Using Assessment and Feedback to Enhance Learning: Examining the Relationship Between Teachers' Reported Use of Assessment and Feedback and Student Performance in AP Biology

Eva Ponte
University of Hawaii at Manoa

Pamela Paek
Center for Assessment

Henry Braun
Boston College

Catherine Trapani, Don Powers
Educational Testing Service

ABSTRACT: This paper analyzes a national sample of teachers' self-reported use of assessment and feedback in Advanced Placement® (AP) biology classrooms. Descriptive statistics of what teachers reported doing in these two areas are discussed, followed by the identification of reported teaching practices variables that were found to be significantly related with student performance on AP biology exams. All the significant variables found are exemplars of authentic assessment and feedback practices and techniques. The study provides valuable information on the types of assessment and feedback practices currently being employed by teachers, and shows the relationship of such practices with actual student outcomes.

KEYWORDS: *assessment; feedback (response); teacher effectiveness; science instruction; advanced placement programs; regression (statistics)*

While research has established that teachers have a measurable effect on student achievement, it has proven much more difficult to pinpoint the specific characteristics of teachers or aspects of their pedagogy that can be linked to higher student achievement (Olson, 2003). This paper reports on our initial attempts to address this issue by examining the relationship of Advanced Placement® (AP) teacher assessment and feedback practices on AP students' outcomes. To do that, we make use of results obtained from a large-scale national study of teaching practices in AP biology.

We believe our efforts are now more relevant than ever; current legislation such as the No Child Left Behind (NCLB) Act requires more accountability for the achievement—or lack thereof—of our nation's students, and teachers have become the center of such accountability. By examining specific teacher practices in relation to students' outcomes, our study attempts to focus on the basis for accountability. One focus of this national legislation is the preparation of a quality teaching force that will provide students with the best education possible. This argument carries the expectation that improvements in the professional development of teachers will promote positive changes in teaching practices, which will in turn enhance student performance and achievement (Cochran-Smith & Fries, 2005). As educational researchers, we are also keenly interested in identifying key elements of this relationship (i.e., teachers' practices and students' outcomes) than can be the basis for much more detailed and extended study.

We surveyed a national sample of AP biology teachers about their teaching practices, asking them which practices they used and how often. The work discussed in this paper focuses specifically on the areas of assessment and feedback. We take as our underlying assumption that teachers must have a better understanding of the types of assessment and feedback available to them in order to achieve ongoing efforts to improve instruction—and, consequently improve student learning (Fullan, Hill, & Crévola, 2006; National Education Association, 2003; Popham, 2008). This assumption frames our survey instruments in that we explore pedagogical practices aligned with it. We discuss the descriptive statistics of what teachers reported doing in the areas of assessment and feedback, followed by an analysis relating these practices with student performance on an AP biology exam.

Theoretical Framework

Research has shown that substantial learning gains are possible when teachers introduce alternative forms of assessments, including formative assessment, into their classroom practice (Black & Wiliam, 1998). Those gains have been demonstrated even in studies where attainment is measured with scores on external tests—as opposed to classroom assessment outcomes (Newmann, Bryk, & Nagaoka, 2001). Studies such as these are useful in pointing out that attention to higher-order goals in teaching can result in higher attainment. However, if we are to draw more direct implications about the utility of alternative assessments, we need to conduct direct studies of the relationship of teachers' use of alternative assessments and student performance where confounding variables—such as overall good teaching—are controlled (Wiliam, Lee, Harrison, & Black, 2004). Below, we offer a brief review of some of those alternative forms of assessment and feedback that have been presented in the literature as more successful in enhancing students' learning and achievement.

Assessment

Assessment is the process of collecting, synthesizing, and interpreting information to aid in educational decision making (Airaisan, 2000). Teachers' use of assessment techniques reflects their views of both appropriate pedagogy and how students learn. For instance, behaviorist approaches to learning tend to consider assessment as a tool to measure how much information students have acquired–assessments *of* learning (Gagne, 1965; Resnick & Resnick, 1992; Shepard, 2001), thus defining assessment as an external activity aimed at grading and sorting students in relation to their acquisition of a specific set of facts and knowledge (Gitomer & Duschl, 1995; Shepard, 2001). The underlying assumption is that there is a clear set of knowledge goals that students ought to master by a certain point in time. As a result, assessment is usually structured in a closed format that produces an indicator of whether or not a student has mastered the material. Usually, this indicator does not provide much information about the processes utilized by students to give answers and thus precludes the teacher from targeting individual student needs (Wolf, Bixby, Glenn & Gardner, 1991).

In contrast, constructivist and sociocultural approaches to learning view assessment as a tool that should inform both student learning and teacher practices—assessments *for* learning (Black, Harrison, Lee, Marshall & Wiliam, 2003; Black & Wiliam, 1998; Brookhart; 2005; Darling-Hammond & Ancess, 1996; Leahy, Lyon, Thompson, & Wiliam, 2005; Wiliam, Lee, Harrison, & Black, 2004). To achieve that goal, assessment is not seen as an external event, but as an embedded classroom activity that is systematically and continuously included in classroom procedures (Black et al., 2003; Darling-Hammond et al., 1994; Greeno, Collins, & Resnick, 1996; Leahy et al., 2005; Shepard, 2001; Wilson & Sloane, 2000). Learning is defined as an effort by students to create meaning out of different things. Therefore, in this approach students are located at the center of the assessment process and are expected to (a) reflect upon their own learning, (b) become familiar and collaborate in the creation of criteria to judge performance, and (c) evaluate their peer's work (Ainsworth & Christinson, 1998; Black et al., 2003; Leahy et al., 2005; McNamara & Deane, 1995; Popham, 2008; Smolen, Newman, Wathen & Lee, 1995; Valencia & Place, 1994). In sum, authentic or alternative forms of assessment attempt to gauge students' learning over time and to embed that assessment in an authentic context, thus providing both the teacher and the students with information about what students know and are able to do (Darling-Hammond, Ancess, & Falk, 1995).

Feedback

Traditional assessment and feedback approaches tend to use close-ended exams such as multiple-choice items. As a result, the amount of information they provide is limited as to whether a student response was correct or not (Broadfoot & Black, 2004; Gitomer & Duschl, 1995; Stiggins, 2002). This type of feedback may be useful for certain situations, but if it is the only feedback available to students, it will eventually be proven as insufficient to successfully guide students' process of learning. Contrary to this, constructivist approaches generate more elaborated feedback, which ideally (a) provides specific evidence of quality of performance, (b) gives information to students and teachers about the progress being made, and (c) uses descriptive language that can be utilized by the teacher to plan instruction accordingly and by the student to improve their learning through self-assessment and self-adjustment (Black et al, 2003; Nicol & Macfarlane-Dick, 2006; Popham, 2008; Shepard, 2000; Shepard, Hammerness, Darling-Hammond, & Rust, 2005; Wiggins, 1998). From a sociocultural perspective, feedback may be interpreted as an essential component of the scaffolding process, with three types of feedback: (a) pointing out the distinction between a child's performance and the ideal, (b) attributing success to effort to encourage academically supportive attributions, and (c) restating the concept that has been learned (Hogan & Pressley, 1997). In this paper, we define feedback as a process with the potential to provide information that moves learners forward and activate instructional resources from both the teacher and the students (Wiliam & Thompson, 2007).

Based on the tenets about assessment and feedback discussed above, our model hypothesizes that teachers' use of more authentic forms of assessment and feedback is likely to lead to improved instruction and learning, and these enhancements would in turn result in higher student performance on the test.

Methods

We sent out 1,874 surveys to a nationally representative sample of AP biology teachers and 1,171 were returned (representing a 62 percent response rate). We were able to run descriptive statistics of teacher practices for all 1,171 surveys. However, to relate teacher data with student performance, we were not able to use all surveys because there were no unique identifiers for teachers (linked to their students) in schools with more than one AP biology teacher. As such, we had to eliminate surveys for which we could not link specific AP teachers with their students' AP exam scores. In addition, after some exploratory analysis, we decided to drop from the study classes with fewer than eight students with usable data to assure reasonable stability of class averages. Our final sample of 667 teachers in the analyses linking teacher and student data included 473 public school teachers and 194 nonpublic school teachers. We distinguish the two types of schools, knowing that public and nonpublic (private and parochial) may have different demographics of students and teachers as well as resources and that such differences need to be accounted for in our analysis.

Demographics

The teachers in our sample of 1,171 teachers tended to be veteran teachers (i.e., 71 percent of them have taught for more than 10 years, and only 6 percent have taught for fewer than 3 years). A survey by NCES on public and private school teachers indicates that only about 61 percent of public and 52 percent of nonpublic school teachers have taught for 10 years or more (NCES, 2002). These data suggest that the AP teachers in our sample tend to have a higher level of teaching experience than the average U.S. teacher.

With respect to background, AP biology teachers in our sample tend to have a high degree of preparation: While many of them (85%) had obtained at least a master's degree, national data indicate that only 47 percent of teachers have a master's or a higher degree (NCES, 2006). Similarly, while the vast majority (84%) of AP biology teachers had a degree in biology and held a regular or standard state certificate degree, just 59 percent of U.S. high school teachers attained this level of preparation (major and certification) to teach biology (NCES, 2006). This finding indicates a higher degree of preparation of this sample when compared to U.S. public school teachers.

Demographic data indicate that most AP biology teachers in this survey are 36 years old or older (79%), Caucasian (93%), and female (56%). When compared with national trends, we see that minorities are less represented in our sample than in comparable U.S. teacher populations, with African Americans and Latino/Hispanics being the most underrepresented in this sample. In addition, there were more males in our sample (44%) than in the current population of secondary school level teachers (35%) (NCES, 2002).

Survey Constructs and Development

Research on teaching continues to grow, as evidenced by a significant increase in the number of articles from the third to the fourth edition of the American Educational Research Association's *Handbook of Research on Teaching*. One important area of this research focuses on factors that influence teacher's effectiveness, as determined by the performance of his or her students. Researchers have shown that there is no single factor that determines teacher effectiveness. Rather, numerous factors govern how successful teachers approach the challenges of teaching (Darling-Hammond, 2002; Marzano, 2003; Porter & Brophy, 1998; Stigler & Hiebert, 1998, 1999; Wenglinsky, 2002). Based on this research, we selected those factors that we deemed most relevant in AP biology courses and that were measurable through a survey. We created a model of teaching practices and the contexts in which teachers work that served as the basis for the creation of the surveys. A detailed discussion of the model and the literature from which it was formulated can be found in Paek, Ponte, Sigel, Braun, and Powers (2005).

Most items in the survey were five-point rating scale items, with a few four-point scales, and one open-ended question. Before the final version was administered to the large sample in spring 2003, the survey was reviewed by three AP biology teachers' focus groups in California, Virginia, and Washington, D.C. and was pilot-tested with 127 AP biology teachers across the country.

A number of factors influencing teachers' effectiveness that are especially prominent in the research literature on teaching were included. These factors can be grouped into five subcategories: (1) substantive expertise and training, (2) school context, (3) classroom context, (4) instructional, assessment, and feedback practices, and (5) content coverage. In addition to these general factors, our view of teaching practices also incorporates one factor that is specifically relevant for this AP study: AP exam preparation practices. The first three factors are considered *context*-related (expertise and training, school context, and classroom context) and are presented first. These are then followed by *teaching* practice, which incorporates instructional, assessment, feedback, content coverage, and AP Exam preparation practices. A brief description of these categories and factors follows.

Dimension 1: Factors Affecting Teachers' Practices—Context Related

Substantive expertise and training. Substantive expertise and training refers to the teacher's experience with the content of the given course. This is a product of numerous factors, such as the educational background of the teacher (including educational level, major, and teaching certification), previous experience teaching courses in this subject area (AP and otherwise), and the teacher's ongoing professional development through workshops, institutes, university classes, and seminars.

School context. School context refers to the nature of the learning environment. It measures a variety of matters related to how the school context provides or does not provide a positive setting for teaching and learning. For instance, this factor provides data for scheduling, the amount of classes and prep time that the teacher has during the day, and the amount of influence that the teacher has in organizing her/his AP class.

Classroom context. Classroom context describes the factors that affect the composition and organization of the classroom, such as the class size.

Dimension 2: Factors Affecting Teachers' Practices—Teaching

Instructional, assessment, and feedback practices. Instructional and assessment practice (teachers' pedagogical practice) may be manifested through the nature of assignments (e.g., how students are configured for in-class and out-of-class activities and assignments). It also concerns the relative role of various styles of instructional delivery that teachers use in their courses. In addition to measuring the teachers' decisions about how to deliver instruction, this factor also reflects the emphasis that the teachers place on various "types" of knowledge realized through their courses. This factor also deals with the ways in which teachers assess students' understanding and provide feedback to students based on those assessments, as well as teachers' use of technology in the classroom. Finally, it covers issues directly related to instructional practice that do not usually take place during instruction, such as teachers' preparation time and students' homework load.

Content coverage. Content coverage addresses the manner in which teachers cover the materials included in the AP course. The first issue addressed is how depth of course concepts is negotiated relative to breadth of course content. Second, teachers report the specific topics and themes that they find more relevant, and thus tend to place more emphasis on, and rank those topics/themes regarding the degree of difficulty students have learning them. Last, content coverage refers to the extent to which the content of the AP class under examination is aligned with the content of the AP Exam.

Test-specific instructional activities and practices. Test-specific instructional activities and practices refer to the instructional activities and pedagogical practices that the teacher uses specifically because he or she is teaching an AP class. This factor addresses teachers' instructional decisions, both inside and outside of class time, related to getting students ready to take and pass the AP Exam. It accounts both for activities, such as after-school review sessions, as well as pedagogical decisions, such as using AP practice tests to familiarize students with the AP exam. It also considers the extent to which the teachers encourage or require students to participate in extracurricular activities, such as districtwide competitions, inasmuch as these activities relate to gaining knowledge about course content and preparing for the AP exam.

Demographic data. This is a source of information not covered by the factors listed above, but used in different ways to gain insight into teacher backgrounds, like teachers' age, race/ethnicity, and gender.

Variables

In the AP Teacher Practices surveys, we aimed to create items that included several practices representative of what little is known about AP teachers' practices, while also including specific practices that we hypothesized to be more effective in terms of enhancing student performance (based on the literature review and model discussed earlier). Within each category, several variables were included. Table 1 below shows the variables we introduce in this study across the two main categories.

Table 1
Overview of Variables Included in the Study

Context for teaching variables	Teaching practice variables
1. Teacher variables related to professional experience (five individual variables, including years of teaching experience, years of AP teaching experience, educational level, major, and teaching certificate)	1. Learning goals (six variables, based on the AP syllabi)
2. Professional development (seventeen variables dealing with participation in specific AP professional development activities [PAPPD])	2. Instructional methods (eight variables, including teacher- and student-centered approaches)
3. School support (six variables, including school policy for assigning teachers to AP classes, number of classes taught, number of students per class, type of teaching schedule, adequacy of different school resources, and influence of AP resources)	3. Assessments (six variables, from more traditional to more authentic assessments).
4. Classroom control (seven variables, including hours dedicated to prepare for AP class, teaching autonomy, school criteria for AP enrollment, school policies determining which students take the AP exam, percentage of students who take the AP Exam, teaching freedom, and class size)	4. Feedback (six variables, from limited (e.g., grades) to more descriptive feedback).
5. Hours of preparation and number of classes taught (two variables and their interaction)	5. Student activities/tasks (eight variables that include types of assignments)
	6. AP exam preparation techniques (seven variables, including focus of reviews, percent of class time dedicated to prepare for the AP exam during the year and during the month prior to the exam, and type and frequency of review activities implemented to prepare for AP exam)

Creation of Response Variable

Higher AP performance is to be expected among those students with higher general developed academic ability. For that reason, we did not use raw AP scores (reported on a 1-5 scale) as a criterion for our examination of the relationship between teachers' reported practices and students' AP biology performance. Instead, we constructed a new criterion from which general academic ability (as indicated by PSAT scores) was removed, at least approximately, as we discuss in the paragraph below. The criterion was calculated both for AP scores of three and above and four and above (hereinafter referred to as mean class residual [MCR] for criteria of scores of three or above—or MCR3, and mean class residual for criteria of scores of four or above—or MCR4, respectively). The rationale for employing these two criteria is that while traditionally scores of 3 or above were accepted for college credit, many colleges are now requiring scores of 4 or above.

To derive the residualized criteria, we ran a logistic regression at the individual student level, predicting success on the AP biology exam (i.e., 3 and above or 4 and above) from that individual's PSAT scores, using the sum of the PSAT verbal and PSAT math scores. With the fitted logistic regression, we were able to compute an estimated probability of success for each student on the AP biology exam. Next, we created an individual residual by subtracting the probability of success from the (coded) score of 1 or 0. This residual is simply the difference between the actual (coded) score and its expected value based on the fitted logistic regression.

This residual is analogous to (though not equivalent to) a gain score for the student. We treat the PSAT/NMSQT score as a proxy for prior achievement as well as unidentified confounding factors (Camara & Schmidt, 1999; Zwick, 2001). Although performance on the PSAT/NMSQT is admittedly a "noisy" indicator of student achievement, it is adequate for this study because we expect that the residual, in comparison to the raw score, will be a more sensitive criterion for ferreting out possible contributions of teacher reported practices to student achievement. Our analyses are conducted at the teacher level; consequently, we constructed a mean class residual by taking the average of the residuals for all the students in the class. Use of the MCR as a criterion places a burden on subsequent analyses inasmuch of the systematic variation has been removed through the preliminary logistic regression. The MCR contains a lot of noise even with averaging over students in a class; this is primarily due to the fact that the dichotomous outcome variable does not have that much information to begin with, but also because we are working with a large number of correlated variables, and what may emerge as significant may be due to the relationships among the predictors.

With regard to the timing of the surveys and the AP and PSAT scores, the surveys were administered in spring 2003, student AP exam scores were obtained from the spring 2003 administration, and PSAT/NMSQT scores for these students were matched from fall 2002 and previous administrations—as such, all students in this study took the PSAT/NMSQT prior to the AP exams.

Creation of Regression Models

We carried out a systematic series of exploratory unweighted regression analyses using the general linear models (GLM) methodology. The independent variables were organized into groups based on a theoretical model of AP overarching teacher practices that was the basis for the AP Teacher Study survey (Paek, Braun, Trapani, Ponte, & Powers, 2007). Every regression was run with the mean school PSAT entered first as a covariate into the model, so as to control for students' general achievement. It should be noted that the mean school PSAT was a significant predictor for all the models.

At each stage, the significant effects were identified using the partial sums of squares (Type III), generally using as a criterion a p-value of .15 for inclusion in subsequent models. The variables at our disposal fell naturally into two general areas: context for teaching and teacher practices. It is within the latter that types and frequency of assessment and feedback used by AP biology teachers were included.

All the teacher practice variables were entered in a new regression with mean school PSAT as a covariate; those that attained significance were retained in a parsimonious "Teacher Practices" model, and the same was done with the "Teacher Context" variables. To construct the final model, the mean school PSAT was entered first, next the parsimonious "Teacher Context" variables, and, lastly, the parsimonious "Teacher Practices" variables. After accounting for individual prior student achievement (by means of the preliminary logistic regression), mean school PSAT/NMSQT score, and other context variables, we are able to estimate the incremental contribution of teacher practices to AP exam results, either MCR3 or MCR4. This is a very stringent standard for identifying teacher reported practices that may be related to improved student achievement.

To compare teachers with different responses to an item, we computed the corresponding least squares means (LSM) of the MCRs, which control for all other terms in the model by accounting for the effects of correlated variables. We compared responses within an item (e.g., excellent, good, fair, and poor are response choices within one item) using a significance level of $p < .01$ to reduce capitalization on chance and thus the possibility of finding a false positive result. This increased the chances that significant differences represent real differences.

Results

In this section, we will first discuss the results of the AP biology survey in terms of teachers' reported use of various types of assessment and feedback practices, then present the findings from our general linear model analysis in which we study the relationship of those assessment and feedback practices to student performance in the AP biology test, using the PSAT to control for achievement and other contextual variables to minimize confounding effects.

Descriptive Statistics

In terms of assessments, multiple-choice tests are reportedly the most often used by AP biology teachers (more than 90 percent of teachers reportedly use them at least once a month), while presentations by students and independent research/projects are less commonly implemented in the classroom (about 30 percent of the teachers reported hardly ever using these assessment methods) (see Table 2 below). These findings align well with recent research showing that even though the use of alternative forms of assessment has been shown to be effective in improving instructional practices and student learning, its practice is not widespread. Some inhibiting factors mentioned by Harlen and Winter (2004) are (a) the widespread use of assessment practices that give more attention to grading and assigning learners to 'levels' rather than giving feedback about how work could be improved; (b) a general lack of awareness among teachers of pupils' learning needs; and (c) the high stakes attached to national test results, which encourage teaching to the test. The reported assessment practices of our sample of AP teachers supports research findings indicating limited use of alternative forms of assessment. However, when teachers commented on their use of more traditional forms of assessment, the reasons they gave for the use of these types of assessments were the vast breadth of the AP biology curriculum and exam, and the limited time they felt they had to cover and assess that material.

Table 2
AP Biology: Kinds of Assessments—Frequency Table

How often do you use the following kinds of assessments with your AP biology students?	*N*	1 Hardly ever	2 Several times a year	3 Once or twice a month	4 Once or twice a week	5 Almost every class session/ period	*M*	*SD*
a. Multiple-choice tests	1166	.9	4.6	67.7	23.4	3.4	3.24	.63
b. Tests requiring sentence- or paragraph-length responses	1159	10.6	9.7	56.6	20.9	2.2	2.94	.90
c. Tests requiring lengthy written responses	1163	6.1	19.7	59.4	13.2	1.5	2.84	.78
d. Laboratory notebooks or journals	1159	25.3	14.5	37.6	20.6	2.0	2.60	1.13
e. Presentations by students	1165	33.5	45.2	18.1	3.0	.2	1.91	.80
f. Independent research/projects by students	1165	50.2	40.9	7.6	.9	.3	1.60	.70

Note: Various cells (9 or 30 percent of all cells) have less than 5 percent of cases. Most of these cells are located under the last level of the scale ("almost every day").

Not surprisingly, and in alignment with our findings regarding teachers' reported use of assessment techniques, the most common type of feedback reportedly utilized by AP biology teachers is to provide students with numerical or letter grades (64 percent of AP biology teachers reported giving this type of feedback to students more than once a week). It should be noted that AP biology teachers rarely report providing students with lengthy descriptions of strengths and weaknesses (64 percent indicated hardly ever using this method) (see Table 3). These findings confirm the limited use of alternative modes of assessment that lend themselves to more descriptive, richer feedback, which tends to be seen as one of the conditions for assessment to support students' learning (Gibbs & Simpson, 2004).

Table 3
AP Biology: Feedback—Frequency Table

How often do students receive each of the following kinds of feedback on their tests or assignments for your AP classes?	*N*	1 Hardly ever	2 Several times a year	3 Once or twice a month	4 Once or twice a week	5 Almost every class session/ period	*M*	*SD*
Numerical or letter grades	1167	0	3.0	32.8	52.4	11.7	3.73	.70
Phrase or sentence-length descriptions of their performance	1161	17.9	20.0	42.0	18.2	1.9	2.66	1.03
Paragraph-length descriptions of strengths and weaknesses	1161	63.7	24.5	9.7	2.0	.2	1.51	.76
Page-length descriptions of strengths and weaknesses	1157	92.0	6.4	1.3	.3	.1	1.10	.38
Discussion of areas needing improvement	1160	14.5	37.4	34.9	10.9	2.3	2.49	.95
Comparison of performance with that of the class as a whole	1162	39.0	17.9	27.7	13.4	2.0	2.22	1.16

Note: Even though several cells (10 or 33 percent of all cells) have less than 5 percent of cases, these cells are not all under one level of the scale. Level 5 includes many cells with less than 2 percent of the cases, but many respondents checked this level for the first subitem (a). Thus, we can conclude that the data in this question is moderately well distributed.

Linking Teachers Reported Practices with Students Performance

We ran four separate analyses to look at the relationship for AP scores of 3 and above and then for AP scores of 4 and above, including all 667 teachers, as well as a separate analysis of only public school teachers. We wanted to see if there were different practices for public school teachers and for the two different criteria. The models below include all of the significant variables from our analyses. We will briefly discuss these models to then present our findings about the assessment and feedback variables, which constitute the focus of this paper.

Overall Model

The first step of the analysis was to determine how the inclusion of context and teacher variables, once the mean PSAT was added into the model, accounted for the variance in the dependent variable (MCR3 and MCR4 in our model, as described earlier in this paper). To that end, we carried a systematic but exploratory analysis by looking specifically for variables that showed differences in student performance. We employed the baseline R-square, using only the mean PSAT to get an idea of how much the context variables and the teacher practice profile contribute. Table 4 summarizes the contribution of the context and teacher variables for each of our analysis.

Table 4
R-Square and RMSE of Models

Model	R^2 (RMSE) of mean PSAT only	R^2 (RMSE) of mean PSAT and context and teacher practices
Biology public schools only (MCR3)	.06 (.182)	.27 (.167)
Biology public schools only (MCR4)	.06 (.179)	.33 (.158)
Biology combined public and nonpublic schools (MCR3)	.09 (.182)	.20 (.175)
Biology combined public and nonpublic schools (MCR4)	.05 (.187)	.40 (.154)

As shown in Table 4, to study the goodness of fit of our models, we looked at the *R*-square, which is not adjusted for the number of predictors. We also employed the RMSE of the fitted model as a diagnostic, which is a kind of generalized standard deviation and an excellent scalar measure of predictive efficacy of the model. We were interested in how much more of the criterion variance we could account for with context and teacher practice variables, as well as how much we could reduce the RMSE error. For example, for the public school teachers model, the initial regression using only mean school PSAT score as a predictor of the MCR3 criterion yielded an R-square of 0.06 and an RMSE of .182. In comparison, the final model combining selected context variables and selected teacher practices variables, along with the mean school PSAT has an R-square of .27 and an RMSE of .167. Thus, the selected survey variables do predict a modest amount of (residual) student performance, accounting for an additional 21 percent of the variance over mean school PSAT alone. There is also a nearly

8 percent reduction in the RMSE. This same trend was found for the other three models, where the context and teacher variables accounted for 33 percent, 20 percent, and 40 percent for the public teacher MCR4 combined public and nonpublic MCR3, and combined public and nonpublic MCR4 models, respectively. The RMSE reduction found in the public school MCR3 model also was similar to the results yielded in the three models (with a RMSE range from .15 to .18).

Assessment and Feedback Variables

For the model of AP biology public school teachers with criterion of MCR3, we included 442 teachers in the analyses; for the criterion of MCR4, we included 435 teachers. The most parsimonious models were both significant using a criteria of $p < 0.01$. The overall test of models shows $F(31, 408) = 4.95$ for MCR3 and $F(35, 402) = 5.70$ for MCR4.

There were three variables that differentiated AP biology public school teachers in both models (MCR3 and MCR4). Only one assessment variable was statistically significant, which was in the MCR3 model: tests requiring lengthy responses (see Table 5[1]). Teachers who reported using such assessments at least once or twice a month had classes perform significantly better on the AP exam than those who did this only several times per year or hardly ever. Part of the AP exam is constructed response, therefore teachers using these types of assessments help students practice what is expected on the high-stakes exam, as well as give teachers opportunities to review students' explanations and thinking. These types of assessments provide strong formative information to teachers on what their students understand or where they need more help, thus providing efficient forms of instructional intervention that can result in higher levels of learning. Our interpretation is that because the AP biology exam, a high-stakes assessment for these students, includes constructed response questions teachers are more prone to also use them in their instruction. These types of assessment tools provide teachers and students with more valuable information about the learning that is taking place, and in that way help teachers shape their practices in more effective ways and students focus their learning, thus resulting in better performance on the AP exam.

[1] The variables from the survey listed in Table 5 are the ones that differentiated more successful AP biology public school teachers from less successful teachers. For simplicity purposes, plus and minus signs have been used to depict relationships between variables. Detailed tables for each GLM can be found in Appendix A.

Table 5
AP Biology Public School Models: Statistically Significant Variables ($p < .10$)

Variables	MCR3	MCR4
Frequency of class meetings	+[a]	+
Percentage of students who take the exam	+	+
Influence of resources: AP exam topics and/or scoring rubrics	+	+
Number of years teaching	+	
Participation in AP professional development activities (PAPPD): Attending an AP institute	–[b]	
PAPPD: Reviewing AP Biology Teachers Guide	–	
Class size	–	
Objective: Learn scientific methods	+	
Assessment: tests requiring lengthy written responses	+	
Influence of resources: Frequency of using exemplary syllabi from other AP biology classes		–
Type and frequency of review activities: teacher estimate of students' time dedicated to study course material on their own		+
Teaching test-taking strategies		X[c]
Computer use: Teacher researching information on the Internet		–
Focus of attention to prepare students for AP exam		X
Percent of class time dedicated to prepare for the AP exam during the month prior to the AP exam		X

Note: Response choices to the variables are listed in increasing order unless the variable is listed as nominal.

Note: [a] '+' represents a positive relationship with the variable'; [b] '–' represents an inverse relationship with the variable; [c] 'X' represents a nonlinear relationship with the variable (e.g. the variable is not ordinal or the relationship is neither positive nor negative)

Our next set of analyses combines public and nonpublic school teachers, with a test for difference in school type. For the model with MCR3, we included 574 teachers in our analysis—132 nonpublic school teachers and 442 public school teachers. For the model with MCR4, we included 548 teachers—113 nonpublic and 435 public. We were analyzing two different types of schools—public and nonpublic, so we also tested whether the differences between school types were significant by testing the interactions of school types with the variables in the model. The asterisks in Table 6 indicate that the variable was significantly different by school type. For both public and nonpublic AP biology teachers, the variables discussed below proved significant ($p < .01$) in the models. The overall test of the model shows $F(32, 554) = 4.73$ for MCR3 and $F(64, 483) = 4.97$) for MCR4. For statistical details, see Tables A3 and A4 in Appendix A.

Both assessment and feedback items are shown as significant in the MCR4 model: assessing students using laboratory notebooks or journals, assessing students using presentations by students, and giving feedback that includes phrase or sentence-length descriptions of students' performance (see Table 6).

Table 6
AP Biology Public and Nonpublic School: Statistically Significant Variables (p < .10)

Variables	MCR3	MCR4
Class size	–[a]	–*[b]
Adequacy of facilities: laboratories, lab tables, sinks, etc.	+ *	
Computer use: Student researching information on the Internet	–	–
Objective: Learn scientific methods	+[c]*	
Student activities: Submitting reports on experiments or observations	–	
PAPPD: Attendance at an AP workshop		*
Number of AP biology classes taught		–*
Percentage of students who take the exam		+*
Influence of resources: Exemplary syllabi from other AP biology classes		–
Influence of resources: AP exam topics and/or scoring rubrics		+
Frequency of class meetings		+
PAPPD: Teach AP Institute		*
Assessment: Laboratory notebooks or journals		+*
Assessment: Presentations by students		+
Frequency of feedback: Phrase or sentence-length descriptions of their performance		+
Student activities: Apply biology concepts to real or simulated real-world problems		+

Note: [a] '–' represents an inverse relationship with the variable'; [b] '*'represents a significant difference by school type; [c] '+' represents a positive relationship with the variable.

Most teachers (77%) reported assessing their students by examining their laboratory notebooks less than one to two times per month. The results of the LSM analysis indicate that, in general, with other pedagogical variables held constant, students whose teachers made more frequent use of laboratory notebooks performed better on the AP exam. For nonpublic school teachers, those who noted they hardly ever assessed their students' lab notebooks had students performing significantly worse than those who claimed to do this more often; this difference proved significant when comparing public and nonpublic school teachers.

Most teachers used student presentations to assess student performance "sometimes" per year (80 percent of public and 87 percent of nonpublic school teachers). In general, students who were more frequently assessed on the basis of such presentations tended to perform better on the AP exam. Public school classes who were engaged in student presentations tended to perform significantly better than nonpublic school classes who engaged in this same type of activity.

Teachers claiming to give more detailed feedback had students with higher than expected performance. The LSM analysis shows the difference in student performance to be significant between those teachers who hardly ever provided such feedback and those who did so between one to two times per month and one to two times per week. Notably, in the models that include public and nonpublic schools, teachers' reported use of feedback was a practice significantly related to student performance, whereas this was not the case in the public-school-only model analysis. We speculate that, given that public and private teachers tend to spend the same amount of hours on the job per week, but public school teachers serve a much larger number of students in need of more attention (e.g., special education students, English-language learners, and students participating in Title II) (Strizek, Pittsonberger, Riordan, Lyter, & Orlofsky, 2006), it may be that teachers in nonpublic schools have more time to provide elaborate feedback to students and thus that variability allows for differences in the effect that variable had on student performance.

Limitations

The survey on AP biology teacher practices was designed to generate a nationally representative sample of teachers. However, considerable nonresponse implies that the obtained sample cannot be so considered. This shortcoming was exacerbated by the difficulty in linking students to their teachers in schools with more than one teacher for the particular AP subject, which further reduced the sample size and representativeness. Moreover, students without PSAT/NMSQT scores could not be included in the analysis. Finally, as we discussed earlier in this paper, classes with fewer than eight scores were eliminated from the analysis to obtain stable estimates of the dependent variable. Consequently, although the sample of teachers is relatively large and the amount of data collected voluminous, our results can only be suggestive of what we might have found had the original survey sample been available.

Several critical decisions made during the course of the analysis lead to focusing on one of many possible perspectives on the structure of the relationship between teacher practices and student achievement. First, rather than using a raw AP score as our dependent variable, we used the difference between the dichotomized score for the student and the corresponding expected value based on the student's PSAT/NMSQT score. Our intention was to minimize the presence of general academic performance on the dependent variable, thus enhancing the analysis' ability to pinpoint reported teacher practices associated with student performance in the AP biology exam. More specifically, we used the difference between the dichotomized score and the expected value based on AP score so that the criterion would more closely resemble a gain score for the student. In comparison to the raw score, the residual should be a more sensitive criterion for revealing possible teacher contributions to student achievement. In creating this residual, we treated the individual PSAT/NMSQT score as a proxy for prior achievement as well as other unmeasured relevant student characteristics. It is certainly possible that by adjusting for PSAT/NMSQT scores we may have removed some criterion variance that could be accounted for by teacher characteristics. Moreover, use of this criterion places a greater "burden" on the subsequent regression analyses, since much of the systematic variation was removed through the preliminary logistic regression.

In order to keep the data analysis manageable, we employed a stagewise approach to variable selection, employing parsimonious GLMs. The different parsimonious GLMs used as a first step were found using the partial sums of squares (Type III), employing a p-value of 0.15 for inclusion in subsequent models. Some would argue that this p-value was too liberal, as the most common p-value used for identifying significance is 0.05. We used this moderately liberal p-value because of possible collinearity among the variables. We did not want to eliminate potentially interesting variables too early in the analysis. Nonetheless, because of the large number of (correlated) variables used in the models, what emerged as significant—or not—may be due in part to the relationships among the predictors.

Conclusions

Although only a few feedback or assessment variables were significantly related to student AP exam performance, that does not mean that these techniques are not useful for improving student learning. A total of 41 variables were included as teaching practices, and due to collinearity, it is possible that we lost other assessment and feedback variables that were related to higher student success. More analyses of the relationship between AP teachers' reported practices and AP biology students' performance are warranted to see which variables were also related to higher performance (even if not significant).

Our analyses indicate that there are assessment and feedback techniques that AP biology teachers can use to improve their students' performance on the AP exams. Interestingly, these practices and techniques are closely associated with the constructivist pedagogical model, which advocates the use of authentic assessment: assessing students via lab notebooks/journals and student presentations and using tests requiring lengthy written responses. The latter seems to be the technique which more effective teachers (as indicated by their students' performance in the AP biology exam) reportedly use to prepare their students for the free-response portion of the AP biology exam.

Similarly, teachers who reportedly provide feedback that is not limited to a grade, but rather explain where students made mistakes and how to address them (i.e., providing students with phrase or sentence-long descriptions of their performance variable), seemed to helped students understand the material better, as evidenced in higher performance on the AP exam, and hopefully provided students with information to perform higher in the future.

These techniques have been researched with a representative sample of AP biology teachers who have a standardized syllabus and curriculum, yet we believe that the results of this study might be extendable to other subject matters and to non-AP teachers. More study is warranted around teaching practices to see how teachers are using assessment and feedback in ways that engage students more actively in their learning. Specifically, it would be beneficial to conduct an examination via extended classroom observations that would afford a better understanding of the relationship between teacher assessment practices and student learning and performance, and the implication of this relationship in terms of the current discussion in the educational arena in relation to teacher accountability and student learning.

Acknowledgements

We gratefully acknowledge the College Board for providing data for this study; however, the point of view or opinions stated in this report are not intended to represent College Board position or policies.

References

Airasian, P. W. (2000). *Classroom assessment* (4th ed.). New York: McGraw Hill.

Ainsworth, L., & Christinson, J. (1998). *Student generated rubrics: An assessment model to help all students succeed.* New York: Dale Seymour Publications.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice.* Berkshire, UK: Open University Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, *5*(1), 7-74.

Broadfoot, P., & Black, P. (2004). Redefining assessment? The first ten years of assessment in education. *Assessment in Education, 11(1),* 7-27.

Brookhart, S. M. (2005). *Research on formative classroom assessment.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Camara, W. J., & Schmidt, A. E. (1999). Group differences in standardized testing and social stratification. *College Board Report No. 99-05.* New York: The College Board.

Cochran-Smith, M., & Fries, K. (2005). Researching teacher education in changing Times: politics and paradigms. In M. Cochran-Smith & K. M. Zeichner (Eds.), *Studying teacher education: The report of the AERA panel on research and teacher education* (pp. 69-109). Washington, DC: AERA.

Darling-Hammond, L., & Ancess, J. (1996). Authentic assessment and school development. In J. B. Baron & D. P. Wolf

(Eds.), Performance-based student assessment: Challenges and possibilities (pp. 52-83). *Ninety-fifth yearbook of the National Society for the Study of Education*. Chicago: University of Chicago Press.

Darling-Hammond, L., Ancess, J., & Falk, B. (1995). *Authentic assessment in action*. New York: Teachers College Press.

Fullan, M., Hill, P., & Crévola, C. (2006). *Breakthrough*. Thousand Oaks, CA: Corwin.

Gagne, R. (1965). *The conditions of learning*. New York: Holt, Rinehart and Winston.

Gibbs, G., & Simpson, C. (2004). Does your assessment support your students' learning? *Journal of Learning and Teaching in Higher Education, 1(1),* 3-31

Gitomer, D., & Duschl, R. (1995). *Moving toward a portfolio culture in education*. Princeton, NJ: Center for Performance Assessment, Educational Testing Service.

Greeno, J. G., Collins, A. M., & Resnick, L. B. (1996). Cognition and learning. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 15-41). New York: MacMillian.

Harlen, W., & Winter, J. (2004). The development for assessment for learning: learning from the case of science and mathematics. *Language Testing, 21*(3), 390-408.

Hogan, K., & Pressley, M. (1997). Scaffolding scientific competencies within classroom communities of inquiry. In K. Hogan & M. Pressley (Eds.), *Scaffolding student learning: Instructional approaches and issues*. Cambridge, MA: Brookline Books.

Leahy, S., Lyon, C., Thompson, M., & Wiliam, D. (2005). Classroom assessment: Minute by minute, day by day. *Educational Leadership, 63*(3), 18-24.

Marzano, R. J. (2003). *What works in schools: Translating research into action*. Alexandria, VA: Association for Supervision and Curriculum Development.

McNamara, M., & Deane, D. (1995). Self assessment activities: Toward autonomy in language learning. *TESOL Journal, 5*(1), 12-14.

National Education Association. (2003). *Balanced assessment: The key to accountability and* improved student learning. Washington, DC: Author.

National Center for Education Statistics. (2002). *Schools and staffing survey, 1999-2000: Overview of the public, private, public charter, and Bureau of Indian Affairs elementary and secondary schools.* (NCES Publication No. 2002-313). Washington DC: Author.

National Center for Education Statistics. (2006). *Characteristics of schools, districts, teachers, principals, and school libraries in the United States 2003-2004 Schools and Staffing Survey.* (NCES 2006-313 REVISED). Washington DC: Author.

Newmann, F. M., Bryk, A. S., & Nagaoka, J. K. (2001). *Authentic intellectual work and standardized tests: Conflict or coexistence?* Chicago, IL: Consortium on Chicago School Research.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199-218

Olson, L. (2003). The great divide. *Education Week, 22*, 9-10, 13-14, 16, 18.

Paek, P. L., Braun, H., Trapani, C., Ponte, E., & Powers, D. (2007). *AP teacher practices: Analysis linking teacher and student data.* (ETS Rep. No. RR-05-10). Princeton, NJ: Educational Testing Service.

Paek, P. L., Ponte, E., Sigel, I., Braun, H., & Powers, D. (2005). *A portrait of Advanced Placement teachers' practices.* (College Board Rep. No. 2005-7). New York: College Board.

Popham, J. (2008). *Transformative assessment.* Alexandria, VA: Association for Supervision & Curriculum.

Porter, A. C., & Brophy, J. (1988). Synthesis of research on good teaching: insights from the

work of the institute for research on teaching. *Educational Leadership, 45*(8), 74-85.
Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford, M. C., O'Connor (Eds.), *Changing assessments. Alternative views of aptitude, achievement and instruction* (pp. 37-75). Norwell, MA: Kluwer Academic.
Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4-14.
Shepard, L. A. (2001). The role of classroom assessment in teaching and learning. In V. Richardson (Ed.), *Handbook of research on teaching (*4th ed.) (pp. 1066-1101). Washington, DC: American Educational Research Association.
Shepard, L., Hammerness, K., Darling-Hammond, L., Rust, F. (2005). Assessment. In L. Darling-Hammond & J. Bransford (Eds.), *Preparing teachers for a changing world: What teachers should learn and be able to do* (pp. 275-326). San Francisco: Jossey-Bass.
Smolen, L., Newman, C., Wathen, T., & Lee, D. (1995). Developing student self-assessment strategies. *TESOL Journal, 5*(1), 22-27.
Stiggins, R. J. (2002). Assessment crisis: The absence of assessment FOR learning, *Phi Delta Kappan, 83,* 758-765.
Stigler, J. W., & Hiebert, J. (1998, Winter). Teaching is a cultural activity. *American Educator*. Retrieved March 14, 2007, from www.aft.org/pubs-reports/american_ educator/winter98/index.html.
Stigler, J. W., & Hiebert, J. (1999). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom*. New York: Free Press.
Strizek, G. A., Pittsonberger, J. L., Riordan, K. E., Lyter, D. M., & Orlofsky, G. F. (2006). *Characteristics of schools, districts, teachers, principals, and school libraries in the United States: 2003-04 schools and staffing survey* (NCES 2006-313 Revised). Washington, DC: U.S. Government Printing Office.
Valencia, S. W., & Place, N. (1994). Literacy portfolios for teaching, learning, and accountability: The Bellevue literacy assessment project. In S. W. Valencia, E. H. Hiebert, & P. P. Afflerbach (Eds.), *Authentic reading assessment: Practices and possibilities* (pp. 134-156). Newark, DE: International Reading Association.
Wenglinsky, H. (2002). How schools matter: The link between teacher classroom practices and student academic performance. *Education Policy Analysis Archives, 10*(12), 1-31.
Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance.* San Francisco: Jossey-Bass Publishers.
William, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: impact on student achievement. *Assessment in Education, 11*(1), 49-65.
Wiliam, D., & Thompson, M. (2007). Integrating assessment with instruction: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 53-82). Mahwah, NJ: Erlbaum.
Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 12(2),* 181-208.
Wolf, D. P., Bixby, J., Glenn III, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. *Review of Research in Education, 17,* 31-74.
Zwick, R. (2001, October 29). Making the grade: The SAT vs. the GPA. *San Francisco Chronicle*, p. A17.

Appendix A: Type III Tables

Table A1
Type III: AP Biology Public School Teachers, MCR3

Source	*df*	Type III SS	Mean square	*F* value	Pr > *F*	η^2
Five-year mean PSAT (math and verbal)	1	1.0556	1.0556	37.58	<.0001	0.0627
Overall experience	2	0.2703	0.1352	4.81	0.0086	0.0169
PAPPD: Attended AP institute	2	0.3462	0.1731	6.16	0.0023	0.0215
PAPPD: Review of AP Biology Teachers Guide	2	0.2597	0.1299	4.62	0.0103	0.0162
PAPPD: Review of AP course description: Biology	2	0.1735	0.0867	3.09	0.0467	0.0109
Influence of resources: AP exam topics and/or scoring rubrics	4	0.2745	0.0686	2.44	0.0462	0.0171
Influence of resources: Schedule	4	0.3772	0.0943	3.36	0.0101	0.0234
Percentage of students who take the exam	3	0.5147	0.1716	6.11	0.0005	0.0316
Class size	3	0.3559	0.1186	4.22	0.0059	0.0221
Objective: learn scientific methods	4	0.2661	0.0665	2.37	0.0521	0.0166
Frequency of assessment: tests requiring lengthy written responses	4	0.2572	0.0643	2.29	0.0592	0.0160

Table A2
Type III: AP Biology Public School Teachers, MCR4

Source	*df*	Type III SS	Mean square	*F* value	Pr > *F*	η^2
Five-year mean PSAT (math and verbal)	1	0.7860	0.7860	31.37	< .0001	0.0496
Influence of resources: Exemplary syllabi from other AP biology classes	4	0.3357	0.0839	3.35	0.0103	0.0218
Influence of resources: AP exam topics and/or scoring rubrics	4	0.4744	0.1186	4.73	0.0010	0.0305
Influence of resources: Schedule	4	0.4908	0.1227	4.90	0.0007	0.0315
Percentage of students who take the exam	3	0.7416	0.2472	9.87	< .0001	0.0469
Teach test-taking strategies	4	0.4215	0.1053	4.21	0.0024	0.0272
Computer use: Teacher researching information on the Internet	1	0.1956	0.1956	7.81	0.0055	0.0128
Focus on multiple-choice, free-response, or both portions of the test	2	0.1753	0.0876	3.50	0.0311	0.0115
Percentage of class time dedicated to prepare for the AP exam during the month prior to the AP exam	4	0.2295	0.0573	2.29	0.0591	0.0150
Type and frequency of review activities: Teacher estimate of time dedicated to study course material on their own	4	0.3815	0.0953	3.81	0.0047	0.0247

Table A3
Type III: AP Biology Public and Nonpublic School Teachers, MCR3

Source	*df*	Type III SS	Mean square	*F* value	Pr > *F*	η^2
Five-year mean PSAT (math and verbal)	1	1.7409	1.7409	58.83	< .0001	0.0836
School type	1	0.0080	0.0080	0.27	0.6018	0.0004
Class size	3	0.6203	0.2067	6.99	0.0001	0.0315
Objective: Learn scientific methods	4	0.3746	0.0936	3.17	0.0137	0.0193
Assessment: tests requiring lengthy written responses	4	0.3603	0.0900	3.04	0.0168	0.0185
Student activities: Submit reports on experiments or observations	4	0.3173	0.0793	2.68	0.0309	0.0164
Computer use: Student researching information on the Internet	1	0.1107	0.1107	3.74	0.0535	0.0058
School type and Adequacy of facilities: Laboratories, lab tables, sinks, etc.	3	0.2053	0.0684	2.31	0.0750	0.0106
School type and Objective: Learn scientific methods	4	0.2773	0.0693	2.34	0.0538	0.0143

Note. When "school type" is included, the results show the interaction of school type with a context for teaching or teacher practice variable.

Table A4
Type III: AP Biology Public and Nonpublic School Teachers, MCR4

Source	*df*	Type III SS	Mean square	*F* value	Pr > *F*	η^2
Five-year mean PSAT (math and verbal)	1	1.1443	1.1443	48.06	< .0001	0.0566
School type	1	0.0148	0.0148	0.62	0.4306	0.0008
Objective: Communicate biological concepts effectively	3	0.2218	0.0739	3.11	0.0263	0.0115
Assessment: Laboratory notebooks or journals	4	0.2716	0.0679	2.85	0.0234	0.0140
Assessment: Presentations by students	4	0.2214	0.0553	2.32	0.0556	0.0115
Feedback: Phrase or sentence-length descriptions of their performance	4	0.3028	0.0757	3.18	0.0135	0.0156
Computer use: Student researching information on the Internet	1	0.1314	0.1314	5.52	0.0192	0.0068
Computer use: Simulation and modeling	1	0.1277	0.1277	5.36	0.0210	0.0067
PAPPD: Attended AP workshop	2	0.2243	0.1121	4.71	0.0094	0.0116
Influence of resources: Exemplary syllabi from other AP biology classes	4	0.4853	0.1213	5.10	0.0005	0.0248
Influence of resources: AP exam topics and/or scoring rubrics	4	0.3167	0.0791	3.33	0.0106	0.0163
Number of AP biology classes taught	4	0.3854	0.0963	4.05	0.0031	0.0198
Influence of resources: Schedule	4	0.4363	0.1090	4.58	0.0012	0.0224
Percentage of students who take the exam	3	0.2376	0.0792	3.33	0.0195	0.0123
Class size	3	0.3678	0.1226	5.15	0.0016	0.0189
School type and Assessment: Laboratory notebooks or journals	4	0.2496	0.0624	2.62	0.0343	0.0129

(Table continues)

Table A4 (continued)

Source	*df*	Type III SS	Mean square	*F* value	Pr > *F*	η^2
School type and PAPPD: Attended AP workshop	2	0.1460	0.0730	3.07	0.0475	0.0076
School type and PAPPD: Taught in an AP institute	1	0.1550	0.1550	6.51	0.0110	0.0081
School type and number of AP biology classes taught	3	0.2572	0.0857	3.60	0.0135	0.0133
School type and percentage of students who take the exam	3	0.2230	0.0743	3.12	0.0257	0.0116
School type and class size	2	0.2263	0.1131	4.75	0.0090	0.0117

Note. When “school type” is included, the results show the interaction of school type with a context for teaching or teacher practice variable.