doi: 10.1111/j.1365-2729.2008.00290.x

Peering into large lectures: examining peer and expert mark agreement using peerScholar, an online peer assessment tool

D.E. Paré & S. Joordens

University of Toronto Scarborough, Toronto ON, Canada

Abstract As class sizes increase, methods of assessments shift from costly traditional approaches (e.g. expert-graded writing assignments) to more economic and logistically feasible methods (e.g. multiple-choice testing, computer-automated scoring, or peer assessment). While each method of assessment has its merits, it is peer assessment in particular, especially when made available online through a Web-based interface (e.g. our peerScholar system), that has the potential to allow a reintegration of open-ended writing assignments in any size class – and in a manner that is pedagogically superior to traditional approaches. Many benefits are associated with peer assessment, but it was the concerns that prompted two experimental studies ($n = 120$ in each) using peerScholar to examine mark agreement between and within groups of expert (graduate teaching assistants) and peer (undergraduate students) markers. Overall, using peerScholar accomplished the goal of returning writing into a large class, while producing grades similar in level and rank order as those provided by expert graders, especially when a grade accountability feature was used.

Keywords critical thinking, expert marking, grading, peerScholar online, peer assessment, writing.

> Upon the education of the people of this country the fate of this country depends (Disraeli 1874).

This quote voices the logical proposition that an educated group of citizens will have the ability to use their knowledge in innovative ways, ultimately resulting in benefits to society such as better health care, successful business ventures, superior technological development, and even enhanced arts endeavours. These benefits do not result from people merely having knowledge, but rather from their abilities to draw on that knowledge to think critically about new information, to arrive at novel perspectives, and to communicate those perspectives clearly to others. In fact, it may be that the specific knowledge learned throughout the educational process is secondary to the critical thinking and communication skills that are transferable to any context.

In this article, we first discuss some of the economic and logistic issues of knowledge assessment in large classes, and briefly review potential solutions to these issues including the use of peer assessment. We then describe the peerScholar[1] system (http://www.peerScholar.com), an online tool that allows written assignments to be employed in any size class in a manner that is not only economical, but more importantly, pedagogically superior to the expert marker approach. Two experiments examining the grading dynamics of the peerScholar system are presented. To anticipate, while development and testing of peerScholar is ongoing, the current research shows that the existing instantiation of the system is pedagogically powerful while resulting in peer grades that are similar

Accepted: 31 May 2008

Correspondence: Dwayne E. Paré, University of Toronto Scarborough, 1265 Military Trail, Toronto ON, Canada, M1C 1A4. Email: dpare@psychexperiments.com

in absolute level and rank order relative to the grades provided by experts.

Background

The best way to ask students to think critically and to communicate clearly is via open-ended writing assignments (Ennis 1993; White 1993), but unfortunately, such assignments have logistic issues that make them difficult to implement on a large scale, at least in their traditional implementation. That traditional implementation is what we will term the 'expert marker' approach, and it involves students first composing a written assignment, and then submitting it to an expert for grading and feedback. The time of experts tends to be valuable and marking is a time-consuming process; as a result, two problems arise when such assignments are carried out in large class settings. The first is economic; written assignments are simply very expensive to implement on a large scale. The second is pedagogical; as time passes between the submission of an assignment and students' receiving feedback on their work, that feedback becomes less meaningful in terms of shaping learning (Heywood 1989). In addition, there is another less obvious pedagogical shortcoming of the expert marker approach. Namely, students do not see any assignments other than their own and therefore do not experience both poorly and well-written work, preventing an occasion to understand what makes a composition relatively strong (or weak), thereby diminishing an opportunity for students to improve the quality of their work (Sadler 1989).

Multiple-choice testing is a solution that many institutions have come to rely on heavily and, in some courses exclusively. Multiple-choice exams, when properly constructed, can assess learning in deep ways, assessing the extent to which students have acquired basic knowledge, comprehension, application, analysis, synthesis, and evaluation (Bloom 1956; Anderson & Sosniak 1994). In fact, tests such as the American Medical College Admission Test represent examples of strong multiple-choice tests that do assess learning at all these levels (Zheng *et al*. 2008).

However, multiple-choice tests remain limited by the concrete manner in which questions must be asked. There simply is no room for students to write freely and creatively, nor to assess the extent to which students are able to communicate these novel thoughts in clear and effective ways (Ackerman & Smith 1988; Raimondo *et al*. 1990). These aspects of written assignments are critical to learning as society moves forward on the basis of new ideas, led by those who can communicate those ideas effectively. Thus, the movement towards heavy reliance on multiple-choice exams, even very good multiple-choice exams, is concerning.

So how does a university retain or reintroduce writing assignments in a climate of very large student to teacher ratios? One solution is to have a majority of courses use primarily multiple-choice assessment, while offering a small number of heavily funded courses that focus on thinking and writing. However, thinking and communication skills, like virtually every other cognitive ability, develop with repeated practice (Schneider & Shiffrin 1977; Shiffrin & Schneider 1977) and develop best with distributed rather than massed practice (Melton 1970; Glenberg 1979; see Seabrook *et al*. 2005 for a demonstration in an educational context). Thus, a superior option would be to find an economical and logistically reasonable method for bringing writing and thinking skills back into the majority of classes, regardless of the enrolment.

There are two methods of assessment that can be offered as potential solutions to retaining or reintroducing writing assignments into large classes: computer-automated scoring systems and peer assessment. Automated marking systems can produce grades that are similar to human markers (Williamson *et al*. 1999; Landauer *et al*. 2003), can be used economically in any class size, and can provide prompt numeric feedback (see Miller 2003 for an overview). Peer assessment has all of these advantages, but more importantly, also has the pedagogical benefits described below that enhance its potential for education. For this reason, we focus primarily on peer assessment throughout this article.

Peer assessment

Peer assessment, sometimes called peer evaluation or peer review, is a process wherein peers evaluate each other's work, usually along with, or in place of, an expert marker (e.g. Topping 1998; Woolhouse 1999; Rada & Hu 2002). There are potential trade-offs associated with the use of peer assessment. Benefits include the logistic and economical advantages mentioned previously. These are especially relevant in courses where there is no money for additional teaching assistants. It is in these courses that writing assignments have been eliminated because of financial constraints, and through the use of peer assessment, these assignments can be reintroduced to the curriculum.

While the functional benefits of peer assessment may allow the inclusion of written assignments in any class context, it is the pedagogical benefits inherent in having students grade the work of their peers that are especially exciting (Sims 1989; Liu *et al*. 2002). Specifically, seeing and critiquing peers' work is said to encourage deeper analyses of the student's own work (Sims 1989), a process that is required for the improvement of work quality (Sadler 1989). The use of peer assessment has also been shown to be a useful tool in exposing students to the real world of scientific discourse (Towns *et al*. 2001; Venables & Summit 2003), a world where peer assessment is a normal and accepted process. Peer assessment can also be used to teach students how to provide both quantitative and qualitative feedback properly (Bloxham & West 2004), and this can be beneficial in any career where giving assessment and feedback might be expected (e.g. business). Finally, the timeliness of peer assessment, at least in the context of electronic peer assessment, allows students to receive feedback on their own work almost immediately, when it is the most effective in shaping learning (Heywood 1989).

These benefits do not come without potential costs. There are concerns with peer assessment that need to be considered, and it is frequently these concerns that cast a shadow over the benefits described above. Concerns include (a) the quality of peer-derived grades; and (b) the practical implementation of peer assessment in the classroom, especially a very large classroom.

The implementation concern is paramount in the context of large classes as any worries concerning quality of grades become mute if an effective system for implementing peer assessment is not available. Successful attempts have been made with students marking each other's work using paper and pencil methods (Sims 1989; Liu *et al*. 2002) and via electronic systems such as online forums (Towns *et al*. 2001; Trautmann *et al*. 2003) in small- to medium-sized class settings, but these approaches become problematic as class size increases. Paper and pencil methods would be very difficult to implement and to monitor with many students. Online forums would be a step in the right direction, but efficiently organizing a method for large numbers of students to mark each other would again be very time consuming. The ideal method would be an easy to use Web-based tool that had the flexibility for instructors to generate customized peer assessment assignments. While conceptualizing these ideas, there were no such tools generally available for use; therefore, we created the peerScholar system.

peerScholar

The peerScholar system was originally developed to address the need for writing and critical thinking assessments in a specific class setting; the Introductory Psychology course at the University of Toronto Scarborough that has enrolments of over 2000 students per year. Prior to the introduction of peerScholar, assessments were based solely on multiple-choice exams for the reasons discussed previously. The challenge was to return the desired writing components to the course and after consideration, the notion of online peer assessment held a high level of promise.

The peerScholar system is an automated online tool used by instructors to manage required readings, writing pieces, assessments, and results for student assignments – all through one Web-based interface. The Internet was chosen as the method of delivery because it ensured a widely accessible system. It also allowed instructors to maintain complete control over assignment content from a single interface. Administering any kind of assessment simultaneously to hundreds of students, let alone thousands, can be daunting. Finding space to physically house all of the students is near impossible and even simple changes in instructions or assignment requirements become a challenge. It was for these reasons that the use of Web-based technologies was decided upon. Without going into technical detail, peerScholar was designed using the Macromedia (now Adobe) Flash application, HTML coding, PHP scripting, and a MySQL database. While discussion as to why each of these software packages was chosen is beyond the scope of the article, it can be noted that cost, security, efficiency, and ease of availability were all factors in the decision process.

Each peerScholar assignment can be fully customized by an instructor, but is always broken into three logical phases emphasizing specific tasks. The first phase involves reading and writing tasks that encourage critical thinking and communication skills. The second phase is where the peer assessment process takes place. Finally, the third phase consists of providing students with the assignment results and feedback. Phase durations are specified by the instructor. Each phase, though housed within the same application, is treated as a separate entity with the second phase opening only after the first closes and the third opening after the second closes. Details of each phase are discussed in the following section.

The reading & writing phase

The reading and writing phase (see Fig 1) is where students access assignment readings along with information about the assignment pertaining to those readings. The details of both the readings and the assignment can be defined by the instructor, and thus, the details provided here correspond to our instantiation. Specifically then, we required our students to read two peer-reviewed articles based on opposing views of the same issue, for example, two articles discussing opposite sides of the topic 'animal use in research'. Articles were made available through an online reader or could be downloaded for offline viewing.

After reading the articles, students proceeded to write two essay-type pieces based on those readings through the peerScholar interface. As a way for students to show their understanding of the readings through summary and to become acquainted with scientific writing practices, they were asked to write an American Psychological Association (APA) formatted abstract for one of the articles. In the second writing piece, students were encouraged to think about the issue presented in the assigned articles and decide where their own opinions lie. They were then asked to write a critical thinking piece that supported their perspective on the issue based on arguments from each article, personal experience, and information acquired from external sources.

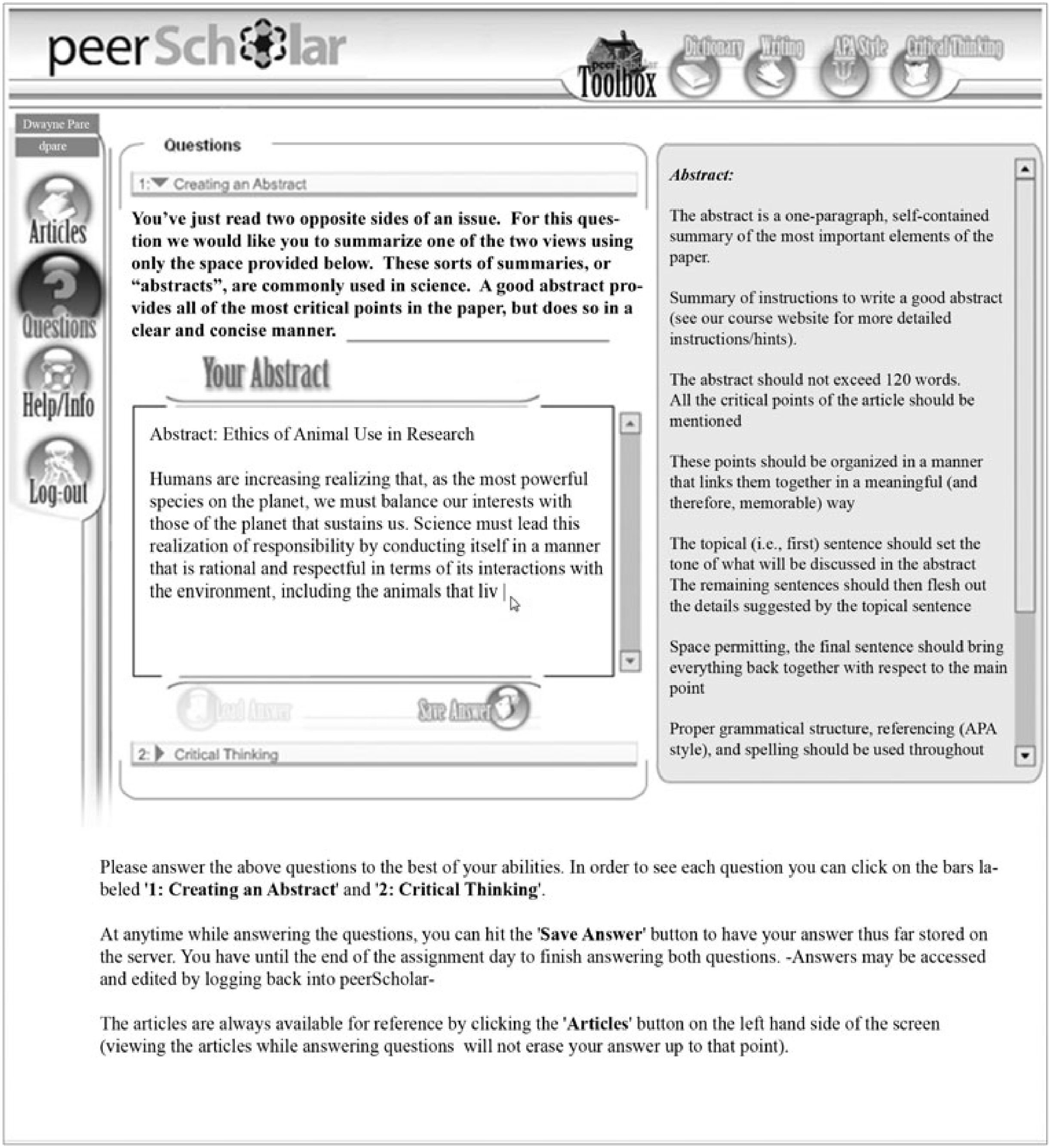
Students were able to save their written pieces to the database at any point during the writing process and could edit their answers freely before the phase's due date. Guidelines for writing an abstract and a critical thinking piece were available throughout the duration of the assignment from peerScholar's help section and the course website. Also available in this and every phase of an assignment was the peerScholar toolbox. The toolbox contained links to resources relevant to the assignment such as a dictionary, APA style guides, the university writing centre, etc. Once the reading and writing phase closed, the system automatically submitted all saved answers in the database, closed access to the first phase, and activated the marking phase of the assignment.

The marking phase

The marking phase (see Fig 2) is where peer assessment occurs. Once this phase opened, peerScholar randomly assigned five abstracts and five critical thinking pieces to each student who participated in the first phase. Written pieces had all identifying information such as student name and student number stripped from them to ensure there were no biases in grading due to any number of factors (e.g. identity, gender, culture, year of study, etc). As each student logged on to peerScholar, they were presented with their assigned 10 anonymous peer-written answers and were required to rate each of them on a scale from 1 to 10. A general rubric was provided for students explaining what the grading criteria were. A number of other references were available from the first phase to serve as a reminder on what was to be included in each written piece. These included the original reading material, the student's own answer, and the abstract and critical thinking guidelines. Students were also asked to provide feedback in the form of a positive and a constructive comment to support the numeric grade given to their peers' assignments. Student marking was expected to be completed in a maximum of two sittings; one for the set of five abstracts and one for the set of five critical thinking pieces. Soon after the marking phase closed, the results and feedback phase was automatically made available.

The results & feedback phase

The results and feedback phase (see Fig 3) is where students accessed the final grade and comments they received on their assignment. Final grades were calculated automatically by peerScholar and were based on averaging the middle three values of the five peer-assigned marks for each written piece, and then averaging the two written piece marks. For example, if a student received marks of 4, 6, 7, 8, and 10 on their abstract, and 3, 7, 8, 9, 10 on their critical thinking piece, the 4 and 10 would be dropped from the abstract mark



You've just read two opposite sides of an issue. For this question we would like you to summarize one of the two views using only the space provided below. These sorts of summaries, or "abstracts", are commonly used in science. A good abstract provides all of the most critical points in the paper, but does so in a clear and concise manner.

Abstract: Ethics of Animal Use in Research

Humans are increasing realizing that, as the most powerful species on the planet, we must balance our interests with those of the planet that sustains us. Science must lead this realization of responsibility by conducting itself in a manner that is rational and respectful in terms of its interactions with the environment, including the animals that liv

Abstract:

The abstract is a one-paragraph, self-contained summary of the most important elements of the paper.

Summary of instructions to write a good abstract (see our course website for more detailed instructions/hints).

The abstract should not exceed 120 words.
All the critical points of the article should be mentioned

These points should be organized in a manner that links them together in a meaningful (and therefore, memorable) way

The topical (i.e., first) sentence should set the tone of what will be discussed in the abstract
The remaining sentences should then flesh out the details suggested by the topical sentence

Space permitting, the final sentence should bring everything back together with respect to the main point

Proper grammatical structure, referencing (APA style), and spelling should be used throughout

Please answer the above questions to the best of your abilities. In order to see each question you can click on the bars labeled '**1: Creating an Abstract**' and '**2: Critical Thinking**'.

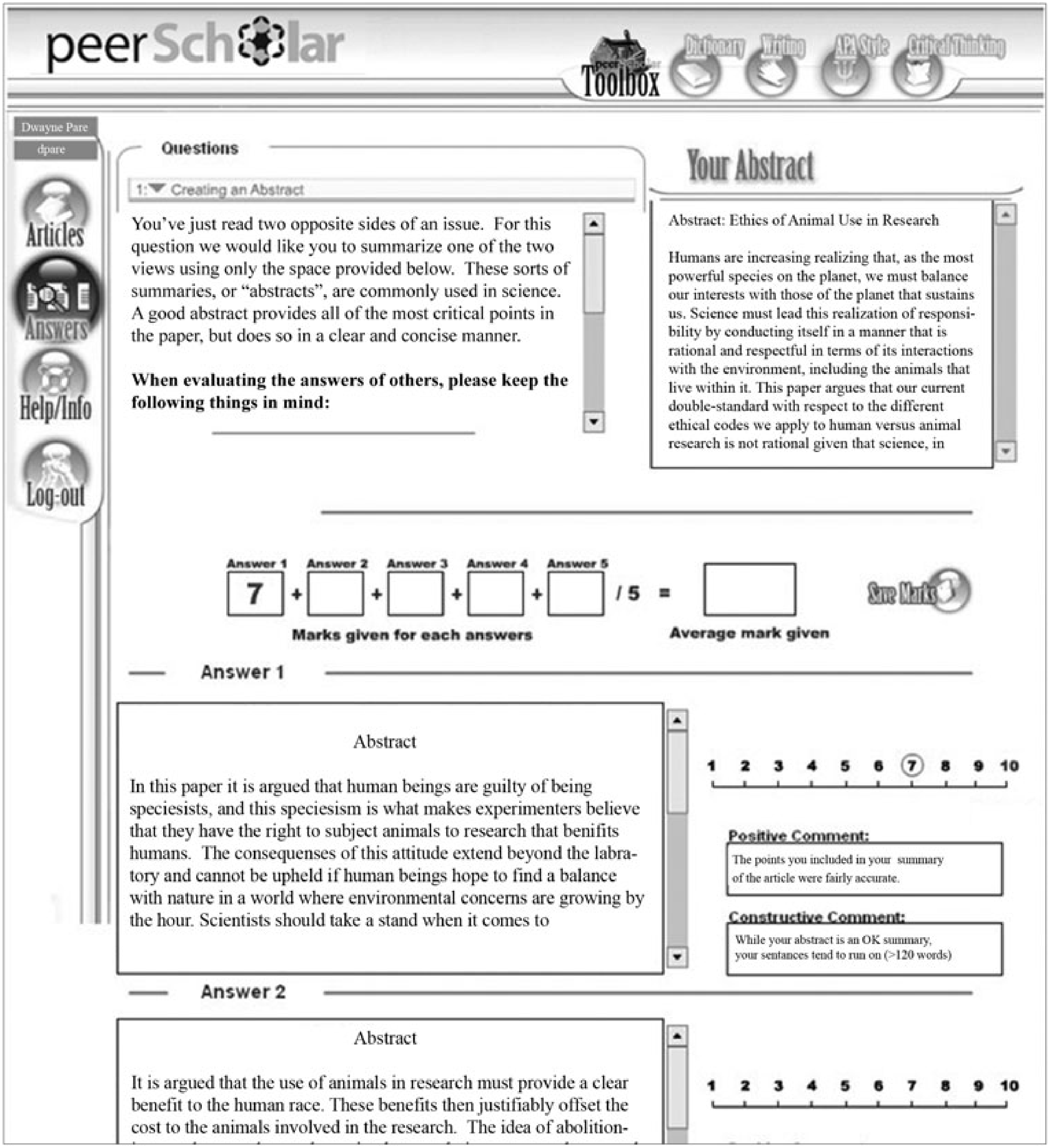
At anytime while answering the questions, you can hit the '**Save Answer**' button to have your answer thus far stored on the server. You have until the end of the assignment day to finish answering both questions. -Answers may be accessed and edited by logging back into peerScholar-

The articles are always available for reference by clicking the '**Articles**' button on the left hand side of the screen (viewing the articles while answering questions will not erase your answer up to that point).

Fig 1 Image of interface for peerScholar: phase 1 (reading and writing).

and the 3 and 10 would be dropped from the critical thinking mark, leaving the student with an average of the other three marks for each exercise, a 7 for the abstract and an 8 for the critical thinking piece. The final grade would then be the average of these two marks, so a 7.5 in this example.

A breakdown of individual peer marks and feedback was also provided for each written piece. This element is provided so students can see peer feedback on each exercise and see how their final grade was calculated. In addition, a bar graph was displayed showing the class averages for the abstract and the critical thinking piece



You've just read two opposite sides of an issue. For this question we would like you to summarize one of the two views using only the space provided below. These sorts of summaries, or "abstracts", are commonly used in science. A good abstract provides all of the most critical points in the paper, but does so in a clear and concise manner.

When evaluating the answers of others, please keep the following things in mind:

Abstract: Ethics of Animal Use in Research

Humans are increasing realizing that, as the most powerful species on the planet, we must balance our interests with those of the planet that sustains us. Science must lead this realization of responsibility by conducting itself in a manner that is rational and respectful in terms of its interactions with the environment, including the animals that live within it. This paper argues that our current double-standard with respect to the different ethical codes we apply to human versus animal research is not rational given that science, in

Answer 1: 7

Answer 1

Abstract

In this paper it is argued that human beings are guilty of being speciesists, and this speciesism is what makes experimenters believe that they have the right to subject animals to research that benifits humans. The consequenses of this attitude extend beyond the labratory and cannot be upheld if human beings hope to find a balance with nature in a world where environmental concerns are growing by the hour. Scientists should take a stand when it comes to

1 2 3 4 5 6 [✓ 7] 8 9 10

Positive Comment:

The points you included in your summary of the article were fairly accurate.

Constructive Comment:

While your abstract is an OK summary, your sentances tend to run on (>120 words)

Answer 2

Abstract

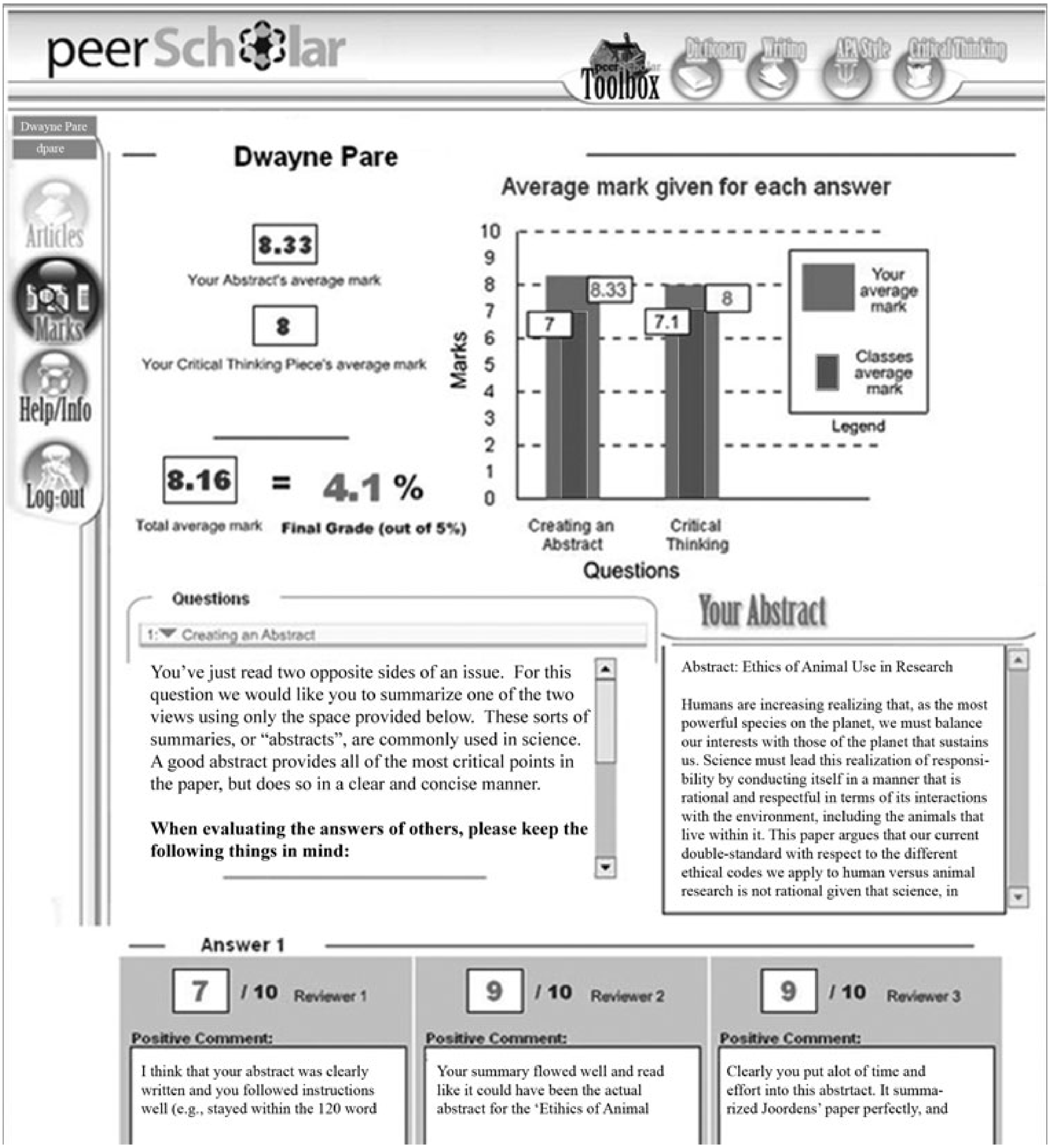
It is argued that the use of animals in research must provide a clear benefit to the human race. These benefits then justifiably offset the cost to the animals involved in the research. The idea of abolition-

1 2 3 4 5 6 7 8 9 10

Fig 2 Image of interface for peerScholar: phase 2 (marking).

alongside the student's own marks for comparison. The final feature of the results and feedback section was an area where the top-ranked abstract and critical thinking pieces, along with their respective marks, were made available for student viewing.

The described peerScholar system was designed to overcome the problems associated with practical implementation of a viable peer assessment process in a large class setting, and it does so well. However, there are still concerns to address before general adoption of such a system would take place in most educational institutions. In his review of past studies of peer assessment, Topping (1998) (see also Falchikov and Goldfinch 2000) concluded that peer assessment is a



Dwayne Pare

8.33 Your Abstract's average mark

8 Your Critical Thinking Piece's average mark

8.16 = 4.1 %

Total average mark Final Grade (out of 5%)

Questions

1: Creating an Abstract

You've just read two opposite sides of an issue. For this question we would like you to summarize one of the two views using only the space provided below. These sorts of summaries, or "abstracts", are commonly used in science. A good abstract provides all of the most critical points in the paper, but does so in a clear and concise manner.

When evaluating the answers of others, please keep the following things in mind:

Your Abstract

Abstract: Ethics of Animal Use in Research

Humans are increasing realizing that, as the most powerful species on the planet, we must balance our interests with those of the planet that sustains us. Science must lead this realization of responsibility by conducting itself in a manner that is rational and respectful in terms of its interactions with the environment, including the animals that live within it. This paper argues that our current double-standard with respect to the different ethical codes we apply to human versus animal research is not rational given that science, in

Answer 1

7 /10 Reviewer 1	9 /10 Reviewer 2	9 /10 Reviewer 3
Positive Comment:	Positive Comment:	Positive Comment:
I think that your abstract was clearly written and you followed instructions well (e.g., stayed within the 120 word	Your summary flowed well and read like it could have been the actual abstract for the 'Etihics of Animal	Clearly you put alot of time and effort into this abstrtact. It summarized Joordens' paper perfectly, and

Fig 3 Image of interface for peerScholar: phase 3 (results and feedback).

reliable and valid method of teaching and assessment. However, the studies considered were conducted using the paper and pencil format. Our first goal is to show that this conclusion also applies to our specific Web-based instantiation of peer assessment. Namely, it is important to both students and faculty that grading done within a peer assessment system is in line with traditional grading practices. Specifically, grades obtained within peerScholar should be similar to expert grades both in terms of their general level, and more importantly, in terms of the manner in which assignments are rank ordered.

Experiment 1

Experiment 1 directly assessed the agreement between peer-given marks obtained within peerScholar and those provided by experts. At the University of Toronto, the standard practice is to have graduate teaching assistants grade undergraduate student assignments. Given this, graduate teaching assistants were used as the expert markers in the following studies. One aspect of this work involved an assessment of the agreement level of these expert markers with each other. This result is interesting in its own right as it provides a reliability index for the traditional approach to marking. Mark agreement between undergraduate peers was also examined and compared to those observed for the expert markers in an effort to determine if peer markers were at an agreement level near that of their expert counterparts.

It was expected that expert markers would show higher levels of agreement than would peer markers because of their previous experience. It was also assumed that individual peer markers could recognize high or low quality assignments, but would not have the experience required to precisely evaluate assignments with fine differences in quality.

Critically though, we assumed that the noise inherent in individual peer marks would be largely eradicated when the marks from several peers were averaged together. Thus, it was hoped that the averaged peer mark, which was the mark a student received for his or her assignment, would correlate well with the expert marks. If this was the case, then the mark generated via the peer assessment process could be considered as providing a good approximation to the marks generated via the traditional grading process.

Methods

Participants

Four psychology graduate students from the University of Toronto participated in the experiment as expert markers for monetary compensation of $100 each. Graduate students were chosen through email and all were currently, or had previously been, a teaching assistant for at least one psychology courses at the University of Toronto. A total of 1143 students from the University of Toronto Scarborough completed the writing assignment as part of their required course work in the Introductory Psychology class (PSYA01). This assignment was worth 5% of participants' final grade in the course.

Apparatus

Students utilized their own computers, or ones freely available on campus, to perform the assignment by linking to the website hosting peerScholar. The peerScholar system used the widely available Macromedia (now Adobe) Flash Player that could be run from most computers with an Internet connection and Web browser with the required Flash Player plug-in[2].

Design and procedure

Each of the 1143 undergraduate students was required to submit two writing pieces during Phase 1 of the first peerScholar assignment. Of the expected 2286 written pieces submitted in the reading and writing phase, 120 were randomly selected to be graded by both expert markers and undergraduate peers. The ratio of evaluators for each written piece remained constant at two expert markers and three peer markers. Pairings of assignment and markers were done randomly with the only stipulation being that no peer mark their own assignment. Since procedures varied for peer markers (i.e. undergraduate students) and expert markers (i.e. graduate teaching assistants), details for each are described in their respective sections.

Peer markers

Through the course syllabus, students in the introductory psychology course were informed that they would be participating in online writing assignments where the primary source of evaluation came from peer assessment. They were also told that expert markers would be involved in marking randomly selected written pieces alongside their fellow students. Information about the peerScholar system was made available to students through the course website. Details regarding the peerScholar assignment procedure were the same as outlined in the previous section of this paper describing peerScholar's implementation.

Instructions for marking were made available online straightforward (see Appendix), and all students were encouraged to ask questions if the need arose. Two points in the instructions should be emphasized. First, students were grading on a scale ranging from 1 to 10 and were asked to try and maintain an average mark of approximately 7 across each set of written pieces they were grading. However, students were told that if presented with 5 poorly written pieces, they would be justified in an average mark below 7 and similarly if they were presented with 5 above average pieces they should have an average above 7. These instructions were included as an attempt to keep peer-given grades in line with the class average without forcing students to grade assignments on the extreme ends of the quality scale with an undeserved mark to meet the average. Second, students were told to take the marking phase seriously given the success of the system, and their own grade, relied on the students' support. To encourage proper grading, students were told that their marks would automatically be monitored by the system for inconsistencies such as patterns (e.g. giving out all 7s) and deviations (e.g. giving a 3 when all other markers gave 9s).

Those undergraduate students whose data were used in this experiment were randomly assigned to grade written pieces in groups of three. For comparison purposes, the average of all three marks was taken to represent the averaged peer mark. The group was also broken up into three sets so that mark agreement comparisons could be made within undergraduate peer markers. Groups of three undergraduate students were different for all of the 120 written pieces.

Expert markers

Each of the four expert markers was asked to log on to peerScholar through a Web link using a password and username provided in an informational email (see Appendix). Detailed instructions on all aspects of the assignment were also included in the email and available on the course website. Before grading assignments, expert markers were asked to take the same assignment preparation steps as undergraduate peer markers in the class by reading all the available information provided by the professor on the course website as well as watching the instructional video. As was the case with the students in the course, all expert markers were encouraged to ask questions if clarification was needed on any aspect of the instructions. Once expert markers had completed the preparation, they were required to log on to the marking phase of peerScholar a total of six times. Each time the expert markers logged on, they were presented with the interface and the same peerScholar marking phase procedures as the peer markers. The written pieces that a specific expert marker received were randomly selected from all available student assignments in the database. In total, each expert marker graded 60 written assignments.

Expert markers were randomly paired so that any writing assignment randomly selected to be marked by one expert marker was guaranteed to be marked by a second expert marker. Marking pairs stayed constant over the course of the experiment.

Results and discussion

The mean mark given by expert markers and by peer markers to the assignment writing pieces, along with percentage of marks in the top and bottom halves of the distribution, can be found in Table 1. An α level of 0.05 was used for all statistical tests in this study. As a preliminary measure, mark means were examined to see

Table 1. Mark descriptives for graduate teaching assistants and undergraduate students in Experiment 1.

	Group	
	Graduate teaching assistants' marks	Undergraduate students' marks
Mean	6.69	6.95
SD	1.03	0.90
n	114	114
% of marks given from:		
1–5	19.2%	11.6%
6–10	80.8%	88.4%

SD, standard deviation.
Note: marks are based on a 10-point scale.

if the average mark given by expert markers was any different than that given by peer markers. An independent-samples *t*-test was conducted using the expert and peer mark means and a significant difference was found, $t(226) = 2.00$, $P = 0.047$, with the peers' mean mark being slightly higher than that of the experts' mean mark.

Expert markers are somewhat tougher markers than peer markers as shown by the statistical difference in mark means. Though the difference is small, experts are prepared to give a greater range of marks as can be seen by the larger standard deviation. More specifically, expert markers appear more willing to give a mark of 5 or below to a written piece as compared to peer markers.

Although these results suggest peer markers are less inclined to give a failing grade, it does not speak to the issue of whether these two groups are giving the same marks to the same types of written pieces. Said another way, if expert markers decide one written piece is of better quality than another, do peer markers' assessments reflect this as well? To address this question, agreement between the averaged expert marks and the averaged peer marks can be examined.

Before discussing agreement levels between peer and expert markers, it is essential to determine what an expected mark agreement level should actually look like. Levels of mark agreement within pairs of graduate teaching assistants (i.e. our expert markers) are the current standard one would expect if looking in a classroom setting where expert markers are doing the grading. Thus, it is important we examine mark agreement between the expert markers participating in this experiment.

To establish this standard, mark agreement within both pairs of expert markers was inspected using Pearson's correlation coefficient. The first pair of experts' marks was found to have a significant positive correlation, $r(59) = 0.46$, $P < 0.001$. Similarly, the second pair of experts' marks was found to have a significant positive correlation, $r(59) = 0.44$, $P < 0.001$. These results indicate that there is significant agreement between expert markers on marks given to each writing assignment. Although these correlations are only moderate, they are typical of agreement levels when grading open-ended written assignments (Blok 1985; Miller 1996).

With an index of mark agreement levels within groups of expert markers, mark agreement within the three sets of peer markers should also be examined. Here, an intraclass correlation coefficient (ICC) was calculated because each assignment was graded by a set of three random peer markers. The ICC for the averaged marks from the sets of three peer markers was 0.41, $P < 0.001$.

Given we now have a standard agreement level to contrast, we can establish mark agreement between expert markers and peer markers to determine if the two agreement levels are comparable. An analysis was done using Pearson's correlation coefficient. A significant positive correlation was found between averaged expert marks and averaged peer marks, $r(113) = 0.27$, $P < 0.003$. It should be pointed out that although this correlation is not high, the fact that it is statistically reliable implies that expert markers and peer markers have a tendency to agree on the quality of written pieces being marked.

Although we have found significant agreement between expert markers and significant agreement between expert and peer markers, our next step was to compute the significance of the difference between the two correlations to determine if the agreement levels are significantly different from each other. To do this, we used Fisher's *z*-score transformation of Pearson's *r* for both pairs of expert markers' correlations and the expert/peer correlations and found that the *z* value of the difference was 1.20 and 1.35, respectively, both smaller than 1.96 and thus not significant at the 0.05 level.

These results tell us that statistically, there is no significant difference between the agreement levels of expert markers with each other, nor is there a significant difference between the agreement levels of expert and peer markers. While this finding is promising in our attempt to create a peer assessment system that is on par with the current standard of grading, we are not yet prepared to make the claim that we have achieved our goal with an agreement level of 0.27 between the expert and peer markers.

Looking again at mark agreement between peers and experts, it might be asked, even though significant, why agreement levels are so low between groups given that within-group agreement is fairly consistent. There may be a few explanations for this, the first being the obvious inexperience on the part of peers with respect to marking written assignments. Though detailed grading instructions were provided to combat this, as any expert grader can attest to, marking subjective written pieces consistently and objectively is hard work even with experience. So even though peer markers may be giving similar grades, they may not be recognizing the same features that are determining marks given by the expert markers.

However, a second possibility is that some students may not have put an appropriate amount of time or effort into the marking phase of the assignment because of the perceived anonymity of the marks they gave to their peers. While students were warned that their marking behaviour might be monitored for inconsistencies and lack of variation, there was no way to prevent a student from hurriedly giving a mark that may not have been as precise as it would have been if more time was taken on assessment.

To ensure the findings in the first experiment were replicable and to account for some of the questions raised by the lack of agreement between peer and expert markers, a replication of the first experiment was done with an additional accountability feature. This feature was labelled ‘Mark the Marker’, and allowed an outlet for students to show praise or concern for a mark they were given by an anonymous marker. It was also expected to add an element of perceived accountability by peers for the marks they gave.

Experiment 2

Methods

Participants

A new group of four psychology graduate teaching assistants from the University of Toronto participated in the experiment as expert markers for monetary compensation of approximately $100. Graduate participants were chosen through email as was described in Experiment 1 and all were currently, or had previously been, a teaching assistant for one or more psychology courses at the University of Toronto. Similar to Experiment 1, 1104 students from the University of Toronto Scarborough’s Introductory Psychology class (PSYA01) participated in Experiment 2 as part of their required course work, where the assignment was worth 5% of the their final grade in the course.

Apparatus

A modified version of the peerScholar system was used in this experiment. The modification involved the addition of a feature to the third phase of the peerScholar interface, labelled ‘Mark the Marker’. This feature allowed students to submit a mark for each of the grades and comments they received from their peers based on how useful they felt the assessments were. Three mark choices were offered to the students through the use of a drop down menu with labels: ‘1 – Not Useful’, ‘2 – Useful’, and ‘3 – Very Useful’.

Design and procedure

The general design and procedure for Experiment 2 was the same as Experiment 1. Differences in procedures for peer markers (i.e. undergraduate students) and expert markers (i.e. graduate teaching assistants) are described in their respective sections.

Peer markers

The design of Experiment 2 was very similar to that of Experiment 1 with the exception of the additional accountability feature built into the peerScholar system (described above). In addition to the instructions provided in the first experiment, students were informed that if their mark was consistently flagged as not useful, the course instructor would examine their marking habits. Submission of a marker’s mark was optional for students and was not required to get full credit for the assignment.

Expert markers

The design of Experiment 2 was nearly identical to that of Experiment 1 for expert markers since they were only required to participate in Phase 2 of the peerScholar assignment. Through assignment instructions, they were made aware of the new Mark the Marker feature that the peer markers would have available in Phase 3 of the assignment.

Results and discussion

The mean mark given by expert markers and by peer markers to the assignment writing pieces, along with percentage of marks in the top and bottom halves of the distribution, can be found in Table 2. An α level of 0.05 was used for all statistical tests in this study. Mark means were again examined to see if the average mark given by expert markers was statistically different than the mean mark given by peer markers. An independent-samples t-test was done using the expert and peer

Table 2. Mark descriptive for graduate teaching assistants and undergraduate students in Experiment 2.

	Group	
	Graduate teaching assistants' marks	Undergraduate students' marks
Mean	6.52	6.85
SD	1.40	0.88
n	116	116
% of marks given from:		
1–5	23.8%	9.6%
6–10	76.2%	90.4%

SD, standard deviation.
Note: marks are based on a 10-point scale.

mark means and a significant difference was found, $t(194) = 2.10$, $P = 0.037$, showing that peers' mean mark was marginally higher than the mean mark given by expert markers.

As was observed in the first experiment, expert markers appear to be marking tougher than peer markers. The difference in mark means is still quite small, but the disparity between expert and peer markers' willingness to give a mark below 5 seems to still be present. It should be reiterated that even though an expert marker is slightly more likely to give a failing mark than a peer marker, this does not address the question of whether there is regular agreement between expert and peer markers on quality of written pieces. To do that, we must once again look at correlations between expert and peer marks.

As was mentioned in Experiment 1, levels of mark agreement within pairs of expert markers is an important measure to have because it gives us the standard agreement we would expect to find in the classroom. Using Pearson's correlation coefficient once again to examine agreement within pairs of experts' marks, we found a significant positive correlation for both pairs of expert marks, $r(59) = 0.55$, $P < 0.001$ and $r(59) = 0.54$, $P < 0.001$. These findings support the notion that expert markers tend to grade similar qualities of written pieces comparably.

As was found in the first experiment, mark agreement within pairs of expert markers was moderate and appears to stay consistent across studies. Mark agreement within sets of peer markers can now be looked at. Here we will use the ICC for our analysis given we have three sets of randomly assigned peer markers for each assignment. The ICC found in this experiment for the averaged marks from the sets of three peer markers was 0.60, $P < 0.001$.

With a standard agreement level between expert markers in place, we can again compare this with the agreement level between peer and expert markers to determine if the levels are analogous. Following from the first experiment, Pearson's correlation coefficient was calculated to determine mark agreement between expert and peer markers. Averaged expert marks and averaged peer marks were analysed and found to have a significant positive correlation, $r(115) = 0.45$, $P < 0.001$. Again, the positive correlation suggests that there is a level of regular agreement between expert and peer markers on quality of written pieces. Of note though is the fact that this correlation is quite a bit larger than the corresponding correlation from Experiment 1.

Although we have once again found significant agreement levels between expert markers and significant agreement levels between expert and peer markers, we wanted to compute the significance of the difference between the two correlations to determine if the agreement levels are significantly different from each other. This was again done using Fisher's z-score transformation of Pearson's r for both pairs of expert markers' correlations and the expert/peer correlations. In this experiment, we found that the z value of the difference was 0.74 and 0.82 respectively, again both smaller than 1.96 and, therefore, not significant at the 0.05 level.

This experiment was designed with the assumption that an accountability feature may translate into increased peer mark agreement. The addition of this feature appears to have played a role in achieving the desired result. Peer markers were much more likely to agree with each other than they had in the previous experiment. In fact, peer-markers agreement levels are surpassing that of expert markers agreement levels. And while high individual peer agreement is not an imperative feature in having a peer assessment system that works, as was discussed in the context of the first experiment, it is important nonetheless in showing that peers are learning to identify and grade features of the writing pieces comparably.

In seeing that the accountability feature increased agreement within sets of peer markers as hoped, mark agreement between expert and peer markers can be revisited. Here it was found that mark agreement between the groups had increased to a moderate correlation from the low one found in the first experiment. Actually, we can now confidently say that the agreement between expert and peer marks has increased in this experiment to a level that is comparable to the one between expert markers, the goal of this research. This increase may be related to the fact that mark agreement between the peers has gone up, which is not to say that increased mark agreement within peer markers automatically equals increased mark agreement between expert and peer markers. In order for mark agreement to correlate between groups, it is required that assignments judged well written by peer markers (even when they all agree) are also judged well written by the expert markers, and the same must be true for poorly written pieces. Following this logic, one might reasonably conclude that peer markers are beginning to recognize details in writing assignments that allow them to distinguish levels of quality comparable to their expert marker counterparts; leading to increased mark agreement within peer marker sets and increased agreement between expert markers and peer markers.

General discussion

At the onset of this paper, we emphasized the importance of open-ended writing assignments to support the teaching of novel thought and clear communication skills. It is these knowledge-use skills that we argue are as important, if not more so, after graduation than the facts acquired in the classroom, regardless of the depth to which they are acquired. However, economic and logistic ‘realities’ have made the inclusion of open-ended written assessments more of a challenge in the context of ever-growing class sizes.

We suggested that peer assessment may provide a viable method for keeping writing assignments in, or returning them back into, large classes. Though benefits of peer assessment were discussed, it was the concerns that needed to be addressed; specifically, concerns regarding the similarity of peer-generated grades to grades provided by expert markers. Across two experiments, we showed that when an accountability feature was added to the peerScholar system, the average grade given by a set of peer markers was similar to the grade given by experts in terms of overall level and rank ordering of assignments. These findings suggest that peerScholar can be used both effectively and fairly in any class context, as has been found previously to be true of pencil and paper versions of peer assessment (e.g. Topping 1998).

A final point regarding the peerScholar system was the acceptance of grades on the students’ part. If the students do not accept this system of grading, there is no way a peer assessment system could work. It was for this reason that while developing peerScholar, student input was encouraged through a number of outlets such as email and questionnaires. Although not detailed here, results from the questionnaires found that students strongly recognized the need for writing assessments in the course. It was also found that students generally liked the concept of peerScholar, they were pleased with the look and feel of the system, and they would like to use it in other university courses. However, there were concerns with peer grading and its fairness. It was partially because of these concerns that the described research was born. Having the ability to tell students the mark they received through peerScholar is comparable to the one that they would have been given by an expert marker lends immense credibility to the process, while simultaneously easing student concerns.

It should be emphasized that concerns for students’ education, and society as a whole, were the motivators for trying to bring critical thinking and communication skills back into the classroom. It is our hope that the peerScholar system, or others like it, will be widely adopted by educational institutions, and will support a re-emphasis of critical thinking and communication skills. To the extent these skills generalize to post-educational successes, the ultimate impact of this effort could be substantial.

Appendix

All instructions and information related to the peerScholar system – including emails, videos, and website material – has been digitally archived at the Web address: http://www.peerScholar.com/articles/grading

Notes

[1] The development name for the peerScholar system was the Reading, Thinking, and Communication Portal (RTCP). The RTCP title was used during the experiments described in this article.

[2] According to a Millward Brown survey conducted in March 2007, the Flash Player is the world's most pervasive software platform, reaching 98% of Internet-enabled desktops in the USA, Canada, UK, France, Germany, and Japan.

References

Ackerman T.A. & Smith P.L. (1988) A comparison of the information provided by essay, multiple-choice, and free-response writing tests. *Applied Psychological Measurement* **12**, 117–128.

Anderson L.W. & Sosniak L.A. (1994) *Bloom's Taxonomy: A Forty Year Retrospective*. University of Chicago Press, Chicago, IL.

Blok H. (1985) Estimating the reliability, validity, and invalidity of essay ratings. *Journal of Educational Measurement* **22**, 41–52.

Bloom D.S. (1956) *Taxonomy of Educational Objectives: The Classification of Educational Goals, Handbook I: Cognitive Domain*. MacKay, New York.

Bloxham S. & West A. (2004) Understanding the rules of the game: marking peer assessment as a medium for developing students' conceptions of assessment. *Assessment & Assessment in Higher Education* **29**, 721–733.

Disraeli B. (1874) *Speech, House of Commons, June 15, 1874*. *Parliamentary Debates* (*Commons*), 3d series, vol. 219, col. 1618.

Ennis R.H. (1993) Critical thinking assessment. *Theory into Practice* **32**, 179–186.

Falchikov N. & Goldfinch J. (2000) Student peer assessment in higher education: a meta-analysis comparing peer and teacher marks. *Review of Educational Research* **70**, 287–322.

Glenberg A.M. (1979) Spacing effects in memory: evidence for a two-process account. *Journal of Experimental Psychology: Learning, Memory and Cognition* **15**, 371–377.

Heywood J. (1989) *Assessment in Higher Education*. John Wiley & Sons, London.

Landauer T.K., Laham D. & Foltz P. (2003) Automatic essay assessment. *Assessment in Education, Principles, Policy & Practice* **10**, 295–308.

Liu J., Pysarchik D.T. & Taylor W.W. (2002) Peer review in the classroom. *Bioscience* **52**, 824–829.

Melton A.W. (1970) The situation with respect to the spacing of repetitions and memory. *Journal of Verbal Learning and Verbal Behaviour* **9**, 596–606.

Miller R. (1996) Mark my words, part 1: teachers. *South African Journal of Higher Education* **10**, 13–24.

Miller T. (2003) Essay assessment with latent semantic analysis. *Journal of Educational Computing Research* **29**, 495–512.

Rada R. & Hu K. (2002) Patterns in student–student commenting. *IEEE Transactions of Education* **45**, 262–267.

Raimondo H.J., Esposito L. & Gershenberg I. (1990) Introductory class size and student performance in intermediate theory courses. *Journal of Economic Education* **21**, 369–381.

Sadler D.R. (1989) Formative assessment and the design of instructional systems. *Instructional Science* **18**, 119–144.

Schneider W. & Shiffrin R.M. (1977) Controlled and automatic human information processing: I. Detection, search and attention. *Psychological Review* **84**, 1–66.

Seabrook R., Brown G.D. & Solity J.E. (2005) Distributed and massed practice: from laboratory to classroom. *Applied Cognitive Psychology* **19**, 107–122.

Shiffrin R.M. & Schneider W. (1977) Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review* **84**, 127–190.

Sims G.K. (1989) Student peer review in the classroom: a teaching and grading tool. *Journal of Agronomic Education* **18**, 105–108.

Topping K.J. (1998) Peer assessment between students at colleges and universities. *Review of Educational Research* **68**, 249–276.

Towns M.H., Marden K., Sauder D., Stout R., Long G., Waxman M., Kahlow M. & Zielinki T. (2001) Interinstitutional peer review on the internet: crossing boundaries electronically in a student-refereed assignment. *Journal of College Science Teaching* **30**, 256–260.

Trautmann N.M., Carlsen W.S., Eick C.J., Gardner F.E. Jr, Kenyon L., Moscovici H., Moore J.C., Thompson M. & West S. (2003) Online peer review: learning science as it's practiced. *Journal of College Science Teaching* **32**, 443–447.

Venables A. & Summit R. (2003) Enhancing scientific essay writing using peer assessment. *Innovations in Education and Teaching International* **40**, 281–290.

White E.M. (1993) Assessing higher-order thinking and communication skills in college graduates through writing. *Journal of General Education* **42**, 105–122.

Williamson D.M., Bejar I.I. & Hone A.S. (1999) 'Mental Model' comparison of automated and human scoring. *Journal of Educational Measurement* **36**, 158–184.

Woolhouse M. (1999) Peer assessment: the participant's perception of two activities on a further education teacher education course. *Journal of Further and Higher Education* **23**, 211–219.

Zheng A.Y., Lawhorn J.K., Lumley T. & Freeman S. (2008) Application of Bloom's taxonomy debunks the 'MCAT myth'. *Science* **319**, 414–415.