The Reliability and Validity of Peer Review of Writing in High School AP English Classes

Christian Schunn, Amanda Godley, Sara DeMartino

The authors report on the reliability and validity of peer assessment of writing in high school Advanced Placement English classrooms. Students used a task-specific rubric to anonymously assess their classmates' writing.

Current literacy policies stress the need for high school students to develop argumentative writing skills in order to be prepared for college, for many careers, and for being critically engaged citizens (Newell, VanDerHeide, & Olsen, 2014). However, recent National Assessment of Educational Progress results suggest that many secondary students struggle with analytical, argument-driven writing (National Center for Education Statistics, 2012). One reason for this may be that high school students are provided with few opportunities to write and receive feedback on their writing (Applebee & Langer, 2013; Kiuhara, Graham, & Hawken, 2009). Surveys of teachers have shown that one reason teachers do not assign significant amounts of writing is due to the amount of time needed to grade and provide feedback on it (Gilbert & Graham, 2010; Kiuhara et al., 2009).

One approach to writing instruction that has been shown to improve secondary and college students' academic writing across disciplines, school settings, and grade levels without increasing the demands on teachers' time is peer review (Cho & MacArthur, 2011; Simmons, 2003). Although there are various ways that English language arts and other subject area teachers design such peer review activities, students typically read their peers' essays and provide feedback in response to teacher-generated questions or prompts. Writers then use the feedback to revise and ideally improve their essays before submitting them to the teacher for a final grade.

However, many teachers and students worry that students' feedback and assessment of their peers' writing is less accurate than teachers' (Gielen, Peeters, Dochy, Onghena, & Struyven, 2010; Hovardas, Tsivitanidou, & Zacharia, 2014; Kaufman & Schunn, 2011). Two studies of adolescents' perceptions of peer review of writing found that students perceived teacher feedback as more accurate and valuable than that of their peers even when the students used their peers' feedback to revise their essays (Gielen et al., 2010; Hovardas et al., 2014).

This study investigated whether high school Advanced Placement (AP) students in diverse school settings can accurately assess their peers' writing if given a carefully designed rubric to guide their assessment and feedback. We first explain the logic and method behind the construction of our rubric, a student-friendly version of the College Board's (2013) rubric for argument-driven AP English Language and Composition essays. We then examine the reliability and validity of the students' assessments by comparing them with their teachers' and

CHRISTIAN SCHUNN is a professor of psychology, learning sciences and policy, and intelligent systems and a senior scientist at the Learning Research and Development Center at the University of Pittsburgh, PA, USA; e-mail schunn@pitt.edu.

AMANDA GODLEY is an associate professor of English education and language, literacy, and culture at the University of Pittsburgh, PA, USA; e-mail agodley@pitt.edu.

SARA DEMARTINO is a doctoral student in language, literacy, and culture and a senior instructional designer at the Institute for Learning at the University of Pittsburgh, PA, USA; e-mail smd94@pitt.edu.

 doi: 10.1002/jaal.525 © 2016 International Literacy Association

trained AP scorers' assessments. Finally, we briefly discuss students' and teachers' perceptions of the peer review process.

Research on Peer Review

Peer review has been a common instructional approach in middle and high school English classes since the 1980s and is considered a key element of process approaches to writing (Atwell, 1987; Graham & Perin, 2007; Hillocks, 1984). However, little research on the validity and reliability of high school students' assessments of their peers' writing has been conducted (Brookhart & Chen, 2015).

At the secondary level, the design of peer review tasks and their effectiveness can vary greatly by classroom. In the absence of specific guidance, such as scoring rubrics and targeted prompts for qualitative feedback, students often provide their peers with empty praise (Tsivitanidou, Zacharia, & Hovardas, 2011; VanDeWeghe, 2004), avoid critiques of their peers' work because of fear of negative social consequences (Freedman, 1992; VanDeWeghe, 2004), or focus only on editing sentence-level errors (Freedman, 1992; Simmons, 2003). Additionally, peer review guidelines that ask students to provide yes/no answers or answer low-level questions (e.g., "Does the essay have three paragraphs?") lead to low-quality peer feedback that is not useful to writers (Goldberg, Roswell, & Michaels, 1995).

Two of the features essential to productive peer review are holding students responsible for taking the task of providing feedback seriously and lessening adolescents' fear of social repercussions of honest peer review. High school students frequently ignore written guidelines provided by their teachers in their peer reviews and assessments when the reviewing process has little teacher oversight or accountability (Freedman, 1992; Peterson, 2003).

Additionally, in face-to-face peer review, students are often more concerned about social dynamics, such as whether they will be perceived as mean or embarrass their peers, than providing accurate peer feedback (Christianakis, 2010; Peterson, 2003). The online peer feedback system that was used for this study addressed these concerns in two ways: by making the quality of the students' peer reviews part of their grade for the task and by offering a double-blind peer review process in which both authors and reviewers are anonymous.

Moreover, peer review of writing is most effective when guidelines for assessing and providing feedback require students to base their assessments on specific, understandable criteria and offer detailed suggestions for improvement (Gan & Hattie, 2014). Students also gain more from peer review when they are guided to focus more on global issues in writing, such as ideas and evidence, and when the activity develops students' awareness of audiences other than the teacher (Freedman, 1992; Simmons, 2003).

Studies at the college level have demonstrated that when students are guided by a clear rubric and held accountable for the quality of their peer feedback, their assessments of their peers' writing have strong reliability and validity (Cho, Schunn, & Wilson, 2006; Panadero, Romero, & Strijbos, 2013). We use the term *reliability* to refer to the extent to which students' assessments of peers' writing correlate with each other (i.e., inter-rater reliability). We use *validity* to describe the extent to which students can accurately judge what they are asked to assess in one another's writing, such as the quality of a thesis.

However, few studies of the validity and reliability of secondary students' peer assessments exist, and none to our knowledge focus on writing in English language arts. In Sadler and Good's (2006) study of seventh-grade science students' assessments of their peers' work (a task that included both multiple-choice and open-ended responses), the researchers found high correlations between the grades assigned by teachers and peers.

Other studies conducted in secondary computer science classes have shown similar results (Sung, Chang, Chiou, & Hou, 2005; Tseng & Tsai, 2007). Conversely, still other studies have found that secondary students' peer and self-assessments of writing are not consistent with teacher assessments (Chang, Tseng, Chou, & Chen, 2011; Varner, Roscoe, & McNamara, 2013). Our study adds to this body of research by focusing specifically on the reliability and validity of peer assessment of writing in English language arts when students are given a high-quality rubric and incentives for providing helpful feedback and accurate assessments.

Methods

Participants

Twenty-eight AP English Language and Composition teachers from 26 different schools located in 12 states across the United States took part in the study. Schools were primarily located in suburban areas (57%) but also included urban (25%) and rural (18%) areas. Most schools were traditional public (71%), but others were Catholic (11%), independent (11%), or charter schools (7%).

Based on academic performance data available online (e.g., ACT or SAT scores), 68% of the schools were high performing (with school means above national

averages), and 32% were low performing (with school means below national averages). A number of schools also had high proportions of historically underserved students of color and/or students eligible for free or reduced-price lunch.

The teachers were recruited through e-mails sent through the College Board and the National Writing Project. All but two teachers had previously used in-class peer assessment before, but only four had used online peer assessment (e.g., through Turnitin: turnitin.com). All but one teacher had given prior AP exams as practice earlier in the year, and most teachers had given more than four. Teachers participated in a one-hour online training session to learn about the online peer review system Peerceptiv (www.peerceptiv.com) and the structure of the study. Teachers also participated in an "act as a student" exercise in Peerceptiv to experience the system through the eyes of students.

A total of 1,215 students participated in the study. The largest number of students per teacher was 134 (in multiple class sections), and the smallest was 13.

Implementation of the Peer Review System

Peer assessment can be implemented through many different methods, from face-to-face discussions of drafts to Web-based distribution and rating methods. Peerceptiv is a Web-based system in which authors and reviewers are anonymous so students feel comfortable being more honest in their feedback. It has been used by over 50,000 students around the world, and currently approximately half of its users are high school students. Links for accessing Peerceptiv and other online review systems are shared in the More to Explore sidebar at the end of this article.

Participating teachers assigned a peer review activity using the Peerceptiv system to at least one class section of students between early April and early May. Before starting the peer review process, the teachers presented their students with a 30–45-minute lesson, provided by the researchers, on high-quality peer feedback (see Figure 1). The peer review process followed these steps:

1. Students uploaded their essays into the system.
2. The system automatically distributed the essays so each student received five peer essays.
3. Students used the rubrics developed by the researchers (described in the Description of Rubric Design section) to assess their peers' essays. The rubrics include both open-ended prompts (e.g., "Provide feedback on how well the author explained the textual evidence he or she provided") and numerical ratings (e.g., on a scale of 1–7, "How strong is the evidence for each claim about Louv's rhetorical strategies?").
4. After all reviews were completed, authors received two kinds of feedback: open-ended feedback provided by peer reviewers and scores that reflected the mean (average) of the reviewers' numerical ratings.
5. Authors then rated the helpfulness of the comments they received.
6. Peerceptiv automatically generated individual student grades for the peer review task based on the

Figure 1
Researcher-Designed Lesson on Good Peer Review

1. Explain to students, Why are we doing peer review?
   - Writing a sample AP essay will give you practice for the test.
   - Multiple peer reviews will help you see strengths and areas for improvement in your essay.
   - Reviewing other students' papers will give you new ideas about AP essays.
2. Ask students to think about some of the feedback they have received and given.
   - What kind of feedback has been most helpful?
   - What kind of feedback has been least helpful?
3. Whole-class discussion of examples of feedback:
   - Read the sample AP prompt.
   - Read the sample AP essay.
   - Read two different samples of peer feedback on the same essay.
   - Discuss what makes each type of peer feedback helpful (or not).
4. Sum up the discussion. Emphasize that helpful feedback:
   - Points out weaknesses or areas for improvement in the essay.
   - Tells the author WHERE in the paper the problem or weakness is.
   - Provides SPECIFIC suggestions to help the author improve hisor her work.
   - Note that:
     - General praise (aka "strong essay") is not helpful feedback.
     - Point out areas for improvement but be kind and respectful.
5. Practice with a partner (or individually) giving feedback on a different sample AP essay.
6. Discuss in pairs, small groups, or whole class about the feedback you gave to the essay.
7. Introduce Peerceptiv and how it works.
8. Introduce the revised AP rubric and check for student understanding of it.

Note. AP = Advanced Placement.

quality of the essays (as determined by the average of peers' ratings), the quality of the peer reviews, and the on-time completion of all aspects of the task.

7. Students used peer feedback to revise their essays.

The quality of students' peer reviews is calculated by Peerceptiv in two ways: by authors' ratings of the helpfulness of their open-ended feedback comments (step 5) and by the accuracy of the reviewer's numerical ratings of peers' essays (step 3). The accuracy of the reviewer's ratings is determined by comparing his or her ratings with the mean ratings produced by other peers on those same essays. The closer the reviewer is to the mean ratings produced by other students (across rating criteria and essays), the higher the accuracy grade. Because students' review grades decline the more their ratings differ from other reviewers', Peerceptiv provides a strong disincentive for students to cheat the system and give undeserved high or low grades to other students, even if they share their pseudonyms.

Essay Prompts

The essay prompt for this study was taken from the 2013 AP English Language and Composition exam. It presented students with a one-page passage (from *Last Child in the Woods: Saving Our Children From Nature-Deficit Disorder* by Richard Louv, 2008) and asked them to analyze the rhetorical strategies used by the author to develop his argument, with specific references to the text. This particular type of essay prompt is a common feature of the AP English Language and Composition exam and reflects a core instructional goal of the class (i.e., the analysis of written arguments).

Teachers were instructed to give students the suggested 40-minute time period to complete their responses. Because some teachers chose to have their students complete the writing outside of class, we cannot ensure that students limited their writing time to 40 minutes; however, the length of time students were given to write does not affect our findings.

Essay Scoring

Students' assessments of their classmates' essays were compared with those of other students, their teachers, and expert AP essay scorers to study their reliability and validity. We investigated the reliability of students' peer assessments by examining the extent to which students' numerical ratings correlated with one another for the same essays (i.e., inter-rater reliability). We then investigated the validity of students' assessments of their peers' essays by comparing them with teachers' and expert AP scorers' assessments of the same essays. Each participating teacher rated at least 15 of his or her students' essays using the researcher-designed rubric. On average, teachers rated 21 essays, for a total of 489 rated essays. Correlations between the student mean ratings (across the five peer assessments per essay) and the teacher ratings were computed separately for each teacher.

Although much of the research literature has used single-teacher ratings as the typical source of validity data, this approach has two problems. First, a single-teacher rating has unknown reliability itself. Second, high-stakes assessments, such as the AP tests, use multiple expert graders for each essay, and the graders go through careful training processes to ensure reliability.

Thus, three experienced AP graders who had been trained by the College Board also scored a subset of 100 AP essays. The essays chosen were evenly distributed across the teachers, sampled from the essays that both students and teachers had evaluated. Unlike students and teachers who used our revised rating rubric, the expert AP graders used the traditional College Board (2013) holistic rubric (with a scale of 1–9) to produce overall document scores through a scoring method closely aligning with the typical expert grading process. Validity of the student ratings was assessed by the correlation with expert AP ratings. Note that even though the expert AP raters used a 9-point scale and students and teachers used a 7-point scale (described in the Description of Rubric Design section), mathematically, the correlations between ratings are not influenced by these scale differences.

Surveys

At the end of the study, both students and teachers were sent links to an online survey about their perceptions of the benefits and disadvantages of the peer review task, and suggestions for improving the Peerceptiv system. The teacher survey also asked about prior experience with peer review, typical writing instruction, and the likelihood of using Peerceptiv in future years. Twenty-six of the 28 participating teachers (93%) and 343 (28%) of the students (a high rate compared with typical high school student survey responses of 13–20%; Porter & Whitcomb, 2003) completed their respective surveys.

Description of Rubric Design

A rubric is a set of criteria for evaluating student work that includes descriptions of performance levels (not just a numerical scale; Brookhart & Chen, 2015). Rubrics are widely used in K–12 writing instruction and for high-stakes writing assessments. Overall, rubrics have been

shown to increase student achievement and motivation in writing, but only when the rubrics convey clear and specific descriptions of quality levels for task-specific criteria (Andrade, Du, & Wang, 2008; Brookhart & Chen, 2015).

However, in high-stakes writing assessments, the language of the rubrics is often general and typically designed to be used only after extensive training and discussion of multiple benchmark sample essays. In the case of the AP English Language and Composition scoring guide, for instance, the difference between a score of 8, 6, or 4 on the 9-point scale hinges on whether students "effectively," "adequately," or "inadequately" analyze Louv's rhetorical strategies (for the full rubric, see College Board, 2013). Additionally, the terminology used in such rubrics may be unfamiliar to most high school students. In the AP rubric, we noted phrases such as "lapses in diction or syntax" and "mature prose style" (p. 2) as descriptors that would likely be unclear to most high schoolers.

The current study took on these challenges by revising the College Board's (2013) AP scoring guide to be more understandable and user-friendly for students. Our revisions were guided by four key design principles:

1. Students should be able to understand the language of the rubric, including writing terminology.
2. The rubric should focus students' attention on critical, high-value elements of the piece of writing.
3. The rubric should include concrete and specific descriptors or examples, not vague or comparative qualifiers such as *adequate* or *more sophisticated*.
4. Each rubric criterion should focus on only one aspect of the writing task (e.g., on quality of evidence not quality and explanation of evidence).

Following these principles, we first changed the holistic rubric to an analytic rubric. Holistic rubrics provide a single score based on an overall impression of student work, whereas analytic rubrics provide separate feedback on each important characteristic of the task. This focused student reviewers' attention on the most important characteristics of the essay rather than on determining an overall score. We identified eight essential characteristics or assessment criteria in the College Board (2013) rubric and in the scored essays on the College Board website (apcentral.collegeboard.com/apc/members/exam/exam_information/2001.html):

1. Thesis
2. Explanation of Louv's argument
3. Analysis of Louv's rhetorical strategies
4. Evidence for claims
5. Explanation of evidence
6. Organization
7. Control of language
8. Conventions

Next, we generated more concrete and specific descriptions for the performance levels for each criterion. In some places, we substituted phrases (e.g., "analyzes multiple, subtle rhetorical strategies" instead of "analyzes effectively"), and in other places, we added examples. In Figure 2, we share a portion of the original College Board (2013) scoring guide for writers' analysis of Louv's argument and a corresponding example from our revised rubric (the complete rubric is available at www.lrdc.pitt.edu/schunn/sword/jaalrubrics.docx). Our rubric is distinct from the College Board's scoring guide in a number of ways.

First, we separated two elements of the original rubric—explanation of Louv's argument (not shown in Figure 2) and analysis of Louv's rhetorical strategies—so students could separately assess writers' understanding of Louv's argument and their analysis of his use of rhetorical strategies. Second, we added explicit descriptions and examples of subtle and more obvious rhetorical strategies to clarify the characteristics of a high-scoring analysis of Louv's rhetorical strategies. Similarly, for the lowest possible score, we added descriptions of essay topics that might seem related to the essay prompt but did not address the prompt directly.

Third, we quantified the number of rhetorical strategies analyzed by students that the College Board scorers seemed to expect for each rating on the scoring guide. Finally, because Peerceptiv uses a 7-point scale for peer assessments that is nonadjustable, we converted the AP rubric's 9-point scale to 7, providing descriptions of performance levels for scores of 7, 5, 3, and 1 and leaving the intermediate scoring levels (6, 4, and 2) without descriptions so as not to overwhelm students.

In addition, open-ended commenting prompts were created for each rubric criterion so students could provide narrative feedback to one another. The only exception was that a combined commenting prompt was used for academic vocabulary and conventions. An early version of the rubric was piloted with three students currently enrolled in AP English Language and Composition courses, and then revised based on their feedback. This revised version was then tested on the teachers as part

Figure 2
Comparison of College Board[a] and Researcher-Created Advanced Placement (AP) Rubric Criteria for Analyzing Louv's[b] Rhetorical Strategies

College Board AP scoring cuide	Revised AP rubric
No open-ended feedback	Open-ended feedback: Provide feedback on how well the author analyzed how Louv's rhetorical strategies support his argument throughout the essay. Be specific about how the writer could improve his or her thesis analysis, and provide suggestions for improvements.
Ratings:	*Ratings:* What rhetorical strategies did the author analyze in his or her essay?
9	7—The author accurately analyzes multiple, subtle rhetorical strategies that Louv uses (e.g., appealing to a common cause, evoking nostalgia, or other sophisticated strategies).
"8—Effective Essays earning a score of 8 **effectively** analyze* the rhetorical strategies Louv uses to develop his argument about the separation between people and nature." "* For the purposes of scoring, analysis refers to explaining how the author's rhetorical choices develop meaning or achieve a particular effect or purpose."	6
7	5—The author analyzes three or more obvious rhetorical strategies that Louv uses (e.g., using rhetorical questions, anecdotes, or other obvious strategies).
"6—Adequate Essays earning a score of 6 **adequately** analyze the rhetorical strategies Louv uses to develop his argument about the separation between people and nature."	4
5	3—The author analyzes only one or two obvious rhetorical strategies that Louv uses (e.g., rhetorical questions) or misunderstands Louv's strategies.
"4—Inadequate Essays earning a score of 4 **inadequately** analyze the rhetorical strategies Louv uses to develop his argument about the separation between people and nature. These essays may misunderstand the passage, misrepresent the strategies Louv uses, or may analyze these strategies insufficiently."	2
3	1—The author didn't write about Louv's rhetorical strategies (instead discussed a different topic, connected to personal experience, or just summarized Louv's piece).
"2—Little Success Essays earning a score of 2 demonstrate **little success** in analyzing the rhetorical strategies Louv uses to develop his argument about the separation between people and nature. These essays may misunderstand the prompt, misread the passage, fail to analyze the strategies Louv uses, or substitute a simpler task by responding to the prompt tangentially with unrelated, inaccurate, or inappropriate explanation."	
1	

Note. [a]College Board. (2013). *AP® English Language a Composition 2013 scoring guidelines*. New York, NY: Author. [b]Louv, R. (2008). *Last child in the woods: Saving our children from nature-deficit disorder*. Chapel Hill, NC: Algonquin.

of their training and further revised to generate the final version of the rubric.

Findings

Reliability of Student Ratings

Our first research questions were, Did the students tend to agree with one another at sufficiently high levels to produce reliable scores? Did particular characteristics of writing, such as more complex ones like reasoning and warranting, lead to more disagreement in students' scores?

Peerceptiv automatically calculates the reliability of the mean peer quantitative ratings for each essay using an interclass correlation (ICC) and provides this information to teachers. This measure examines how consistently each student's pattern of ratings are with the ratings produced by the other students to determine the stability or trustworthiness of the resulting mean rating across students for each document. If the ICC is high, then assigning a given document to another group of students would produce the identical ratings; if the ICC is low, then another group of students could often produce a different rating.

Figure 3 shows these ICC values for each rubric criterion. All criteria except conventions show acceptable reliability, that is, ICC values above .40 (Fleiss, 1986), which is a common level of agreement among teachers scoring a set of documents (Cho et al., 2006). The low ICC for the conventions criterion shows that students seem to disagree a lot about whether an essay generally follows the conventions of standard written English. The first two criteria, quality of thesis and analysis of rhetorical strategies, have the highest reliability.

Thus, students seem to be able to reliably judge the more higher level or complex aspects of essay quality even more than lower level features, such as mechanics. Additionally, there were no statistically significant differences in reliability across higher versus lower performing schools on any of the rubric criteria, suggesting that students from diverse school contexts can reliably and accurately assess their peers' writing.

Validity of Student Ratings Against Teacher Ratings and Expert AP Ratings

Our next research question was, What is the extent to which students can accurately judge the quality of their peers' essays?—with "accuracy" determined by two sets of experts (AP English teachers and trained AP scorers). To answer this question, student ratings were compared with both teacher ratings and expert AP scorers' ratings to analyze the validity of the student ratings. As noted previously, each teacher rated at least 15 of their students' essays, and some teachers rated over 35 essays.

Correlations between teacher and mean student ratings for each paper were computed for each rating criterion to examine whether some criteria were judged more accurately than others. In addition, the correlations were calculated separately for each teacher to examine whether the alignment between teacher and student ratings varied by context. Mean correlations are presented in Figure 4. Correlations were generally

Figure 3
Mean Reliability (and Standard Error Bars) for Mean Peer Ratings on Each Rubric Criterion

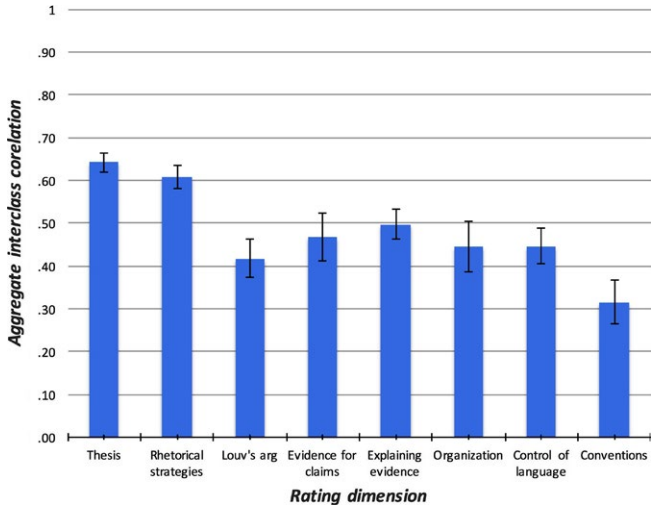
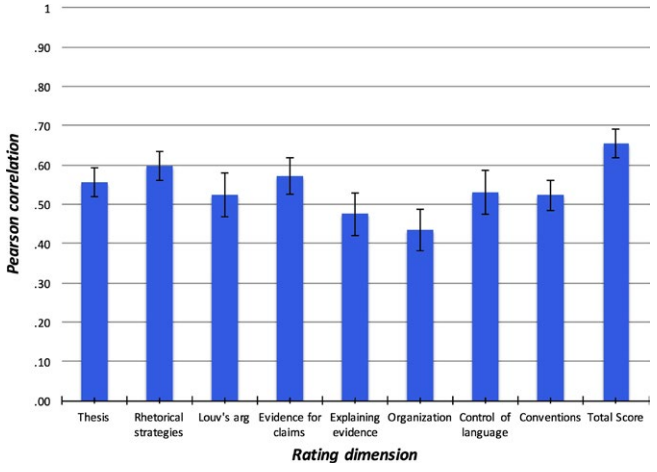


Figure 4
Mean Correlation (and Standard Error Bars) Across Classes of Mean Student Ratings With Teacher Ratings



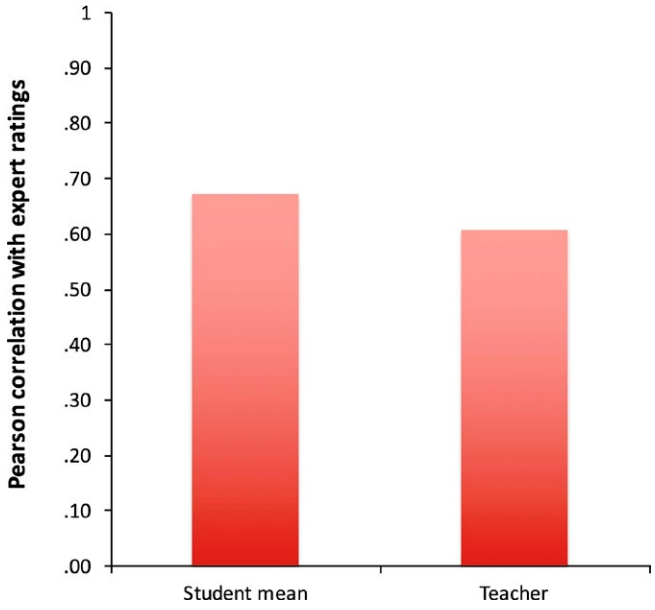
between .4 and .6 across rubrics and almost .7 for the overall essay rating.

These correlations are higher than what is typically seen between two teachers' ratings on a set of essays (Cho et al., 2006), suggesting that the accuracy of students' ratings of their peers' essays, when calculated by taking the average of five students' ratings, is very high. Finally, the student versus teacher correlations were very similar across higher and lower performing schools.

Figure 5 presents the correlations of the mean overall essay scores of students (averaged across student raters and rubrics) and teacher scores (averaged across rubrics) against the expert AP ratings. Pearson correlations are used as the primary metric because the ratings are not all on the same scale (i.e., student and teacher ratings are on the Peerceptiv 1–7 scale rather than the AP 1–9 scale).

There are a number of important outcomes to note from the results presented in Figure 5. First, both correlations between student mean scores and expert AP scores and between teacher and expert AP scores were acceptably high (at or above .5) and relatively similar. Despite the complexity of rating the AP essays even for expert AP raters, this study revealed that both students and teachers could produce useful ratings when provided with a carefully designed rubric. Second, the mean of student ratings correlated with experts' slightly more highly than did the teacher ratings. That is, the mean of five student ratings appears to be even more valid than single-teacher ratings. This suggests that if multiple students assess a peer's essay using a well-designed rubric, the average of the students' ratings could potentially be used in place of a teacher-generated grade.

Figure 5
Correlations With Advanced Placement Expert Ratings for Mean Student Ratings and Individual Teacher Ratings

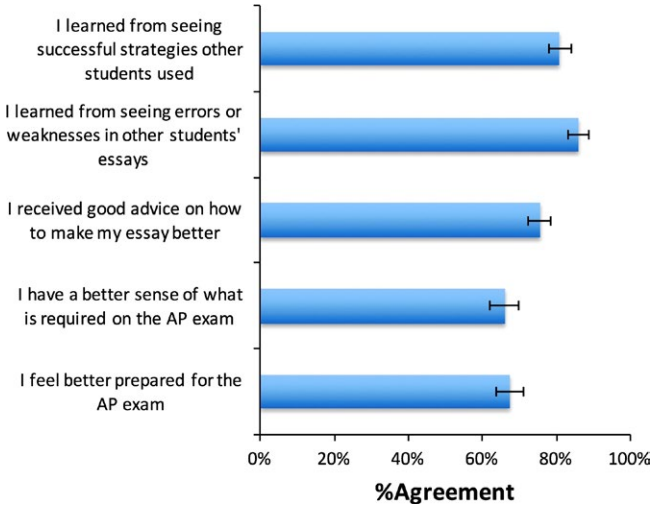


Student and Teacher Perceptions of Peer Review

Our final research question was, What were students' and teachers' perceptions of the peer review process? Overall, both students and teachers thought that using Peerceptiv for peer review was beneficial. A majority of students agreed with each of the possible benefits of peer review on our survey (see Figure 6). Although students generally thought that they had received good advice from their peers, the strongest perceived benefits involved seeing successful strategies and weaknesses in other students' essays. The perception that assessing peers' writing is more beneficial than receiving peer assessments is consistent with the findings of previous studies (Godley, Loretto, & DeMartino, 2014).

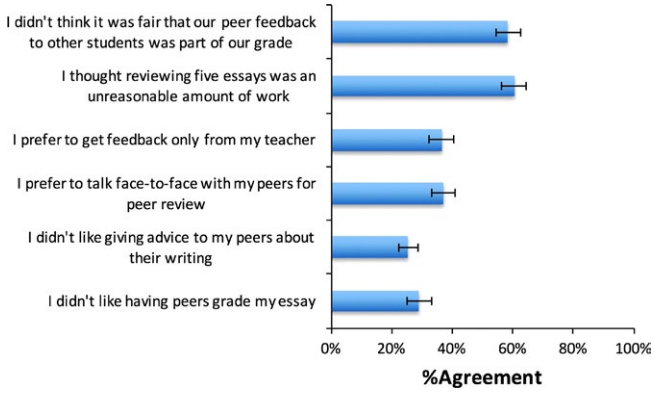
The survey also provided insights into student concerns (see Figure 7). The only concerns endorsed by a majority of students were that they did not like having grades based on peer assessments, and they were concerned about the workload, which mirrors the findings of previous research (Kaufman & Schunn, 2011). Regarding workload, most comments suggested that four peer assessments would be reasonable rather than

Figure 6
Percentage of Students Agreeing or Strongly Agreeing With Each Possible Benefit of Peer Assessment of Writing (With Standard Error Bars)



Note. AP = Advanced Placement.

Figure 7
Percentage of Students Agreeing or Strongly Agreeing With Each Possible Disadvantage of Peer Assessment of Writing (With Standard Error Bars)



five. Also, only a minority of students preferred teacher feedback or face-to-face (rather than online) peer feedback, suggesting that student buy-in for using peer feedback instead of teacher feedback is strong, even if students question whether their peers' assessment of their work is accurate.

The teacher survey also probed the perceived benefits and disadvantages of online peer assessment. In general, all but one of the teachers agreed that students learned from the peer review activity and gave helpful feedback to their peers. Like students, teachers perceived the benefits of giving feedback as even greater than receiving feedback. All but one teacher felt that their students were better prepared for the AP exam as a result of only one round of peer assessment.

Finally, in terms of the feasibility of using online peer assessment, all but one teacher felt that this was a convenient way for students to receive detailed feedback on their writing, and 80% of the teachers felt that they could easily implement this kind of peer assessment on a regular basis.

Only one teacher did not want to use online peer assessment in the next year. Of all the teachers, 35% wanted to do so a few times per year, and almost 60% wanted to use online peer assessment regularly throughout the year. Overall, teachers generally saw that the benefits outweighed the difficulties and that online peer assessment of this form significantly adds value over more informal peer feedback that can be done face to face in class.

Conclusion

Can high schoolers reliably assess their peers' academic writing? The results of our study suggest yes. Given teachers' and students' concerns about the validity and reliability of peer assessment of writing, this study demonstrated that when using a carefully designed rubric, high school students were able to provide ratings that were more valid than the ones provided by a single teacher and just as valid as the ones provided by expert AP scorers. Further, most students perceived their peers' feedback as helpful.

Across school contexts (urban, suburban, and socioeconomic status levels) and student writing achievement levels, students' ratings were acceptably correlated with teachers' and experts', and students were able to consistently rate the higher level aspects of academic writing, such as explaining evidence. Thus, it does not seem to be the case that only the most capable students are able to participate meaningfully in peer assessment activities. These positive results of peer assessment across contexts mirror the findings of prior research examining the peer reviews of stronger and weaker students (Patchan, Hawk, Stevens, & Schunn, 2013). Finally, both teachers and students perceived many benefits of participating in peer assessment, and most participating teachers believed that online peer assessment should be used regularly in their classes.

Our study has a number of limitations. First, it was limited to one approach to peer review that included three key features: anonymity, a student-friendly rubric, and accountability for the quality of feedback given to peers. It is possible that different results would be obtained from another approach to peer review, whether online or face to face. Second, we focused on a writing task that was familiar to the AP students: an analysis of rhetorical strategies. Although the rubric used for peer assessment was new to the students, it is likely that the familiarity of the task increased their understanding of the expectations and, thus, the reliability and validity of their peer assessments.

Third, the study involved only AP students, who tend to be in 11th and 12th grades and have developed relatively stronger academic literacy skills than other high school students. However, in recent years, there has been a nationwide movement to enroll younger students with a wider range of academic preparedness in AP classes (Godley, Monroe, & Castma, 2015). Additionally, our other studies have demonstrated the benefit of peer review in diverse, non-AP high school classes (Godley et al., 2014).

Fourth, our study did not tackle the issue of how to convince students that peer-generated grades are fair and can be used in place of teacher grading. As the research on peer assessment at the college level has shown, students' perceptions of fairness are a significant

TAKE ACTION!

These steps can be done without technology, but online tools simplify the logistics:

1. Find online peer review tools that are available for your school or district.
2. Create a rubric with student-friendly language.
3. Create guidelines for how students should participate in the assessment process.
4. Teach students about good peer review. Allow them time to practice assessing a shared piece of writing in pairs or small groups.
5. Anonymity may be achieved by having students select pseudonyms or assigning students numbers that only you and they know and that they will put on each document used during peer review, keeping the writers' and reviewers' identities a secret.
6. Assign three or four peers to review each student's paper. Research has shown that students learn as much or more from giving reviews as from receiving them.
7. Before delivering the peer feedback to students, scan it for common issues and average reviewers' ratings.
8. In class, discuss papers that received conflicting reviews.
9. Allow students time to reflect on the process of peer review before asking them to revise. This will give students time to consider their own understanding of the rubric and synthesize feedback from multiple peers before revising.

challenge to the use of peer assessment in determining task or course grades (Kaufman & Schunn, 2011). Finally, some students were unhappy with the workload required to review five peers' essays. Mathematically, fewer peer assessments per essay, such as three or four, would generate mean essay scores with slightly lower validity; however, it is expected that four peer assessments per essay would still produce ratings as valid as those produced by a single teacher.

Despite these limitations, our study suggests that carefully designed peer review tasks can provide students with helpful feedback on and fair assessments of their writing without increasing English language arts teachers' workload. We believe that peer assessment can be a feasible and productive solution to the challenge of asking high school students to write and revise more often without increasing English teachers' paper load. Given the evidence that students learn from peer review and from using rubrics, peer assessment can enhance writing instruction in high schools.

NOTES

This study was funded by The College Board and by the Institute of Education Sciences, U.S. Department of Education (R305A120370). The first author has a significant financial interest in the company that makes Peerceptiv available.

REFERENCES

Andrade, H.L., Du, Y., & Wang, X. (2008). Putting rubrics to the test: The effect of a model, criteria generation, and rubric-referenced self-assessment on elementary school students' writing. *Educational Measurement: Issues and Practice*, *27*(2), 3–13. doi:10.1111/j.1745-3992.2008.00118.x

Applebee, A.N., & Langer, J.A. (with Wilcox, K.C., Nachowitz, M., Mastroianni, M.P., & Dawson, C.). (2013). *Writing instruction that works: Proven methods for middle and high school classrooms*. New York, NY: Teachers College Press; Berkeley, CA: National Writing Project.

Atwell, N. (1987). *In the middle: Writing, reading, and learning with adolescents*. Portsmouth, NH: Heinemann.

Brookhart, S.M., & Chen, F. (2015). The quality and effectiveness of descriptive rubrics. *Educational Review*, *67*(3), 343–368.

Chang, C.-C., Tseng, K.-H., Chou, P.-N., & Chen, Y.-H. (2011). Reliability and validity of Web-based portfolio peer assessment: A case study for a senior high school's students taking computer course. *Computers & Education*, *57*(1), 1306–1316. doi:10.1016/j.compedu.2011.01.014

Cho, K., & MacArthur, C. (2011). Learning by reviewing. *Journal of Educational Psychology*, *103*(1), 73–84. doi:10.1037/a0021950

Cho, K., Schunn, C.D., & Wilson, R.W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology*, *98*(4), 891–901. doi:10.1037/0022-0663.98.4.891

Christianakis, M. (2010). "I Don't Need Your Help!": Peer status, race, and gender during peer writing interactions. *Journal of Literacy Research*, *42*(4), 418–458. doi:10.1080/1086296X.2010.525202

College Board. (2013). *AP® English Language a Composition 2013 scoring guidelines*. New York, NY: Author. Retrieved from media.collegeboard.com/digitalServices/pdf/ap/apcentral/ap13_english_language_scoring_guidelines.pdf

Fleiss, J.L. (1986). *The design and analysis of clinical experiments*. New York, NY: John Wiley & Sons.

Freedman, S.W. (1992). Outside-in and inside-out: Peer response groups in two ninth-grade classes. *Research in the Teaching of English*, *26*(1), 71–107.

Gan, M.J., & Hattie, J. (2014). Prompting secondary students' use of criteria, feedback specificity and feedback levels during an investigative task. *Instructional Science*, *42*(6), 861–878. doi:10.1007/s11251-014-9319-4

Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction*, *20*(4), 304–315. doi:10.1016/j.learninstruc.2009.08.007

Gilbert, J., & Graham, S. (2010). Teaching writing to elementary students in grades 4–6: A national survey. *The Elementary School Journal*, *110*(4), 494–518. doi:10.1086/651193

Godley, A., Loretto, A., & DeMartino, S. (2014, December). *The dialogic construction of academic writing through peer review in secondary classrooms*. Paper presented at the annual meeting of the Literacy Research Association, San Marco Island, FL.

Godley, A., Monroe, T., & Castma, J. (2015). Increasing access to and success in Advanced Placement English in Pittsburgh Public Schools. *English Journal*, *105*(1), 28–34.

Goldberg, G.L., Roswell, B.S., & Michaels, H. (1995). Can assessment mirror instruction? A look at peer response and revision in a large-scale writing test. *Educational Assessment*, *3*(4), 287–314. doi:10.1207/s15326977ea0304_1

Graham, S., & Perin, D. (2007). *Writing next: Effective strategies to improve writing of adolescents in middle and high schools*. Washington, DC: Alliance for Excellent Education.

Hillocks, G., Jr. (1984). What works in teaching composition: A meta-analysis of experimental treatment studies. *American Journal of Education*, *93*(1), 133–170. doi:10.1086/443789

Hovardas, T., Tsivitanidou, O.E., & Zacharia, Z.C. (2014). Peer versus expert feedback: An investigation of the quality of peer feedback among secondary school students. *Computers & Education*, *71*, 133–152. doi:10.1016/j.compedu.2013.09.019

Kaufman, J.H., & Schunn, C.D. (2011). Students' perceptions about peer assessment for writing: Their origin and impact on revision work. *Instructional Science*, *39*(3), 387–406. doi:10.1007/s11251-010-9133-6

Kiuhara, S.A., Graham, S., & Hawken, L.S. (2009). Teaching writing to high school students: A national survey. *Journal of Educational Psychology*, *101*(1), 136–160. doi:10.1037/a0013097

Louv, R. (2008). *Last child in the woods: Saving our children from nature-deficit disorder*. Chapel Hill, NC: Algonquin.

National Center for Education Statistics. (2012). *The Nation's Report Card: Writing 2011* (NCES 2012-470). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Newell, G.E., VanDerHeide, J., & Olsen, A.W. (2014). High school English language arts teachers' argumentative epistemologies for teaching writing. *Research in the Teaching of English*, *49*(2), 95–118.

Panadero, E., Romero, M., & Strijbos, J.W. (2013). The impact of a rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort. *Studies in Educational Evaluation*, *39*(4), 195–203. doi:10.1016/j.stueduc.2013.10.005

Patchan, M.M., Hawk, B., Stevens, C.A., & Schunn, C.D. (2013). The effects of skill diversity on commenting and revisions. *Instructional Science*, *41*(2), 381–405. doi:10.1007/s11251-012-9236-3

Peterson, S. (2003). Peer response and students' revisions of their narrative writing. *L1-Educational Studies in Language and Literature*, *3*(3), 239–272. doi:10.1023/B:ESLL.0000003605.45224.b5

Porter, S.R., & Whitcomb, M.E. (2003). The impact of contact type on Web survey response rates. *Public Opinion Quarterly*, *67*(4), 579–588. doi:10.1086/378964

Sadler, P.M., & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment*, *11*(1), 1–31. doi:10.1207/s15326977ea1101_1

Simmons, J. (2003). Responders are taught, not born. *Journal of Adolescent & Adult Literacy*, *46*(8), 684–693.

Sung, Y.-T., Chang, K.-E., Chiou, S.-K., & Hou, H.-T. (2005). The design and application of a Web-based self- and peer-assessment system. *Computers & Education*, *45*(2), 187–202. doi:10.1016/j.compedu.2004.07.002

Tseng, S.-C., & Tsai, C.-C. (2007). On-line peer assessment and the role of the peer feedback: A study of high school computer course. *Computers & Education*, *49*(4), 1161–1174. doi:10.1016/j.compedu.2006.01.007

Tsivitanidou, O.E., Zacharia, Z.C., & Hovardas, T. (2011). Investigating secondary school students' unmediated peer assessment skills. *Learning and Instruction*, *21*(4), 506–519. doi:10.1016/j.learninstruc.2010.08.002

VanDeWeghe, R. (2004). "Awesome, dude!": Responding helpfully to peer writing. *English Journal*, *94*(1), 95–99. doi:10.2307/4128855

Varner, L.K., Roscoe, R.D., & McNamara, D.S. (2013). Evaluative misalignment of 10th-grade student and teacher criteria for essay quality: An automated textual analysis. *Journal of Writing Research*, *5*(1), 35–59. doi:10.17239/jowr-2013.05.01.2

MORE TO EXPLORE

Resources for online peer review:

- Peerceptiv: www.peerceptiv.com
- Calibrated Peer Review: cpr.molsci.ucla.edu/cpr/cpr_info/index.asp
- Eli Review: https://app.elireview.com
- PeerMark (a feature of Turnitin): https://guides.turnitin.com/01_Manuals_and_Guides/Student/Student_User_Manual/19_PeerMark

Resources for using rubrics:

- Edutopia's "Resources for Using Rubrics in the Middle Grades": www.edutopia.org/rubrics-middle-school-resources

Print resources:

- Andrade, H.G. (2000). Using rubrics to promote thinking and learning. *Educational Leadership*, *57*(5), 13–18.
- Topping, K.J. (2009). Peer assessment. *Theory Into Practice*, *48*(1), 20–27.