestation is 3056.07 grams.	Yes
Phase III		
Review of Text Entry 1	The estimated birthweight for children born between 38 and 40 weeks of gestation is 3056.07 g.	Yes
Review of Text Entry 2	If a child is born between 38-40 weeks gestation, the expected birth weight is 3056.07 grams.	Yes
Review of Text Entry 3	The mean of birth weight of the children whose gestational age are between 38 to 40 weeks is 3056.07.	Yes
Phase IV		
Self-Assessment	Same as Phase I	No

* Instructor's grade for student B in question 14.

Student B made a mistake common in intermediate and low students choosing the wrong reference group for the categorical variable coded by two dummy indicator variables in this

model. This was the last topic learned prior to the CPR module, so we were not surprised to see so many people with an imperfect understanding of the reference group in the presence of dummy variables. The feedback provided by student B for review 3 correctly identified a similar mistake: "…in part C, interpretation of b2, didn't understand the comparison between reference group (gestational week 38-40) and group with gestational age > 40 weeks (it is the difference between those groups)…."

All three peer reviewers also correctly identified this error in the feedback provided to student B, both in their response to question 14 and in the narrative of their review. Thus even if student B had not learned how to correct this mistake during the assignment, the peer reviewers would have served as one last opportunity for learning.

Student C

The course had already spent a significant amount of time on the concept of interpreting regression parameters as average values prior to this assignment. Consequently, only two students (8%) made the mistake highlighted in Table 4. We chose to explore this mistake because it is such a critical error and one typical of students doing poorly in the course.

Table 4. Student C: Assessment of Question 4
Was b1 interpreted in terms of a "mean", "average", "estimated", "predicted", or "expected" value of the correctly-specified outcome variable (birth weight)?

Stage	Text	Grade
Phase I		
Text Entry	Mothers whose smoking status is 1 compared to those mothers who have a smoking status of zero, will have a birthweight that is 92.5 grams less.	*No**
Phase II		
Low Calibration	The predicted mean birth weight for babies whose mothers did smoke is 92.5 grams less than babies whose mothers did not smoke.	Yes
Mid Calibration	Smoking mothers have babies who weigh 92.5 grams more than if the mother had not smoked, on average.	Yes
High Calibration	b1 is the slope, or the weight for babies born to non-smoking mothers.	No
Phase III		
Review of Text Entry 1	Our best estimate of the difference in mean birth weight for unborn child whose mother's do not smoke verses newborns whose mother's smoke is 92.5 grams.	Yes
Review of Text Entry 2	The difference in the average birth weight of a child whose mother is a smoker compared to the average birth weight of a child whose mother is a nonsmoker is -92.5 grams	Yes
Review of Text Entry 3	The estimated birth weight of a newborn who's mother is smoking is 92.5 grams less than the estimated weight of a newborn who's mother is not smoking is [sic]	Yes
Phase IV		
Self-Assessment	Same as Phase I	No

*Instructor's grade for student C in question 4.

Student C's response to the author of the second text entry was telling: "I think they want you to use "mean" birthweight for every answer. I think they're trying to drill this into our cerebri for some reason." Student C clearly did not understand the reason for using a word such as "mean" or "average" for every interpretation. The CPR module was never intended to serve as a mechanism for teaching the reason for such words, but instead clarifying when and where they should be included in a regression interpretation. That topic had previously been discussed during four lectures and implied by the use of wording such as this in interpretations in eight of the previous ten lectures. Nevertheless, while the assignment "succeeded" in getting student C to use the appropriate words, this comment shows that student C still had a long way to go to achieve appropriate working knowledge. It also highlights that the CPR grading rubric may in some cases improve scores without improving understanding.

Much of the learning process seemed to take place during the assignment. Even the lowest student caught this mistake on the self-assessment. However, in this case the written feedback from other students did not explicitly state the problem with student C's interpretation. This may be due to the fact that the written feedback is intended to provide an overall review, and student

C had made so many errors (scoring 7 out of 19 points) that overall feedback was insufficient. One reviewer used the feedback section to provide correct interpretations for all questions asked.

5. Discussion

5.1 Summary and Discussion of the Results

We compared students' text entry during phase I to free text responses on their exam; both were graded by the instructor with the same rubric. The students showed dramatic improvement in their interpretations in this analysis; however, the lack of a comparison group made that improvement difficult to interpret. In particular, additional course activities also took place between the CPR assignment and the midterm exam (including one homework assignment) so the improvement cannot be entirely attributed to the use of CPR. Our second evaluation focused on historical controls; we assessed questions on interpreting the estimated coefficients for a continuous predictor from the prior course offering's exam according to the usual grading schema (rather than the detailed grading rubric used in the CPR assignment). The risk of imperfectly interpreting a regression parameter was 6.2 times greater before the introduction of the CPR module ($p=0.04$). However, these results are limited by possible confounding due to differences in students' background and preparation, which our data do not allow us to assess.

We also assessed when learning took place for a specific item for three students of differing capabilities at the start of the assignment. All three demonstrated achievement of the goal of this assignment; that they learn to correctly evaluate their written work to identify mistakes. For each student, we were able to qualitatively identify a time during their CPR assignment in which they demonstrated this understanding. In addition, CPR appears to meet the initial conditions set out by Lou et al. (2003) for feedback to be effective in advancing students' understanding. Corrective messages are sent in the form of calibrations, and the grading rubric increases the chance that these corrective messages are interpretable. Student C's work demonstrates, however, that if the student lacks the prior knowledge needed to succeed, full understanding may not be achieved. Motivation for a CPR assignment needs to be provided by the instructor prior to student entry into the system, since students do not all enjoy using the system. However, the CPR system provides at least seven opportunities to solidify learning by using the grading rubric, thus increasing the chance that students will respond appropriately to the feedback by demonstrating learning after completing the assignment.

The use of peer review is always associated with a concern that students may receive low quality feedback because their work is reviewed by other students. In general, a poor CPR reviewer can only have a strong negative impact if 1) 2/3 reviewers provide low-quality or misleading feedback and 2) the student receiving the feedback lacks the understanding to evaluate it appropriately. In practice, well thought out calibrations should minimize this worst-case scenario by training students before they enter phase III of the assignment. This scenario was not observed in this assignment. Instructors using CPR may watch for ≥ 2 low reviewer competency scores among the three reviewers for each student. This information is provided by CPR and could provide an alert cueing further exploration of the submission and reviews.

Like this study, Pritchard (2005) also qualitatively assessed students' perceptions of CPR. Generally, she found that most students felt that they were providing helpful peer reviews, but many did not recognize that their peer review or writing skills improved. In contrast, students in our course did understand that they were learning during the CPR process and thus were better able to review their own work. Pritchard also found that 27% cited "peer review is too much work" as their primary reason for disliking the system. That matched qualitative responses from our students; the results from both the inter-item reliability and the qualitative student feedback suggested that our grading rubric was too long. The original grading rubric used in this module was 19 questions; after this feedback we have created a shortened 11-question version. Since the rubric is used a minimum of seven times for each student, this corresponds to a total of at least 56 fewer questions per student. In order to shorten the rubric, we combined questions so that students are now asked to assess global issues as well as emphasizing critical points with detailed separate questions. Both versions of the grading rubric are shown in the appendix; the shortened version is included in the now-public version of this CPR module (available through the CPR website at http://cpr.molsci.ucla.edu/; Assignment Library; "Interpreting Linear Regression Coefficients: One Predictor").

Lin et al. (2001) found that learners with high executive function performed well regardless of the type of feedback. Student A is a good example of this; however, student A went even farther than the assignment had anticipated in improving understanding of nuances of interpretation. Lin et al. (2001) also found that those with low executive function learned more from specific feedback than general feedback ($p<0.0001$). Similarly, Moreno (2004) found that novice learners learned more from explanatory feedback (which explains the mistake as well as correcting the answer) than from corrective feedback (which simply corrects the mistake; $p=0.005$). We hypothesize that student C exemplifies this. Student C appears to have learned to identify and correct the original mistake from the calibrations and grading rubric. This information was then reinforced by the grading of peers. One advantage of CPR is the use of a detailed, specific grading rubric which provides specific feedback before the learner is even asked to assess their own work. The rubric doesn't directly provide explanatory feedback; however, if a student fails to master a calibration during Phase II, the student is then provided with feedback detailing the reason for answers on the grading rubric for every calibration. These answers act as explanatory feedback for the student who does not yet understand the issues after using the grading rubric to assess the calibrations.

Jacobs et al. (1975) found that reviews were less accurate when the student being reviewed was at a lower level than the peer reviewer. In this study, student C had the poorest text entry; one of the students reviewing student C gave responses that exactly matched the instructor's review, but the second peer reviewer graded more items as incorrect than the instructor did. This may be an example of Jacobs et al's findings. Indeed, we found this to be generally true of reviewers of students with lower initial performance. This is in direct contrast to poor students' grading of their own work during the self assessment phase.

Figure 2 revealed a pattern within students' self-assessment as compared to the instructor's grade for their text entry from phase I. Students with lower scores tended to rate themselves slightly above the instructor's grade, while advanced students tended to rate themselves at or below the grade given by the instructor. We see hints of this pattern among the three students selected for

qualitative assessment. Student A gave herself a lower score than the instructor gave, because she felt that the direction of association should be more clearly specified. Student C provided a self-assessment slightly higher than the instructor's assessment. While poor students tended to rate themselves more highly than the instructor, they did capture the majority of their mistakes. Student C had a score of 7/19 (37%), but correctly identified all but one mistake in the self-assessment. However, student C, while correctly assessing his response to question 4, did not appear to understand the reason for this response. This illustrates that in some cases, the use of a grading rubric may simply lead to emulation rather than understanding. That issue might be less troubling in a CPR assignment targeted to statistical reasoning.

This CPR assignment was focused on gaining statistical literacy. In the prior course offering, many students had completed the course without this basic skill, which comprised a primary learning objective. Consequently, we felt it was worth the time spent on CPR to see gains in literacy. We have since developed other CPR assignments, but so far all have focused on stubborn but essential statistical literacy topics. It is not clear whether CPR is well suited to statistical reasoning topics. We have not attempted these, but we would envision making use of the two types of rubric questions available within CPR (for "style" and "content") to attempt to assess two different levels of questions. As the time required to develop and complete a CPR assignment is substantial, we recommend that only topics that are not otherwise easily learned be considered as candidates for CPR.

5.2 Development and Use of the CPR Module

Those considering developing a CPR module should be aware that the format (see Figure 1) is restrictive. For instance, for many students, two peer reviews might suffice, but this number cannot be modified. Also, feedback is always provided after the self-assessment rather than in advance. In practice, however, organizing many blinded reviews without such a system would be onerous. The CPR system essentially transfers that burden to module development prior to the course.

Module development took several months of weekly meetings. Most of that time was spent on fine-tuning and integrating the questions, the calibrations, and the grading rubric. In addition to linking the question to the grading rubric, the calibrations need to collectively include both major mistakes and typical minor mistakes. The wording on the grading rubric has to fit all possible correct answers. After development, we pilot tested the module to identify problems. Additional wording changes resulted from the pilot but no substantive changes were made.

In writing the CPR assignment, we discovered that the rules for developing a grading rubric allow two different types of questions; "style" questions and "content" questions to be used at the discretion of the author. We chose to use only content questions, since we wanted to focus only on correct wording for the interpretation of estimated regression coefficients. Students then tended to use the final question "How would you rate this text" to provide an overall summary of the quality of writing as well as content. One drawback of this focus on content was that when language was grammatically or structurally deficient, the only way to identify that concern was through the final question. However, we felt the focus of the assignment should remain on the content rather than the quality of the language. During our pretesting we also discovered that

users who fail to master the calibration section are asked to retake the calibrations that they did not master. However, when they retake the calibrations, the CPR system does not show them their prior possibly incorrect responses. Finally, we were concerned that the CPR text entry system requires students to use html. Since many of our students are not comfortable with this programming language, we provided an attachment inside the CPR module which included appropriate html code with space for students to enter their responses within the template. Students were then instructed to cut and paste from the template into the CPR text editor so that their response would be appropriately formatted. Two students did not use the html template, and their peers tended to grade them more harshly than they deserved.

Since this was our first time using a CPR assignment in any course, we felt it was important to introduce students to the CPR system before they began work with the CPR assignment discussed in this paper. To that end, we developed another assignment on discriminating the differences between correlation and linear regression. We were surprised by two aspects of this experience. First, most students found the CPR system very easy to use even on their first entry. Second, this "easy" assignment was quite frustrating, because the text entry did not require that all good responses would include content assessed through the grading rubric. This emphasized for us the critical interlinked nature of the different aspects of the CPR system. The question has to be written so that all issues listed on the grading rubric will be answered by every competent student.

We were concerned that students would find the CPR system confusing. As a result, we began the assignment during a computer lab so that the instructor and teaching assistants would be on hand to cope with any problems. Although the CPR system itself turned out to be much easier for students to use than we had anticipated, beginning the assignment during a lab also streamlined any difficulties students had with access. Student profiles were set up in advance using student ID numbers, and a CPR ID is assigned for each student ID. The in-person laboratory session greatly facilitated the transmission of this unique information to each student.

We were apprehensive about the timing of the assignment phases, which is normally quite restrictive so that students cannot begin phase II until everyone in the class finishes phase I. We learned that one could trick the CPR system into being more user friendly by changing the timing settings. If the end of phase 1 has passed, the instructor can manually extend students' time to complete these phases. In practice, these students are then able to do phases 1-4 in a continuous sequence without stopping. We have now adopted this as standard practice for our CPR assignments with the modification that several answers from previous years' students are entered into the system so that the first student to enter the system will still have the experience of grading peers' assignments. This has the drawback of making it possible for a student to see the grading rubric (on another student's screen) before completing text entry, but we felt that was the lesser of two evils.

Overall, we found that participation in this CPR assignment improved students' detailed understanding of expectations for linear regression interpretations. Those students who were previously making critical errors, such as failing to describe the estimated coefficient as an average value or citing an incorrect reference group, quickly learned to correct these errors and were able to identify the mistake in their own submission. Moreover, stronger students gained

improved nuance of expression. The rate of mistakes during the midterm examination on interpretation in the presence of a continuous predictor variable was dramatically reduced from the prior year's course offering, thus achieving the objective of the authors. CPR can be successfully used to reinforce correct usage of complex ideas in free text responses. CPR thus has the potential to be of great benefit to the statistics education community as a tool to improve outcomes in the second course in statistics.

Appendix

The Calibrated Peer Review Module

"Interpreting Linear Regression Coefficients: One Predictor**"**

Student Instructions

The results from three linear regression models (using birthweight as the outcome) for each question (a, b, and c) are given below. Your Calibrated Peer Review (CPR) assignment is to interpret all of the coefficients in each of these models.

This CPR assignment:
• Should not be submitted with your regular written assignment
• Should be written in Word and submitted on the CPR website
• Should be submitted no later than October 18, 2006 at midnight
• The review and grading period for this CPR assignment is from October 19 to October 24
• As you review and grade the different assignment solutions, pay careful attention to which part (A, B, or C) and which coefficient (b0, b1, b2) to which a particular question refers.
• If you think ANY part of a question is incorrect in the assignment you are grading, then the whole question should be considered incorrect (mark "no").

In the original data file, the variables are:

- **birth_weight:** child's weight at birth in grams
- **gestation:** child's gestation in weeks
- **smoking_status:** Smoking status of the mother (1=smoker, 0=nonsmoker)

(A) Interpret both of the coefficients in the following model that predicts birthweight using mother's smoking status:
\hat{y}_i = b_0 + b_1 (smoking mom_i)
\hat{y}_i = 3066.13 – 92.5 (smoking mom_i)

(B) Interpret both of the coefficients in the following regression model that predicts birthweight using child's gestation in weeks (as a continuous variable):
\hat{y}_i = b_0 + b_1 (gestation in weeks_i)
\hat{y}_i = -2037.01 + 130.82 (gestation in weeks_i)

(C) Interpret only b_0 and b_2 in the following regression model that predict birthweight using child's gestation in weeks (as a categorical variable: <38, 38-40, or over 40):
\hat{y}_i = b_0 + b_1 (gest age<38_i) + b_2 (gest age>40_i)
\hat{y}_i = 3056.07 - 414.51 (gest age<38_i) + 321.56 (gest age>40_i)

High Quality Calibration

Part A response:

- Interpretation of b0: The estimated mean birth weight for babies whose mothers did not smoke is 3066.13.

- Interpretation of b1: The predicted mean birth weight for babies whose mothers did smoke is 92.5 grams less than babies whose mothers did not smoke.

Part B response:

- Interpretation of b0: The estimated mean birth weight for babies with zero weeks of gestation is -2037.01 grams (does not make sense).
- Interpretation of b1: The difference in mean birth weight for 2 babies whose gestation times differ by 1 week is 130.82 grams, where babies with longer gestation time have a greater mean birth weight.

Part C response:

- Interpretation of b0: The estimated mean birth weight for babies with a gestational age between 38 and 40 weeks is estimated to be 3056.07.
- Interpretation of b2: The estimated mean birth weight for babies with more than 40 weeks of gestational age is 321.56 grams more than babies with 38 to 40 weeks of gestational age.

Mid Quality Calibration

Part A response:

- Interpretation of b0: 3066.13 is the average weight for babies whose mothers did not smoke.
- Interpretation of b1: Smoking mothers have babies who weigh 92.5 grams more than if the mother had not smoked, on average.

Part B response:

- Interpretation of b0: -2037.01 grams is the overall mean birth weight.
- Interpretation of b1: 130.82 grams is the difference in birth weight for each 1 unit increase in gestational age.

Part C response:

- Interpretation of b0: The birth weight for babies with 38 to 40 weeks of gestation is 3056.07 grams.
- Interpretation of b2: The estimated average of birth weight for babies with the longest gestation age is 321.56 grams more.

Low Quality Calibration

Part A response:

- Interpretation of b0: b0 is the y-intercept of the line, or the average value of birth weight in grams for a baby born to a mother who smokes.
- Interpretation of b1: b1 is the slope, or the weight for babies born to non-smoking mothers.

Part B response:

- Interpretation of b0: A baby born at zero grams is expected to have had gestation of -2037.01 weeks (note: does not make sense since negative).

- Interpretation of b1: If one baby has one gram greater birth weight than another, we predict that the larger baby will have about 131 weeks longer gestation.

Part C response:

- Interpretation of b0: The overall mean birth weight is 3056.07 grams, on average.
- Interpretation of b2: Babies born more than 40 weeks weigh 321.56 more than babies who weigh less than or equal to 40 weeks.

Original Grading Rubric with Correct Calibration Responses

Question	**High**	**Mid**	**Low**
1. Part A (b0): Was b0 interpreted in terms of a "mean", "average", "estimated", "predicted", or "expected" value of the correctly-specified outcome variable (birth weight)?	Yes	Yes	Yes
2. Part A (b0): Was b0 interpreted in terms of the correctly-specified reference group (mothers who do not smoke)?	Yes	Yes	No
3. Part A (b0): Was the value correctly given (3066.13) in the units of the outcome variable (grams)?	No	No	No
4. Part A (b1): Was b1 interpreted in terms of a “mean”, “average”, "estimated", "predicted", or “expected” value of the correctly-specified outcome variable (birth weight)?	Yes	Yes	No
5. Part A (b1): Was b1 interpreted in terms of the difference between mothers who do smoke as compared to mothers who do not smoke?	Yes	Yes	No
6. Part A (b1): Was this value correctly provided (92.5) in the correct units of the outcome variable (grams), AND was the direction correctly specified (mothers who do not smoke have higher birth weight babies than mothers who do smoke)?	Yes	No	No
7. Part B (b0): Was b0 interpreted in terms of a “mean”, “average”, "estimated", "predicted", or “expected” value of the correctly-specified outcome variable (birth weight)?	Yes	Yes	No
8. Part B (b0): Was b0 interpreted in terms of babies with zero weeks of gestation?	Yes	No	No
9. Part B (b0): Was the value correctly given (-2037.01) in the correct units of the outcome variable (grams)?	Yes	Yes	No
10. Part B (b1): Was b1 interpreted in terms of “means”, “averages”, "estimated", "predicted", or “expected” values of the correctly-specified outcome variable (birth weight)?	Yes	No	No
11. Part B (b1): Was the value interpreted as a "difference between 2 babies whose gestation time differ by 1 week” or as a “change in birthweight for a 1 week change in gestation time"?	Yes	No	No
12. Part B (b1): Was the value correctly specified (130.82) in the units of the outcome variable (grams), AND was the direction correctly specified (babies with longer gestation times have a greater estimated mean birth weight)?	Yes	No	No
13. Part C (b0): Was b0 interpreted in terms of a “mean”, “average”, "estimated", "predicted", or “expected” value of the correctly-specified outcome variable (birth weight)?	Yes	No	Yes

14. Part C (b0): Was b0 interpreted in terms of the correctly-specified reference group (babies with a gestational age between 38 and 40 weeks of gestation)?	Yes	Yes	No
15. Part C (b0): Were both the value AND the units correctly given (3056.07 grams) in terms of the correctly-specified outcome variable (birth weight)?	No	Yes	Yes
16. Part C (b2): Was b2 interpreted in terms of a "mean", "average", "estimated", "predicted", or "expected" value of the correctly-specified outcome variable (birth weight)?	No	Yes	No
17. Part C (b2): Was b2 interpreted as comparing babies with more than 40 weeks of gestation versus babies with between 38 and 40 weeks of gestation?	Yes	No	No
18. Part C (b2): Was the value provided correctly (321.56) in the units of the outcome variable (grams)?	Yes	Yes	No
19. Part C (b2): Was the direction of the difference in birth weight correctly-specified (babies with >40 weeks gestation have greater birth weight than babies with 38-40 weeks gestation)?	Yes	No	No

Revised Grading Rubric with Correct Calibration Responses

Question	**High**	**Mid**	**Low**
1. Was each coefficient throughout the entire assignment interpreted in terms of a "mean", "average", "estimated", "predicted", or "expected" value of the outcome variable (birthweight)?	No	No	No
2. Throughout the entire assignment, was b0 interpreted in terms of the correct reference group (part A: mothers who do not smoke, part B: babies with zero weeks of gestation, and part C: babies with gestational age between 38-40 weeks)?	Yes	No	No
3. Part A (b0): Was the value correctly given (3066.1) in the units of the outcome variable (grams)?	No	No	No
4. Part A (b1): Was b1 interpreted in terms of the difference between mothers who do smoke as compared to mothers who do not smoke?	Yes	Yes	No
5. Part A (b1): Were the value and direction correctly provided in the units of the outcome variable (non-smoking mothers have babies who weigh 92.5 grams more than smoking mothers, on average)?	Yes	No	No
6. Part B (b0): Was the value correctly given (-2037.0) in the units of the outcome variable (grams)?	Yes	Yes	No
7. Part B (b1): Was the value interpreted as a "difference between 2 babies whose gestation time differ by 1 week" or as a "change in birthweight for a 1 week change in gestation time"?	Yes	No	No
8. Part B (b1): Were the value and direction correctly provided in the units of the outcome variable (babies with 1 week longer gestational time are estimated to be 130.8 grams heavier at birth)?	Yes	No	No
9. Part C (b0): Was the value correctly given (3056.1 grams) in the units of the outcome variable (grams)?	No	Yes	Yes

10. Part C (b2): Was b2 interpreted as a comparison of babies with more than 40 weeks of gestation versus babies with between 38 and 40 weeks of gestation?	Yes	No	No
11. Part C (b2): Were the value and direction correctly provided in the units of the outcome variable (babies with >40 weeks of gestation are predicted to be 321.6 grams heavier at birth than babies with 38-40 weeks of gestation)?	Yes	No	No

Acknowledgement

Part of this manuscript was presented at the Joint Statistics Meetings in Salt Lake City, UT, August 2007.

References

Berwick, D.M., Fineberg, H.V., and Weinstein M.C. (1981). "When doctors meet numbers." *American Journal of Medicine*, 71(6):991-998.

Boud, D. (1990). "Assessment and the Promotion of Academic Values." *Studies in Higher Education*. 15(1):101-111.

Carlson, P.A. and Berry, F.C. (2003). "Calibrated Peer Review and Assessing Learning Outcomes." Paper presented at the American Society for Engineering Education (ASEE) / Institute of Electrical and Electronics Engineers (IEEE) Frontiers in Education Conference, Boulder, CO.

Carlson, P.A. and Berry, F.C. (2005). "Calibrated Peer Review: A Tool for Assessing the Process as well as the Product in Learning Outcomes." Paper presented at the Annual Conference and Exposition: The Changing Landscape of Engineering and Technology Education in a Global World, Portland, OR.

Davies, P. (2006). "Peer Assessment: Judging the Quality of Students' Work by Comments Rather than Marks." *Innovations in Education and Teaching International*. 43(1):69-82.

delMas, Robert C. (2002). "Statistical Literacy, Reasoning, and Learning: A Commentary." *Journal of Statistics Education* 10(3)
http://www.amstat.org/publications/jse/v10n3/delmas_discussion.html

delMas, R., Garfield, J., Ooms, A., and Chance, B. (2006). "Assessing Students' Conceptual Understanding After a First Course in Statistics," presented at the Annual Meetings of The American Educational Research Association, San Francisco, CA April 9, 2006
https://app.gen.umn.edu/artist/articles/AERA_2006_CAOS.pdf

delMas, R., Garfield, J., Ooms, A., and Chance, B. (2007). "Assessing Students' Conceptual Understanding After a First Course in Statistics," *Statistics Education Research Journal* 2007 6(2):28-58. http://www.stat.auckland.ac.nz/~iase/serj/SERJ6(2)_delMas.pdf

Falchikov, N. and Goldfinch, J. (2000). "Student Peer Assessment in Higher Education: A Meta-Analysis Comparing Peer and Teacher Marks." *Review of Educational Research*. 70(3):287-322.

Garfield, J. (2002). "The Challenge of Developing Statistical Reasoning," *Journal of Statistics Education*, 10(3). http://www.amstat.org/publications/jse/v10n3/garfield.html

Gunersel, A.B. and Simpson, N. (2009). "Improving Writing and Reviewing Skills with Calibrated Peer Review." *International Journal for the Scholarship of Teaching and Learning.* (3)2:1-9. academics.georgiasouthern.edu/ijsotl/v3n2/articles/_GunerselSimpson/index.htm

Hartberg, Y., Gunersel, A.B., Simspon, N.J., and Balester, V. (2008). "Development of Student Writing in Biochemistry Using Calibrated Peer Review." *Journal of the Scholarship of Teaching and Learning.* 2(1):29-44.

Heise, E.A., Palmer-Julson, A., and Su, T.M. (2002). "Calibrated Peer Review Writing Assignments for Introductory Geology Courses." Presented at the Science at the Highest Level Conference, Denver, CO.

Jacobs, R.M., Briggs, D.H., and Whitney, D.R. (1975). "Continuous-Progress Education: III. Student self-evaluation and peer evaluation." *Journal of Dental Education*. 39(8):535-541.

Kaplan, J.J., Fisher, D.G., and Rogness, N.T. (2008). "Lexical Ambiguities in Statistics." Presented at the Joint Statistical Meetings, Denver, CO.

Kim, S.H., Wise, J., and Hillsey, M. (2005). "Learning Technical Writing Skills through Peer Review: Use of Calibrated Peer Review in Unit Operation Lab." Presented at the Free Forum on Engineering Education, Cincinnati, OH.

Lepper, M.R. and Chabay, R.W. (1985). "Intrinsic Motivation and Instruction: Conflicting Views on the Role of Motivational Processes in Computer-Based Education." *Educational Psychologist.* 20(4):217-230.

Lin, S.S.J., Liu, E.Z.F., and Yuan, S.M. (2001). "Web-Based Peer Assessment: Feedback for Learners with Various Thinking-Styles." *Journal of Computer Assisted Learning.* 17:420-432.

Lou, Y., Dedic, H., and Rosenfield, S. (2003). "A Feedback Model and Successful E-Learning." In Learning and Teaching with Technology: Principles and Practices, ed. Naidu S, Kogan Page, London and Sterling, VA.

MacFarlane, G., Markwell, K., and Date-Huxtable, L. (2005). "Encouraging Students to 'Think as Biologists': Independent Field-Based Projects and Peer Assessment as a Deep Learning Strategy." Presented at the Uniserve Science Conference, Sydney, Australia.

McCarty, T., Parkes, M.V., Anderson, T.T., Mines, J., Skipper, B.J., and Grebosky, J. (2005). "Improved Patient Notes from Medical Learners during Web-Based Teaching Using Faculty-Calibrated Peer Review and Self-Assessment." *Academic Medicine.* 80(10):S67-S70.

Moreno, R. (2004). "Decreasing Cognitive Load for Novice Learners: Effects of Explanatory versus Corrective Feedback in Discovery-Based Multimedia. *Instructional Science.* 32:99-113.

Orsemond, P., Merry, S., and Callaghan, A. (2004). "Implementation of a Formative Assessment Model Incorporating Peer and Self-Assessment." *Innovations in Education and Teaching International.* 41(3):273-290.

Pope, N.K. (2005). "The Impact of Stress in Self- and Peer Assessment." *Assessment and Evaluation in Higher Education.* 30(1):51-63.

Prichard, J.R. (2005). "Writing to Learn: An Evaluation of the Calibrated Peer Review Program in Two Neuroscience Courses." *The Journal of Undergraduate Neuroscience Education.* 4(1):A34-A39.

Pridemore, D.R. and Klein, J.D. (1995). "Control of Practice and Level of Feedback in Computer-Based Instruction." *Contemporary Educational Psychology.* 20:444-450.

Reese-Durham, N. (2005). "Peer Evaluation as an Active Learning Technique." *Journal of Instructive Psychology.* 32(4):338-343.

Rumsey, D. J. (2002). "Statistical Literacy as a Goal for Introductory Statistics Courses" *Journal of Statistics Education* 10(3). http://www.amstat.org/publications/jse/v10n3/rumsey2.html

Topping, K. and Ehly, S., Eds. (1998). Peer-Assisted Learning. Lawrence Erlbaum Associates, Inc., Mahwah, NJ.

Trautmann, N., Carlsen, W., Yalvac, B., Cakir, M., and Kohl, C. (2003). "Learning Nature of Science Concepts through Online Peer Review of Student Research Reports." Presented at the annual meeting of the National Association for Research in Science Teaching, Philadelphia, PA.

Van de Ridder, J.M.M., Stokking, K.M., McGaghie, W.C., and ten Cate, O.T.J. (2008). "What is Feedback in Clinical Education?" *Medical Education.* 42:189-197.

Van den Berg, I., Admiraal, W., and Pilot, A. (2006). "Design Principles and Outcomes of Peer Assessment in Higher Education." *Studies in Higher Education.* 31(3):341-356.

Weiss, S.T. and Samet, J.M. (1980). An assessment of physician knowledge of epidemiology and biostatistics. *Journal of Medical Education,* 55(8):692-697.

Windish, D.M., Huot, S.J., and Green, M.L. (2007). "Medicine Residents' Understanding of the Biostatistics and Results in the Medical Literature." *The Journal of the American Medical Association.* 298(9):1010-1022.

Wise, J.C. and Kim, S. (2004). "Better Understanding Through Writing: Investigating Calibrated Peer Review." Presented at the Annual Conference and Exposition: Engineering Education Reaches New Heights, Salt Lake City, UT.

Wulff, H.R., Andersen, B., Brandenhoff, P., and Guttler, F. (1987). "What do doctors know about statistics?" *Statistics in Medicine,* 6(1):3-10.

Felicity B. Enders, PhD, MPH
Division of Biomedical Statistics & Informatics
Department of Health Sciences Research
Mayo Clinic
200 First Street, SW
Rochester, MN 55905
Enders.Felicity@gmail.com
507-538-4970 (phone)
507-284-9542 (fax)

Sarah Jenkins, MS
Division of Biomedical Statistics & Informatics
Department of Health Sciences Research
Mayo Clinic
200 First Street, SW
Rochester, MN 55905
mailto:Jenkins.Sarah@Mayo.edu

Verna Hoverman
Division of Biomedical Statistics & Informatics
Department of Health Sciences Research
Mayo Clinic
200 First Street, SW
Rochester, MN 55905
mailto:Hoverman@Mayo.edu
