2012

Computer-Supported Peer Review in a Law School Context

Kevin D. Ashley
University of Pittsburgh School of Law, ashley@pitt.edu

Ilya Goldin
Carnegie Mellon University - Human-Computer Interaction Institute

Computer-Supported Peer Review in a Law School Context

Kevin D. Ashley and Ilya M. Goldin

I. Introduction

Legal instructors have been urged to incorporate peer reviewing into law school courses for a variety of reasons: the long perceived need to improve law students' writing abilities, persistent calls based on learning theory for more useful feedback in legal education, and the increasing focus on assessing student learning outcomes.[1] Peer review, sometimes also referred to as "peer editing", is "a form of collaborative learning in which students review and critique each other's work."[2] As a pedagogical technique, peer review has the potential to improve legal instruction even with the high student-teacher ratios of most law school courses. It should be especially helpful in legal education, where writing skills are regarded as crucial in analyzing and resolving ill-defined or open-ended problems with alternative reasonable answers that students should be able to explain, compare, evaluate, and justify.[3]

Peer review may bring important benefits to legal education, but only if law school instructors adopt peer review on a large scale, and for that, *computer-supported peer review systems* are crucial. These web-based systems orchestrate the mechanics of students submitting written assignments

---

Kevin D. Ashley is a Professor of Law, University of Pittsburgh School of Law, Professor of Intelligent Systems, Graduate Program in Intelligent Systems, and Senior Scientist, Learning Research and Development Center.

Ilya M. Goldin is a Postdoctoral Researcher at the Human-Computer Interaction Institute, Carnegie Mellon University.

The authors gratefully acknowledge the very helpful comments of Professor David Herring and valuable assistance of Branden Moore.

[1] *See* Cassandra L. Hill, *Peer Editing: A Comprehensive Pedagogical Approach to Maximize Assessment Opportunities, Integrate Collaborative Learning, and Achieve Desired Outcomes*, 11 NEV. L. J. 667 (2011). This article presents a convincing argument for incorporating peer review in legal education and a comprehensive explanation of how to do so. The potential of computer-supported peer review, however, is not addressed.

[2] *Id*. at 671; *see also*, Kirsten K. Davis, *Designing and Using Peer Review in a First-Year Legal Research and Writing Course*, 9 J. LEGAL WRITING INST. 1–18, 1 (2003).

[3] Ill-definedness relates closely to "indeterminacies, or discourse-specific ambiguities and vagueness of cases" in DOROTHY H. EVENSEN ET AL., LAW SCH. ADMISSIONS COUNCIL, DEVELOPING AN ASSESSMENT OF FIRST-YEAR LAW STUDENTS' CRITICAL CASE READING AND REASONING ABILITY: PHASE 2, at 1, 3 (2008): "[W]hen lawyers are researching and reading cases in anticipation of advocacy on behalf of a client, their reading is likely to be sensitive to indeterminacies that are potentially problematic or helpful, either for their contentions or for those of their opponents. These indeterminacies may turn out to be ones that help direct further legal research, that suggest strategically valuable concessions to be made in argument, or that provide viable counterarguments to positions taken by their opponents or by the judiciary in court." That study detected no significant differences between first- and third-year scores on tests measuring skills of dealing with such indeterminacies.

on-line and distributing them to other students for anonymous review, making it considerably easier for instructors to manage. With peer review systems like the Web-based SWoRD (Scaffolded Writing and Rewriting in the Disciplines) system[4] or the Comrade system used in the work reported here, (1) students write their compositions as per the instructor's assignment and submit them to the system. (2) The system distributes the compositions to a group of *N* student peers for review. (3) Using instructor-specified review criteria and forms, reviewers assess the authors' works as per the review criteria by assigning numerical ratings for each criterion and providing written justifications. The reviewers submit their feedback via the system, authors receive the anonymous reviews, and (4) the authors rate the helpfulness of the reviews in so-called back-reviews. (5) Finally, depending on the instructor, authors may revise their drafts.[5]

Beyond the problem of orchestrating mechanics, however, a deeper obstacle to widespread acceptance of peer review in legal education is the question of how well student peer feedback focuses on the legal analytical aspects of the student-authored texts. Can students receive useful feedback on substantive legal issues from peers who, after all, are law students taking the same course and learning the material for the first time, too? Since peers review each other's written work, they are exposed to alternative ways to address the same problems about which at least some of the reviewers are likely to have already thought deeply. Students receive feedback from multiple anonymous reviewers, much more feedback than they are likely to receive from an instructor. The fact remains, however: the feedback is from other students, not from the instructor.

This article reports an experiment assessing the pedagogical utility of computer-supported peer review in a legal educational context. We used a web-based, computer-supported peer review system in an Intellectual Property (IP) survey course to help administer an essay-type take-home midterm exam. We have collected and analyzed data concerning the relationships between instructor- and peer-feedback both on substantive legal issues, such as the claims and issues raised in the law school essay examination problem, and on more general legal writing criteria.

As reported below, our experiments indicate that the student feedback is correlated with independently-provided instructor scores both on the more problem-specific substantive legal criteria and on the more general criteria relevant to legal domain writing. Computer-supported peer review systems enable the efficient collection and analysis of pedagogically useful data about the peer review exercise that can also inform an instructor about how well the exercise and the review criteria met instructional goals.

Our findings, that peer-generated review scores are consistent with those of an instructor, support the belief that peer review can serve as an additional source of useful feedback, including on substantive legal issues. Though generated by students, this feedback is guided by the instructors' explicit grading rubrics. By making their grading rubrics explicit, legal instructors help students to better understand instructors' expectations and receive more and more informative feedback.[6]

This article briefly surveys the promise of peer review for legal education and from the viewpoint of cognitive science and learning theory, and introduces the important topic of reviewing rubrics (Section II). The workings of computer-supported peer review systems are set out in Section III in greater detail along with the advantages of a computerized, web-based approach. In Section IV, two kinds of review rubrics studied here are distinguished: legal domain-related versus problem-specific rubrics and criteria. An experiment using a computer-supported peer-review

---

[4] SWoRD, http://sword.lrdc.pitt.edu. *See generally* Kwangsu Cho & Christian D. Schunn, *Scaffolded Writing and Rewriting in the Discipline: A Web-based Reciprocal Peer Review System*, 48 COMPUTERS & EDUC. 409 (2007).

[5] The systems provide options for making the reviews anonymous and soliciting back-reviews.

[6] Sophie Sparrow, *Describing the Ball: Improve Teaching by Using Rubrics – Explicit Grading Criteria*, 2004 MICH. ST. L. REV. 1, 2–6.

system in a legal course, comparing instructor scores versus peer ratings and analyzing the use of the two rubrics, is presented in Section V, along with the hypotheses tested and measures employed (Section VI). Results likely to interest legal instructors are summarized and their pedagogical implications discussed in Section VII. Conclusions and advice on how legal instructors can use computer-supported peer reviewing with the SWoRD program, are presented in Section VIII.

II. Promise of Peer Review for Legal Education

Landmark critiques such as *Best Practices for Legal Education* document the shortcomings of current assessment practices in American law schools: their over-reliance on the 3-hour end-of-the-semester essay exam,[7] lack of formative feedback during the semester,[8] and lack of detailed summative feedback beyond grades, even on final exams, a tradition deemed "inconsistent with best practices, because it misses an opportunity to use final examinations to enhance student learning."[9] That law school economics have traditionally been based on high student-teacher ratios does not mitigate the fact that "students … link assessment and learning. They want to learn to take exams, and they want feedback so they can improve."[10] They also "need feedback on their ability to self-assess so that they can improve."[11] "Students would benefit from instruction in and application of peer-assessment and self-assessment methods."[12] The authors of the influential *Educating Lawyers* recite that "arguments can be written down, then rehearsed, analyzed, criticized, and in the process, improved…. Feedback from more accomplished performers directs the learner's attention toward improved attempts to reach a goal." The authors note with approval as an educational activity the use of iteration and peer review in preparation of a short advice letter to clients summarizing a 150-page judicial opinion.[13]

Peer review can address some of these shortcomings to a substantial extent.[14] In writing instruction research generally, peer review is frequently recommended as an instructional technique for improving students' writing.[15] From the viewpoint of learning theory, peer review is

---

[7] ROY STUCKEY, BEST PRACTICES FOR LEGAL EDUCATION: A VISION AND A ROADMAP 236 (Clinical Legal Education Association, 2007), *available at* http://law.sc.edu/faculty/stuckey/best_practices/best_practices-full.pdf.

[8] *Id.* at 237; *Id.* at 191 ("The purpose of an assessment can be formative, summative, or both. Formative assessments are used to provide feedback to students and faculty. Their purpose is purely educational, and while they may be scored, they are not used to assign grades or rank students. A summative assessment is one that is used for assigning a grade or otherwise indicating a student's level of achievement").

[9] *Id.* at 261.

[10] *Id.* at 244 (citing Judith Wegner, Thinking Like a Lawyer About Law School Assessment 26 (2003) (unpublished manuscript) (on file with Roy Stuckey)).

[11] GREGORY S. MUNRO, OUTCOMES ASSESSMENT FOR LAW SCHOOLS 124 (Institute for Law School Teaching, Gonzaga University School of Law, 2000).

[12] STUCKEY, *supra* note 7, at 255.

[13] WILLIAM M. SULLIVAN ET AL., THE CARNEGIE FOUND. FOR THE ADVANCEMENT OF TEACHING, EDUCATING LAWYERS: PREPARATION FOR THE PROFESSION OF LAW 98 (2007).

[14] *See* Hill, *supra* note 1, at 670–679.

[15] *See* STEVE GRAHAM & DELORES PERIN, ALLIANCE FOR EXCELLENT EDUC., WRITING NEXT: EFFECTIVE STRATEGIES TO IMPROVE WRITING OF ADOLESCENTS IN MIDDLE AND HIGH SCHOOLS – A REPORT TO THE CARNEGIE CORPORATION OF NEW YORK 15 (2007); Keith Topping, *Peer Assessment Between Students in Colleges and Universities*, 68 REV. EDUC. RES. 249, 269 (1998); Keith J. Topping, *Peer Assessment*, 48 THEORY INTO PRAC. (SPECIAL ISSUE) 20, 23 (2009).

related to approaches that promote active learning, including, repeated opportunities for scaffolded practice in relevant domain-specific tasks, provision of feedback, and reciprocal teaching.[16] Peer review helps focus students on the communicative and rhetorical aspects of their writing. Students receive feedback from audiences who frequently may not already know the content being conveyed[17] and who, having experienced first-hand the consequences of any poor writing strategies, will let the author know. Peer review can also improve student attitudes towards writing.[18]

A pedagogically crucial component of peer review is the use of an instructor-provided rubric according to which reviewers will evaluate the peer author's work.[19] A rubric comprises a suite of prompts focusing reviewers on a set of pedagogically relevant reviewing criteria. The rubric can focus reviewers on domain-independent writing criteria (*e.g.*, insight, logic, and style). More interestingly for legal education, it can also focus reviewers on writing criteria more specific to the legal domain or course (*i.e.*, a "domain-related" rubric or criteria) or even on substantive issues raised in a particular assignment in a law course (*i.e.*, a "problem-specific" rubric or criteria.)

While some instructors may assume that peer feedback is likely to be less helpful than instructor feedback, evidence (from non-law-school courses) shows that combined evaluations from multiple student peer reviewers have moderate to high validity, at least as high validity as single instructor ratings, and moderate to high reliability.[20] *Validity* means the extent to which the rubrics really measure what they are intended to measure. *Reliability* means whether the instrument produces a consistent result when used by different assessors or on different occasions.

In other words, the use of multiple peer reviewers per paper can make up for the peers' lack of expertise relative to an instructor concerning the criteria assessed in these studies. Moreover, students may respond better to feedback from peers than from the instructor.[21] Student peer feedback can be as effective as or more effective than teacher feedback in improving novice's writing,[22] especially when well-developed rubrics and review incentives are employed. Even weaker writers can provide feedback that is useful to stronger writers.[23] There is also some

---

[16] *See generally* L. S. VYGOTSKY, MIND IN SOCIETY: THE DEVELOPMENT OF HIGHER PSYCHOLOGICAL PROCESSES (Michael Cole et al. eds., Harvard University Press,1978) (scaffolded practice). *See also* Hollis Ashbaugh Skaife et al., *Outcome Assessment of a Writing-Skill Improvement Initiative: Results and Methodological Implications*, 17 ISSUES IN ACCT. EDUC. 123, 125 (2002) (domain-specific tasks); Ineke van den Berg et al., *Peer Assessment in University Teaching: Evaluating Seven Course Designs*, 31 ASSESSMENT & EVALUATION IN HIGHER EDUC. 19, 34 (2006) (role of feedback); Annemarie Sullivan Palincsar & Ann L. Brown, *Reciprocal Teaching of Comprehension-Fostering and Comprehension-Monitoring Activities*, 1 COGNITION & INSTRUCTION 117 (1984).

[17] Moshe Cohen & Margaret Riel, *The Effect of Distant Audiences on Students' Writing*, 26 AM. EDUC. RES. J. 143, 152 (1989).

[18] Joyce Katstra et al., *The Effects of Peer Evaluation on Attitude Toward Writing and Writing Fluency of Ninth Grade Students*, 80 J. EDUC. RES. 168, 171 (1987) (dealing with high school students).

[19] *See* Hill, *supra* note 1, at 689–691.

[20] Kwangsu Cho et al., *Validity and Reliability of Scaffolded Peer Assessment of Writing from Instructor and Student Perspectives*, 98 J. EDUC. PSYCHOL. 891, 898, 900 (2006).

[21] Kwangsu Cho et al., *Peer-Based Computer-Supported Knowledge Refinement: An Empirical Investigation*, 51 COMM. ACM 83, 84-87 (2008).

[22] Kwangsu Cho & Charles MacArthur, *Student Revision with Peer and Expert Reviewing*, 20 LEARNING & INSTRUCTION 328, 335 (2010).

[23] Melissa M. Patchan, Peer Review of Writing: Learning from Revision using Peer Feedback and Reviewing Peers' Texts (2011) (unpublished Ph.D. dissertation, University of Pittsburgh) *available at http:// http://d-scholarship.pitt.edu/7542/*.

evidence that the process of providing feedback leads to improvements in the feedback-provider's own writing especially when students provide constructive feedback[24] and explain their ratings.[25]

III. Computer-supported Peer Review Systems

As noted, peer review may realize these benefits for legal education, but only if law school instructors adopt peer review on a large scale. Especially in larger courses, however, it is a logistical challenge for legal instructors to supervise a complex thinking, problem-solving, and writing process like peer review.[26]

In this respect, computer-supported peer review systems help instructors to manage the challenges while adding value to the peer review experience of both students and instructor. Systems like SWoRD or Comrade implement reciprocal peer reviewing of writing. They are designed to support writing practice, especially in large content courses where writing skills are critical but may not be adequately addressed due to class size. As noted, these programs scaffold an entire cycle of writing, reviewing, reviewing reviews, and rewriting; and perform a variety of statistical analyses. They orchestrate the mechanics of students submitting papers, assigning papers to reviewers, returning reviews to authors, and enabling back-reviews where authors comment on the helpfulness of critiques. Both papers and reviews can be anonymized for double-blind review.

Computer support of peer review mechanics is especially important when class size is large, or where peer review reliability is important, which can only be achieved, as the research mentioned above shows, through multiple reviews (*i.e.*, from four to six) for each paper. The above-mentioned systems make it easy to set up multiple reviews per paper by setting a parameter; the systems handle allocating papers to reviewers in a manner that ensures that each author receives an equivalent number of reviews and that the reviewing burden is distributed fairly.

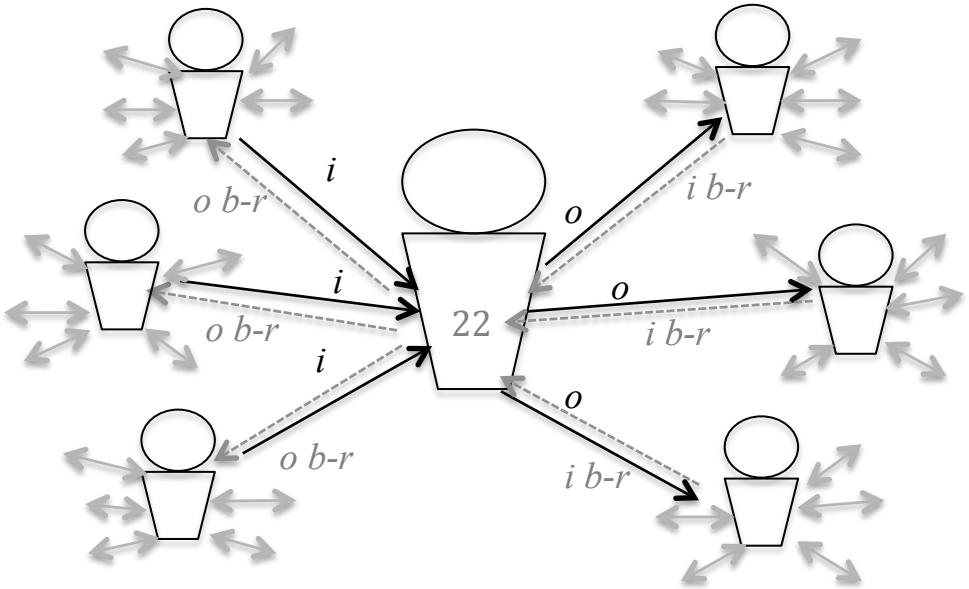
Because each student's paper is reviewed by from four to six other student reviewers, a network of such incoming and outgoing information between and among all the students exists. The peer review network can be visualized as a graph of agent nodes connected by arrows, each of which represents an interaction prescribed by the peer-review process. In Figure 1, every student can be thought of as a "node." When a student reviews other students' essays, there are "outbound" arcs or edges connecting the reviewing student node in the graph to the nodes representing the student authors whose work she is reviewing. In a complementary way, there are "inbound" arcs from the student reviewers to the student author whose work is being reviewed. The feedback a student gives is that student's outbound feedback; the feedback a student receives is referred to as that student's inbound feedback. Similarly, an author's back-reviews of reviewers are the author's outbound (and the reviewers' inbound) back-reviews.

---

[24] Ryan S. Wooley et al., The Effects of Feedback Elaboration on the Giver of Feedback, Poster presented at the 30th Annual Meeting of the Cognitive Science Society 2378 (July 26, 2008), *available at* http://csjarchive.cogsci.rpi.edu/Proceedings/2008/pdfs/p2375.pdf.

[25] Kwangsu Cho & Christian Schunn, *Developing Writing Skills Through Students Giving Instructional Explanations*, *in* INSTRUCTIONAL EXPLANATIONS IN THE DISCIPLINES 207, 214f (Mary Kay Stein & Linda Kucan eds., 2010). Young Hoan Cho & Kwangsu Cho, *Peer Reviewers Learn from Giving Comments*, 39 INSTRUCTIONAL SCIENCE 629, 640 (2011).

[26] See Hill, *supra* note 1, at 671–678.



Author's:
inbound arcs (*i*): N reviewers rate *student22*
outbound arcs (*o*): *student22* rates N others
inbound back-review arcs (*i b-r*): N students rate *student22*'s reviews
outbound back-review arcs (*o b-r*): *student22* rates N others' reviews

Figure 1: Peer-review as Network

Computer-supported peer review systems also assist instructors in defining, reusing, adapting, sharing, and evaluating reviewing rubrics. Instructors who use SWoRD frequently make their rubrics available anonymously for the use of other instructors. The result is a shared resource of rubrics, providing instructors new to peer review systems with sample criteria. These spur users to improve the phrasing of existing criteria, to add new ones, and to adapt the rubrics to the varying requirements of different courses and instructional domains. The systems also make it easier to automatically generate forms, providing ratings scales with which reviewers evaluate an author's work on each criterion in the rubric and fill in their written comments.

Since the systems are computer-supported and web-based, they provide a technological base for adding pedagogically valuable functions, some available currently, such as automated plagiarism-detection and others the subject matter of research and development. For example, as described below, natural language processing, machine learning, intelligent tutoring, and other Artificial Intelligence (AI) techniques are being developed to help reviewers prepare better reviews and authors to better apply the reviews in improving their drafts.

While law school faculty certainly know about classroom applications of peer reviewing,[27] and some faculty may have used computer-supported peer review systems,[28] legal instructors may not

[27] *See, e.g.*, Hill, *supra* note 1; Davis, *supra* note 2; Roberta K. Thyfault & Kathryn Fehrman, *Interactive Group Learning in the Legal Writing Classroom: An International Primer on Student Collaboration and Cooperation in Large Classrooms*, 3 J. MARSHALL L.J. 135 (2009); Cara Cunningham & Michelle Streicher, *The Methodology of Persuasion: A Process-Based Approach to Persuasive Writing*,13 J. LEGAL WRITING INST. 159 (2007); Susan M. Taylor, *Students As (Re)visionaries: Or, Revision, Revision, Revision*, 21 TOURO L. REV. 265, 282 (2005); Sophie Sparrow, *Taking a Small Step toward More Assessments*, 16 L. TCHR. 1, 2 (2009).

yet have understood that statistical analysis of data collected by such systems should provide insights into how well students have grasped the underlying lesson in an exam or writing assignment and how well the review criteria have performed. As illustrated in Figure 1, *student22*'s activities as both generator and receiver of reviews (i.e., the outbound as well as inbound arcs) provide positive and negative evidence of whether he or she understands the concepts employed in the review criteria. The independent assessments of the multiple peer reviewers enhance the reliability of the evidence from the outgoing and incoming arcs. In related work, we have shown that computer-supported peer review can very efficiently process the voluminous and pedagogically informative data generated in the reviewing process, and argued that such systems should be able to summarize and present the data to instructors as the reviewing proceeds.[29]

IV. Review Rubrics: Legal Domain-Related versus Problem-Specific Criteria

As noted, legal instructors may decline to adopt computer-supported peer review systems unless student peer reviews focus on the legal analytical aspects of the student-authored texts being evaluated. The question is whether students can receive reasonable and effective feedback on substantive legal issues from their law school peers taking the same course and learning the material for the first time. An affirmative answer could encourage legal instructors to adopt peer review as an instructional technique, especially in doctrinal courses where class sizes are large and the need for feedback great.[30]

Aside from our work, the studies cited above, concerning the validity and reliability of student peer reviews, did not specifically address legal writing; and the rubrics employed did not focus on criteria specific to the legal domain. The Cho, Schunn, and Wilson study, for example, focused on

---

[28] *See, e.g.*, TURNITIN PEERMARK, https://turnitin.com/static/products/peermark.php (last visited Jan. 19, 2012).

[29] *See* Ilya Goldin, A Focus on Content: The Use of Rubrics in Peer Review to Guide Students and Instructors (2011) Chapter 4 (unpublished Ph.D. dissertation, University of Pittsburgh) (on file with the author) *available at http://d-scholarship.pitt.edu/8375/*; *See also*, Ilya M. Goldin & Kevin D. Ashley, *Eliciting Informative Feedback in Peer Review: Importance of Problem-Specific Scaffolding*, *in* 6094 LECTURE NOTES IN COMPUTER SCIENCE, INTELLIGENT TUTORING SYSTEMS: 10TH INTERNATIONAL CONFERENCE, ITS 2010, PITTSBURGH, PA, USA, JUNE 2010, PROCEEDINGS, PART I, 95 (Vincent Aleven et al., eds.); Ilya M. Goldin & Kevin D. Ashley, *Peering Inside Peer Review with Bayesian Models*, 6738 LECTURE NOTES IN ARTIFICIAL INTELLIGENCE, ARTIFICIAL INTELLIGENCE IN EDUCATION: 15TH INTERNATIONAL CONFERENCE, AIED 2011, AUCKLAND, NEW ZEALAND, JUNE/JULY 2011, 90 (Gautam Biswas et al., eds.); Kevin D. Ashley & Ilya M. Goldin, *Toward AI-Enhanced Computer-Supported Peer Review in Legal Education*, *in* 235 FRONTIERS IN ARTIFICIAL INTELLIGENCE AND APPLICATIONS, LEGAL KNOWLEDGE AND INFORMATION SYSTEMS – JURIX 2011: THE 24TH ANNUAL CONFERENCE 3 (Katie M. Atkinson ed., 2011).

[30] See Hill, *supra* note 1, regarding the desirability of using peer review in doctrinal courses. "And as the ABA's assessment mandate grows, doctrinal faculty also should experiment with incorporating peer editing in their courses to increase student feedback in a workable manner." *Id.* at 669.

> By introducing peer editing in doctrinal or casebook courses such as contracts, torts, or commercial law…, students will learn to work together and professors can strengthen and refine students' collaborative skills over time. But how can doctrinal professors incorporate peer editing in their courses? First, professors must give students more opportunities to write, create, perform, and actively participate in their educations. Second, professors should use peer editing as part of their feedback and assessment plans.

Id. at 704. "To complicate matters further, doctrinal professors tend to have a large number of students in each class…. Thus, to accomplish assessment goals and provide students with much-needed feedback, doctrinal professors can use peer editing as one of the many tools at their disposal." *Id.* at 706.

three writing criteria: flow, logic, and insight.[31] While not irrelevant to legal writing, these criteria fail to address a legal instructor's doctrinal and substantive concerns in evaluating law students' writing.

A. Two Review Rubrics for Legal Instruction

In this work, we explored the use of review rubrics adapted to the domain of legal instruction, particularly legal domain-related writing criteria and problem-specific substantive legal criteria.[32] These are only two ways of developing more specific rubrics adapted to the needs legal education.[33]

In a law school context, for instance, from a pedagogical viewpoint, legal "domain-related" writing prompts are intuitively useful. Our peer-review system prompted reviewers to rate (on a seven-point anchored rating scale and in written critiques) how well the paper under review addresses:

- **issue identification** (identifies and clearly explains all relevant legal issues; does not raise irrelevant issues),
- **argument development** (for all legal issues, applies principles, doctrines, and precedents; considers counterarguments),
- **justified overall conclusion** (assesses strengths and weaknesses of parties' legal positions in detail; recommends and justifies an overall conclusion), and
- **writing quality** (makes insightful, clear arguments in a well-organized manner). [34]

In addition, or instead, legal instructors could use legal "problem-specific" prompts, such as those addressing each legal claim in the problem.[35] For instance, the IP midterm exam problem described below involved five legal claims:

---

[31] *See* Cho et al., *supra* note 20, at 895 ("The *flow* dimension… concerns the extent to which the prose of a paper is free of flow problems…. The *logic* dimension addresses the extent to which a paper is logically coherent (i.e., … links arguments and evidence in a well-organized fashion?). The *insight* dimension accounts for the extent to which each paper provides new knowledge to the reviewer, where *new knowledge* is operationally defined as knowledge beyond course texts and materials."). In Cho & MacArthur, *supra* note 22, the second dimension was "Argument". It dealt with the quality of the claims and support, including the relevance and consideration of counter-arguments.

[32] See Hill, *supra* note 1, at 690, for a discussion of legal instructors' checklists reflecting "the professor's grading rubric for the assignment, … the concepts or material the professor expects to see in the students' work product, … [and] specific criteria students should use to critique their partners' work."

[33] *See* Sparrow, *supra* note 6 (providing an appendix of sample rubrics for legal writing and examinations).

[34] Appendix A.1 provides details concerning the domain-related legal writing criteria.

[35] *See* Sparrow, *supra* note 6, at 1–2 (In a number of respects the problem-specific condition in the experiment described below resembles a computer-supported version of the pedagogical exercise the author recommends here: "Reuse a question from last year's final exam as a writing exercise.

1. Modify the question so that it is limited to the topics students have been studying to date.

2. Assign students to write the answer as homework. …

3. During class, have them … first read through each other's answers. Notice similarities and differences. They are often surprised at how differently they respond to the same material.

4. Give students a checklist – your scoring sheet or list of material you were expecting to see in last year's exams. Have them use the checklist to review another person's answer in greater depth.

5. Walk them through the checklist, e.g. "Find the issue of duty in your neighbor's essay. Identify the standard of care. Compare it to the checklist. Here's what I was looking for and why it is important….").

- breach of nondisclosure/noncompetition agreement ("nda"),
- trade-secret misappropriation ("tsm"),
- two idea misappropriation claims ("idea1", "idea2") and a
- right-of-publicity claim ("rop").

For each claim, in the problem-specific rubric condition, the system prompted reviewers to rate (on a seven-point anchored rating scale and in written critiques) the extent to which the author analyzes the claim, all arguments pro/con, and supporting facts; and cites relevant legal standards, statutes, or precedents. (Appendix A.2 provides details concerning the problem-specific criteria (*i.e.*, the legal claims) and the common set of rating prompts used for each of those.)

Both kinds of rubrics have potential pedagogical value but they also represent tradeoffs. The legal domain-related criteria are quite general and could be applied in many legal writing assignments, yet they do not necessarily capture how well students have mastered applying the course's substantive content to analyze problems. The problem-specific rubrics have to be created specially for, or at least modified to fit, each new problem; but they do reflect the substantive content and application instructors regard as important.

B. Research Questions re Review Rubrics

We investigated a variety of research questions potentially interesting to legal instructors. Since the course instructor (Ashley) graded the papers independently of the peer-reviewing process, it presented an opportunity to assess how closely the student ratings elicited by these rubrics correlated with the instructor's scores, that is, the validity of the peer reviewer ratings. We also examined reliability of the peer ratings across multiple reviewers.

In addition, we wanted to determine whether any differences existed in the effects on student reviewers and authors of supporting reviewing with problem-specific, as opposed to domain-related, rubrics. How sensitive were reviewers to the differences between the two kinds of rubrics, and to the differences among criteria within each rubric? How likely were the ratings from each rubric to be helpful and informative to peer authors? Could we document that students learned by virtue of their experience of peer-reviewing as measured by an objective test assessing knowledge of relevant legal concepts, and was there any differential effect on learning due to being in one condition or the other?

V. An Experiment with Computer-supported Peer Review in a Legal Class

In order to empirically investigate these questions, the authors collected and analyzed data from computer-supported peer-review exercises in connection with the midterm take-home examination in a 2009 Intellectual Property course at a major US law school.[36] Students analyzed a complex legal scenario mainly involving state IP issues. The students then gave and received feedback on each other's analyses using a computer-supported peer review system.

The experiment was designed as a between-subjects treatment. That is, it concerned measuring the values of the dependent variables (described below) for distinct and unrelated groups subjected to each of two experimental conditions. Students in each condition used one rubric to review the work of their peers, either the domain-related rubric (domain-related condition) or the problem-

[36] The experiment was conducted as part of Ilya Goldin's Ph.D. dissertation at the University of Pittsburgh Graduate Program in Intelligent Systems (ISP). *See* Goldin, *supra* note 29. *See also*, Goldin & Ashley 2010, 2011, *supra* note 29.

specific rubric (problem-specific condition). Students only received feedback from reviewers within the same condition. Students received no other training in evaluating peer works.

Fifty-eight (58) second- or third-year law students participated in the study in connection with the Intellectual Property course in which they had enrolled.[37] Students used Comrade, a web-based peer review application that we developed. Beside supporting the peer review cycle described above, Comrade enforces deadlines, checks for acceptable file formats for the submitted papers and reviews, supports students in choosing nicknames to facilitate anonymous peer review, and randomly assigns students to review each other's work using an algorithm that ensures a balanced assignment of papers to review. Conveniently, Comrade also administers objective tests (described below) and user surveys, and maintains all of the data generated in the review activities.

In the course of the study, the students engaged in the following activities:

Day 1: Students picked up the take-home midterm exam and wrote their essay answers.

Day 4: Students turned in paper copies of their midterm exam answers to the law school registrar and uploaded digital copies of their anonymized answers to Comrade from wherever they had an Internet connection. Based on LSAT scores, the investigators made random but balanced assignments of participants to one of the two conditions. Students completed a multiple-choice test (described below); in each condition, half of the students received form A of the multiple-choice test and half received form B.

Day 7 to 11: Students logged in to Comrade to review other students' papers. Comrade assigned each student four papers to review. It was estimated that each review would take about two hours. After completing the reviews, but before receiving reviews from other students, each student completed a second multiple-choice test; those students who earlier completed test form A, now completed test form B, and vice versa.

Day 8: Students logged in to Comrade to receive reviews from their classmates.

Day 10: Students provided back-reviews to the reviewers explaining whether the feedback was helpful. Students also took a brief survey about their peer review experience.

The peer review exercise focused on the essay-type, IP midterm examination question reproduced in Appendix B. The single question involved a fairly elaborate factual scenario designed by the instructor to raise a broad set of legal claims and issues addressed in the first third of the course. In the scenario, a student, Jack, gives his computer programming instructor, Professor Smith, an idea for a musical game iPhone application. Professor Smith hires another student, Barry, to help him develop a variation of Jack's idea. Upon leaving Professor Smith's employ, Barry develops on his own yet another variation of the musical game idea, and sells the application to VeeGames, Inc. for a large sum. In addition, Professor Smith's game application uses visual imagery that evokes the identity of a now-deceased celebrity rock musician, Jimi Hydrox.

Students were expected to apply the material in the casebook readings and classroom discussions in providing advice concerning a particular party's (*i.e.*, Professor Smith's) legal rights and liabilities, given the factual developments. Student's answers were limited to no more than four

[37] For purposes of IRB rules, the study was conducted as an "exempt" evaluation of an educational strategy under 45 CFR Part 46. Students were required to take the midterm examination and to participate in good faith with the peer-reviewing activities. Participation in the study, however, in the sense of allowing their data to be used for research, was voluntary. The syllabus did warn that, "a lack of good-faith participation in the peer-reviewing process as evidenced by a failure to provide thoughtful and constructive peer reviews may result in a lower grade on the mid-term." Only students who gave permission to use their data were included in the study, and students were advised that they could withdraw from the study (*i.e.*, withdraw permission to use their data) at any time. Students were assured that the instructor would grade a physical copy of the midterm examination answer independently of anything they did in the research study. All 58 students in the course gave their permission, and no one withdrew.

typed pages (double- or 1.5-spaced with one-inch margins). Students had a weekend to complete this open-book, take-home midterm examination.[38]

Students were expected to analyze the facts, identify the claims and issues raised, make arguments pro- and con-resolution of the issues in terms of the concepts, rules, and cases discussed in class, and make recommendations accordingly. From the instructor's viewpoint, it was important for students to identify the different kinds of ideas and information that could be protected under the relevant IP laws, and to consider not only the IP claims of a property claimant against others, but also the IP claims of others against the claimant. The instructor focused on plausible claims by the primary intellectual property claimant in the problem, Professor Smith, against various other parties for breach of nondisclosure and noncompetition agreements, trade secret misappropriation, idea misappropriation, unfair competition, and passing off under Section 1125(a) of the federal Lanham Act. In addition, various parties (*i.e.*, Jack, Barry, and the estate of Hydrox) had plausible claims against Professor Smith for idea misappropriation and violating the right of publicity.

Each claim involves somewhat different legal interests and requirements, and thus presents a somewhat different framework for viewing the problem's facts. Since the instructor was careful to include factual weaknesses as well as strengths for each claim, the problem was ill-defined; strong arguments could be made for and against each party's claims. Certain general issues are common across IP claims; for instance, the extent and nature of the alleged infringer's use of the ideas or information required for misappropriation or infringement to exist, and the degree of similarity required between the claimant's ideas and information and those used by the alleged infringer. Students were expected to consider the differences concerning these general issues of use and similarity presented by the different kinds of IP claims, ideas, and information.

This type of essay assignment is perhaps typical of American law school examinations. As is common in legal practice (at least, in litigation), students need to explain the legal issues that arise in a novel factual situation, to connect the issues to the facts of the case, and to make arguments and counterarguments in light of relevant legal principles, doctrines, and precedents.[39]

VI. Hypotheses and Measures

Given the experimental set-up and the nature of the different types of rubrics, domain-related versus problem-specific, we evaluated a number of hypotheses concerning their validity, reliability, whether reviewers were responsive to the rubrics, and whether authors found reviewers' feedback based on the rubrics to be helpful. In this context, *validity* and *reliability* are defined as above.[40] *Responsiveness* refers to whether, in giving their ratings, reviewers treated the rubric's criteria independently or holistically. If a rubric is constructed in such a way that different criteria evaluate

---

[38] The course text was PAUL GOLDSTEIN & R. ANTHONY REESE, COPYRIGHT, PATENT, TRADEMARK AND RELATED STATE DOCTRINES: CASES AND MATERIALS ON THE LAW OF INTELLECTUAL PROPERTY (6th ed., Foundation Press, 2008).

[39] *See* STUCKEY, *supra* note 7, at 180 ("Judith Wegner determined that 'law school exams can best be understood as attempts to measure students' law-related problem-solving expertise.' Problem-based essay exams require students to perform three principle functions – spotting issues, identifying relevant authorities, and applying legal authorities to complex fact patterns – and on occasion a possible fourth, evaluating competing policies or principles. Wegner concluded that such exams, ..., 'appear forthrightly directed to discerning the existence of student expertise as legal analysts confronted with a problem-solving task'") (citations omitted).

[40] That is, *validity* means the extent to which the rubrics really measure what they are intended to measure; *reliability* means whether the instrument produces a consistent result when used by different assessors or on different occasions.

the effect of the same underlying cause,[41] or if reviewers do not differentiate among criteria (*e.g.,* because the differences are too subtle) or interpret criteria differently from the way the instructor intended, then responsiveness will be compromised. *Helpfulness* simply means whether the reviews based on a rubric had a helpful (*i.e.*, formative) effect on authors.

Hypotheses

We tested the following hypotheses.

Rubric Validity: We expected that using each type of rubric, reviewers would produce valid feedback on written works. We expected that three different constructs used to measure student understanding of relevant concepts (*i.e.*, peer ratings, instructor scores on essays, and objective test performance) would converge. We also expected that student peer ratings of the written works would be correlated with those of an independently-trained rater.

Rubric reliability: We expected that reviewer ratings applying the problem-specific rubric would be more reliable than those applying the domain-related rubric. This seemed reasonable because it appeared to be easier to objectively apply the problem-specific criteria, than the domain-related criteria. Reviewers are more likely to agree that an essay is missing a key problem-specific claim or concept, than that an essay lacks good issue identification, good argument development, or another of the more subjective domain-related criterion.

Reviewer responsiveness to rubric: We expected that peer reviewers would be responsive to both types of rubric; that is, in giving their ratings, reviewers would apply the rubric's criteria independently, and not holistically. We did not expect that reviewers would be more responsive with respect to one type of rubric than the other.

Rubric helpfulness: We expected that peer reviewers would produce helpful feedback according to either rubric, since the problem-specific and domain-related rubrics each address important aspects of course material, although in distinctive ways.

Rubric as conceptual scaffold: Finally, we hypothesized that when a rubric supports a reviewer in evaluating another student's work, the rubric may act as a scaffold in focusing the reviewer on key domain concepts; thus, making it more likely that the reviewer will understand these concepts.

In more detail, the experimental set-up generated four basic types of data that served as dependent variables:

1. The peer reviewer's ratings on the seven-point anchored rating scales for each rubric's criteria, domain-relevant and problem-specific (described above and in Appendix A.1 and A.2);
2. The instructor's scores on each student's midterm examination, assigned independently of the peer reviewing activities. The instructor assigned one score to each exam and these scores were not broken down by criteria. As a check on validity, given that instructor scores had not been broken down by criterion, we also trained as a rater a former student, who excelled in the same course in the previous year. The trained rater used the problem-specific rubric to independently rate all papers in the problem-specific condition.
3. The authors' back-review ratings for helpfulness of peer reviews; and
4. The students' scores on two multiple-choice objective tests.

The instructor scored each paper by reading it, making marginal notes based on an answer key (see Appendix C for excerpts), reviewing the answer key to refresh the instructor's memory, and

[41] PAUL B. DIEDERICH ET AL., EDUCATIONAL TESTING SERVICES, FACTORS IN JUDGMENTS OF WRITING (1961); Kristin A. Gansle et al., *The Technical Adequacy of Curriculum-Based and Rating-Based Measures of Written Expression for Elementary School Students*, 35 SCH. PSYCHOL. REV. 435, 446f (2006).

then assigning a gestalt score. The instructor did not score individual elements in the answer key. He created the answer key when he formulated the midterm question in early September, 2009.

To ensure that the trained rater's scoring method was similar to the instructor's, in November, 2010, the rater first scored four papers representing various levels of performance of each criterion, using an answer key prepared by the instructor. After the instructor and the rater discussed the few differences of opinion, the trained rater scored the remaining papers.

The authors were asked to rate each reviewer's feedback for helpfulness on each criterion, the back-review ratings. A single condition-neutral back-review seven-point anchored rating scale (Appendix A.3) was applied for each criterion (*i.e.*, four criteria in the domain-related rubric; five criteria in the problem-specific rubric). The possible ratings ranged from (1) "does not substantively address my analysis" to (7) "identifies all key strengths and problems, and suggests useful solutions to the problems."

In order to obtain an objective measure of something that students might learn in connection with the prescribed peer review activities, the instructor designed a multiple-choice test dealing with the legal claims, concepts, and issues addressed in the first third of the IP course, some of which were involved in the midterm exam question. In preparing the objective test, the instructor did not assume that it would measure the same knowledge and skills that he intended to assess via the midterm essay exam question in this doctrinal course, namely, how to apply the concepts taught thus far in the IP course in solving an ill-defined or open-ended problem.[42] Instead, the hope was simply to assess changes in students' ability to answer certain kinds of objective questions involving roughly (although not exclusively) the legal claims, concepts, and issues that would also be relevant in addressing the midterm exam problem. The instructor designed the multiple-choice test in two equivalent forms (A and B), each with 15 questions. A sample question in two variations is shown in Appendix A.4. The variations were intended to enable giving two different tests of equivalent difficulty as a pretest and posttest to try to assess if students learned anything objectively measurable from the activities in the study. The instructor invited several particularly strong students who had taken the same course in prior years to take the multiple-choice tests, and revised the tests based on their answers and other feedback.

Other data included the participants' Law School Admission Test (LSAT) scores (48 of 58 students opted to allow their LSAT scores to be used), as well as the students' midterm papers.

Measures

From an operational viewpoint, in assessing the hypotheses, we measured rubric *validity* as the correlation between aggregated inbound peer ratings and the summative instructor scores of the exam essays. Given the novelty of the problem-specific rubric, we also assessed the validity of its conceptual distinctions by comparing the student peer ratings of written works against the ratings of the trained rater.

---

[42] *See* Phillip C. Kissam, *Law School Examinations*, 42 VAND. L. REV. 433, 439–441 (1989). Query whether there are objective means of assessing students' ability to apply intellectual property law concepts in solving open-ended problems like the one employed in the IP midterm exam. *See* Linda R. Crane, *Grading Law School Examinations: Making a Case for Objective Exams to Cure What Ails "Objectified" Exams*, 34 NEW ENG. L. REV. 785, 787 (2000). *See also* Greg Sergienko, *New Modes of Assessment*, 38 SAN DIEGO L. REV. 463, 493–505 (2001). In other work, Ashley developed an objective test for evaluating skills of legal reasoning with rules and hypothetical examples, and found some support for its validity but that approach was not employed here. *See* David J. Herring & Collin Lynch, *Teaching Skills of Legal Analysis: Does the Emperor Have Any Clothes?* (University of Pittsburgh Legal Studies, Research Paper No. 2011-16, 2011).

As noted, reliability refers to the consistency of individual student ratings of papers. As recommended in a similar study,[43] we used the intraclass correlation coefficient (ICC) to measure reliability. Here, we report "effective reliability (EFR), one of two versions of ICC.[44] EFR measures the consistency of *k* reviewers combined (i.e., the reliability of the average combined ratings given by the reviewers) where both reviewers and papers are treated as randomly interchangeable. EFR focuses on reviewer consistency rather than requiring exact reviewer agreement. By definition, EFR ranges from 0 to 1.

We evaluated reviewer *responsiveness* by testing if the ratings authors received were correlated across rubric criteria, or if they were independent. We measured *helpfulness* directly by asking student authors to rate how helpful the feedback was to them in a back-review. Regarding a rubric as a *scaffold* for student's *conceptual understanding*, we attempted to compare student understanding of key domain concepts before and after reviewing in terms of his or her performance on the objective tests.

VII. Result Highlights and Implications

This section highlights the experimental results, both positive and negative, likely to be of most interest to legal instructors and discusses their pedagogical implications.[45] The results are summarized in two tables: Table 1, Rubric Validity, Reliability, Reciprocity and Helpfulness, and Table 2, Reviewer Responsiveness to Rubrics.

Within each condition, Table 1 (column 1) shows the peer authors' mean inbound peer ratings across all rating criteria and across all reviewers. In the domain-related condition, the mean involved four rating criteria (x 4 reviewers = 16 inbound ratings). In the problem-specific condition, the mean involved five rating criteria (x 4 reviewers = 20 inbound ratings).

---

[43] *See* Cho et al., *supra* note 20, at 896. ICC computes the analysis of variance of the ratings as the dependent variable, treating both reviewers and papers as independent variables. Where the reviewers are consistent, the variance in ratings depends on the quality of the papers (*i.e.*, the paper effect or the "signal"). Where the reviewers are inconsistent, the variance in ratings also depends on the interaction effect of papers and reviewers (*i.e.*, the "noise".) ICC increases with the mean square of the paper effect and decreases with the mean square of the interaction effect of papers and reviewers; ICC is highest where the signal is strong and there is no noise due to reviewer inconsistency. The ICC calculation enables checking if the ratings pursuant to one type of rubric are noisier than the other, that is, if there is more reviewer inconsistency or, in other words, a greater interaction effect of reviewers and papers.

[44] The other version of ICC, single-rater reliability (SRR) is similar to EFR but it estimates the reliability of one, typical, single reviewer. By definition, SRR also ranges from 0 to 1, and EFR is always greater than SRR.

[45] For a formal presentation of results, *see* Goldin, *supra* note 30; Ilya M. Goldin & Kevin D. Ashley, *Conceptually Focusing Peer Feedback Using Rating Dimensions* (under review).

Table 1: Rubric Validity, Reliability, Reciprocity and Helpfulness

	1. Mean Inbound Peer Rating (Standard Deviation)	2. Validity Correlation of Peer vs. Instructor [95% Confidence Interval]	3. Validity-Check Correlation of Peer vs. Trained Rater [95% Confidence Interval]	4. Reliability (EFR) [95% Confidence Interval]	5. Reciprocity	6. Helpfulness [i.e., Mean Inbound Back-Review Rating] (Standard Deviation)
Domain-relevant Criterion		$r(27) = 0.46$ $p = 0.011$ [0.12, 0.71]				
argument	5.37 (1.20)		N/A	0.65 [0.37, 0.82]	0.28	5.49 (1.38)
conclusion	5.48 (1.09)		N/A	0.34 [-0.2, 0.68]	0.26	5.64 (1.47)
issue	5.37 (1.12)		N/A	0.8 [0.64, 0.9]	0.25	5.34 (1.56)
writing	5.74 (1.36)		N/A	0.48 [0.03, 0.75]	0.38	5.82 (1.31)
Problem-specific Criterion		$r(26) = 0.73$ $p < 0.001$ [0.49, 0.87]				
idea1	4.82 (1.37)		0.43 [0.07, 0.69]	0.51 [0.09, 0.77]	0.27	5.23 (1.65)
idea2	1.98 (1.59)		0.33 [-0.04, 0.63]	0.3 [-0.33, 0.67]	0.21	4.23 (1.93)
nda	4.53 (1.66)		0.71 [0.45, 0.85]	0.86 [0.74, 0.94]	0.23	4.99 (1.73)
rop	2.79 (2.15)		0.81 [0.62, 0.91]	0.95 [0.91, 0.98]	0.24	4.89 (1.84)
tsm	4.84 (1.41)		0.47 [0.12, 0.72]	0.73 [0.49, 0.87	0.24	4.95 (1.79)

Validity

We evaluated the validity of peer-review with the two rubrics by correlating the peer ratings of the students' exam essays to the instructor's independently-assigned essay scores (not shown). Since the instructor's scores were per paper, not per criterion, it was not possible to perform a more fine-grained comparison. For each condition, Table 1 (column 2) shows the Pearson correlations of the means with the instructor's score for the same papers. This measure of the linear dependence between two variables ranges in value between +1 and −1; a value of 0 indicates no correlation, and the closer the value is to +1 (or -1) the stronger the positive (or negative) correlation.

As shown in column 2, for both rubrics the correlations were positive and statistically significant. Thus, each type of rubric is *valid*, as determined by similarity to instructor scores.[46]

[46] Despite the greater value of *r* for the problem-specific rubric and the narrower confidence interval, a Fisher transformation test showed that the problem-specific rubric is not "more *valid*" than the domain-relevant one. This one comparison of two rubrics is too small a sample to recommend the use of a problem-specific rubric over a domain-relevant one.

Since the instructor scores were not broken down by problem-specific criterion, we checked the validity of the peer ratings of the papers from the problem-specific condition in a different way using the trained rater's scores. For each criterion, we assessed the correlation between the trained rater's scores of each paper against the mean inbound peer ratings (See Table 1, column 3). Peer ratings for all but one problem-specific criterion were significantly correlated to the ratings of the trained rater. The sole exception was the second idea misappropriation claim for which peer rating reliability was particularly low.

In sum, the domain-related and problem-specific rubrics were valid in that the mean inbound student peer ratings each correlated strongly with the instructor's aggregate scores on the Intellectual Property midterm exam.

Implications: This is an important result for a course in law. Law is a domain of open-ended problems, where achieving reproducible assessments with plausibly valid criteria is problematic. With instructor-provided review criteria and multiple peer reviews per paper, however, computer-supported peer-review can harness law students' developing expertise and achieve valid peer assessments.

While the validity findings are important regarding each kind of review rubric, it is especially noteworthy with respect to the problem-specific rubric. For each of five legal claims implicit in the facts of the midterm scenario, the problem-specific criteria assess the extent to which the student author "analyzes the claim, all arguments pro/con and supporting facts; cites relevant legal standards, statutes, or precedents". This lies at the heart of what an instructor of a substantive law course seeks to evaluate: how well students understand the substantive law and can apply it to analyzing a complex problem. The results suggest that a legal instructor can construct problem-specific claim- and concept-related substantive criteria for assessing students' written analyses of the problem, and students can use them successfully in the computer-supported peer review to validly assess their fellow students' essays.

Two cautionary notes concerning validity are in order. First, although both types of rubrics may be used for valid peer assessment, further experiments are required before one can conclude that a problem-specific rubric is likely to be valid more often than a domain-related rubric. Second, we found that neither instructor scores nor peer ratings were related to students' LSAT scores.[47] This lack of correlation to LSAT performance is, arguably, inconsistent with the validity results. Our population, however, comprised second- and third-year students, not first-year students. The LSAT is conventionally validated against "the average grade earned by the student in the first year of law school".[48] Bar exam performance, for example, is better predicted by law school grade point average than by the LSAT.[49]

Reliability

As noted above, reliability is measured in terms of the ICC calculation and enables comparing the consistency of the reviewers' ratings using the domain-related and problem-specific rubrics. As indicated in Table 1, column 4, each rubric elicited inconsistent ratings across peer reviewers. While there is no firm dividing line between "good" and "bad" ICC values, both rubrics had some

---

[47] Correlation of LSAT scores with instructor scores was $r(45) = -0.12$, $p = 0.43$, and correlation of LSAT scores with peer ratings was $r(44) = 0.03$, $p = 0.82$.

[48] ANDREA E. THORNTON ET AL., LAW SCH. ADMISSION COUNCIL, LSAT TECHNICAL REPORT 06-02, THE VALIDITY OF LAW SCHOOL ADMISSION TEST SCORES FOR REPEATERS: 2001 THROUGH 2004 ENTERING LAW SCHOOL CLASSES (Oct. 2006), *available at* http://lsacnet.org/LsacResources/Research/TR/TR-06-02.pdf.

[49] LINDA F. WIGHTMAN, LAW SCH. ADMISSION COUNCIL, LSAC NATIONAL LONGITUDINAL BAR PASSAGE STUDY (1998), *available at* http://www.lsac.org/lsacresources/Research/RR/Wightman-LSAC-98.pdf.

criteria that were not reliable. Effective reliability for the domain-related criteria ranged from 0.34 to 0.8 with only issue identification (issue) approaching reasonable effective reliability. Argument was a somewhat distant second. For the problem-specific criteria, effective reliability ranged from 0.3 to 0.95. Only the right of publicity claim (rop) had reasonable effective reliability with nondisclosure agreement (nda) and trade secret misappropriation (tsm) close behind. Effective reliability was relatively low for the two problem-specific criteria dealing with idea misappropriation. The results tend to confirm our expectation that the problem-specific rubric may be easier to objectively apply, and thus be more reliable; but our sample is too small to draw a firm conclusion. We did confirm that low ratings were not a cause of low reliability among reviewers.[50]

In sum, for both problem-specific and domain-related criteria, reliability was problematic; a few criteria in each rubric approached reasonable effective reliability but most did not.

Implications: Reliability is generally desirable in peer review, especially with respect to summative assessment: "[F]or students to take the feedback seriously, the ratings need to count for actual grades, and the validity and reliability of the grades depends upon there being ratings from multiple reviewers."[51] In general, reliability may be improved by increasing the number of peer reviewers,[52] and by calibrating their rating techniques.[53] In this respect, it may be easier to teach students how to apply problem-specific criteria (*e.g.*, the elements of an idea misappropriation claim or a claim for breach of the right of publicity), rather than the more abstract domain-related criteria (*e.g.*, what constitutes a good legal argument in general).

The importance of reliability in peer review may be overstated, however.[54] The lack of reliability of the peer assessments across criteria in each rubric (i.e., the lack of consistency of the peer ratings – their noisiness) is not necessarily a surprising result where the rubrics are applied to comparatively ill-defined problems, like those in a law school essay exam. Since ill-defined problems present alternative reasonable answers that need to be explained, compared, evaluated, and justified, the problems and answers may typically be framed in multiple ways, and disagreements with respect to conceptual issues are frequently legitimate. In addition, reviewers may encounter new ways of viewing the same material in which the reviewers have already invested considerable time and energy (because they also took the exam). Their reviews of other authors' exam essays may call into question their own answers, especially since, given the range of instructor-assigned scores, some authors may have completely missed certain legal issues implicit

---

[50] The domain-related rubric elicited ratings at the high end of the anchored rating scale in all criteria. The problem-specific rubric elicited ratings at both high and low ends of the rating scale, and sometimes with a high variance (Table 1). Despite this, the effective reliability of problem-specific ratings did not suffer in comparison to the reliability of the domain-related ratings. The two problem-specific concepts that had the lowest mean inbound peer ratings, namely the second idea misappropriation claim (idea2) and right of publicity (rop), were, respectively, the least and most reliable problem-specific concepts.

[51] *See* Cho & Schunn, *supra* note 4, at 414. The general tension between validity and reliability has been noted in peer assessment: peer review may demonstrate a "convergence of different raters on a 'single truth'", or it may "uncover the presence of multiple perspectives about the performance being assessed, which do not necessarily have to agree." Peter J. Miller, *The Effect of Scoring Criteria Specificity on Peer and Self-assessment*, 28 ASSESSMENT & EVALUATION IN HIGHER EDUC. 383, 390 (2003).

[52] *See* Cho & Schunn, *supra* note 4, at 419.

[53] Arlene A. Russell, *Calibrated Peer Review: A Writing and Critical-Thinking Instructional Tool*, *in* INVENTION AND IMPACT: BUILDING EXCELLENCE IN UNDERGRADUATE SCIENCE, TECHNOLOGY, ENGINEERING AND MATHEMATICS (STEM) EDUCATION 67 (American Association for the Advancement of Science, 2004), *available at* http://www.aaas.org/publications/books_reports/CCLI/PDFs/03_Suc_Peds_Russell.pdf.

[54] Ngar-Fun Liu & David Carless, *Peer feedback: the learning element of peer assessment*, 11 TEACHING IN HIGHER EDUC. 279, 282 (2006).

in the exam problem. Under these circumstances, a high degree of reliability might actually be more surprising than the lack of it.[55]

Reviewer responsiveness to rubric

For each type of rubric, we tested whether reviewers applied the rubric's criteria independently and not holistically, or whether the criteria were redundant.

In order to assess responsiveness to criteria, the mean inbound peer ratings per criterion within each student were correlated, yielding six pairwise correlations for the domain-related rubric and ten for the problem-specific rubric (Table 2).[56] As indicated, all of the correlations between mean inbound ratings in the domain-related condition were statistically significant. By contrast, in the problem-specific condition, only two pairs of criteria were highly correlated (the trade secret misappropriation claim (tsm) and the first idea misappropriation claim (idea1), and the tsm claim and breach of non-disclosure agreement (nda)).

Table 2: Reviewer Responsiveness to Rubrics

Domain-related criterion	**argument**	**conclusion**	**issue**	**writing**
argument		0.72*	0.69*	0.65*
conclusion			0.61*	0.77*
issue				0.67*

Problem-specific criterion	**idea1**	**idea2**	**nda**	**rop**	**tsm**
idea1		-0.04 (0.14)	0.31 (-0.03)	-0.01 (0.22)	0.70* (0.39*)
idea2			-0.07 (0.11)	-0.11 (0.00)	-0.21 (0.15)
nda				0.18 (0.02)	0.46* (0.21)
rop					0.16 (-0.09)

We confirmed that the relative lack of correlation among the problem-specific criteria was not due to the fact that peer reviewers missed important relationships among the criteria, as could happen if the conceptual issues were too difficult for peer reviewers to assess. We computed the correlations among the trained rater's scores for each pair of criteria in the same manner as for the mean inbound peer ratings. As indicated in Table 2 (in parentheses for problem-specific criteria) no significant pairwise correlation, found by the trained rater, was missed by the peer reviewers. Of the two significant pairwise correlations that were present according to the peer reviewers, one

[55] *See* STUCKEY, *supra* note 7. "As currently used, the end-of-the-semester essay exam is neither valid, nor reliable, nor fair." *Id.* at 177. "We join Judith Wegner and other scholars in encouraging law professors to develop and apply explicit grading criteria to minimize the risk of unreliability in assigning grades." *Id*. at 181.

[56] Pairwise Pearson correlations are shown between mean inbound peer ratings for domain-related criteria and problem-specific criteria. For problem-specific criteria, correlations are shown (in parentheses) of author work according to the trained rater's ratings. Asterisks indicate correlations significantly higher than zero at $\alpha = 0.05$.

was also significant according to the trained rater (idea1 vs. tsm); the other was not significant according to the trained rater (nda vs. tsm). In the main, the peer reviewers distinguished among problem-specific criteria similarly to the trained rater.

In sum, the higher correlations for the domain-related criteria indicate less independence (i.e., more redundancy). The lower correlations for the problem-specific criteria indicate more independence; the problem-specific criteria are less redundant.

Implications: Since problem-specific support to reviewers leads to ratings that largely do not correlate with each other across criteria, such ratings are less likely to be redundant, and are more likely to be informative. Authors receiving domain-related ratings may have found them to be redundant.

Several possible explanations exist for inter-criteria correlations. First, some of the criteria may be intrinsically interdependent. The peer reviewers appeared not to discriminate when applying the domain-related criteria; justifying overall conclusions may necessarily depend on developing arguments, which, in turn, may depend on identifying issues. Regarding the problem-specific criteria, claims of trade secret misappropriation (tsm) frequently arise in the context of breach of non-disclosure and non-competition agreements (nda), one of two correlations found in that condition.

Second, the criteria may be correlated in terms of the behavior or characteristics of student writing they describe. The students' essays may have been quite homogeneous with regard to rhetorical aspects of writing quality. If a student writes in a well-organized way, it is likely that the student will also identify issues, develop arguments and justify overall conclusions, even if one does not directly cause the other. Arguably, students acquire the skills of identifying issues, developing arguments, justifying overall conclusions, and organizing writing at the same time and in combination.

Third, peer reviewers could have rated each other inaccurately, although this is unlikely since both types of ratings are valid with respect to instructor scores. Second and third year law students are likely to be sufficiently familiar with what it means to make legal arguments to apply the domain-related rubric accurately. Of course, they were novices in the subject matter of Intellectual Property; but comparison with the trained rater's scoring indicated that students did not miss important relationships among the problem-specific criteria, which could have led to the low inter-criterion correlations.

In any event, that the domain-related criteria elicit more redundant ratings than the problem-specific ones is a reason for instructors assigning peer review exercises to include the latter as providing students more informative ratings.

Reciprocity and rubric helpfulness

As noted, we measured *helpfulness* using back-review ratings. Before accepting these ratings at face value, however, we first assessed the effects of author-reviewer reciprocity. Sometimes, peer reviewers and authors engage in tit-for-tat reciprocal behavior.[57] That is, authors who receive high inbound peer ratings may respond with high back-review ratings. Those who receive low inbound peer ratings may respond with low back-review ratings. We tested for apparently reciprocal behavior. In addition, we expected to see less reciprocal behavior among authors receiving problem-specific feedback. Since problem-specific criteria may be easier to apply objectively than domain-related criteria, authors may find it easier to evaluate such objective feedback on its own merits.

[57] *See* Cho & Schunn, *supra* note 4, at 415.

Reciprocity was operationally defined as the correlation between the peer ratings given by reviewers and back-review ratings given by authors in response. It was computed using Kendall's τ, "the difference between the probability that the observed data are in the same order for the two variables versus the probability that the observed data are in different orders for the two variables."[58] Reciprocity was fairly constant across rating criteria (Table 1, column 5). Thus, we found a small but statistically significant amount of reviewer-author reciprocity for each rubric.

Given this small but statistically significant bias due to tit-for-tat back-review ratings, when we compared feedback helpfulness between the two conditions, we adjusted for inbound peer ratings. After this reciprocity adjustment, each rubric was found to elicit helpful feedback most of the time as indicated by the mean helpfulness ratings (on a seven-point scale) for each criterion. (Table 1, column 6). An ANOVA, comparing all problem-specific versus all domain-related back-review ratings, showed that rubric type was not a significant predictor of the back-review rating,[59] that is, the helpfulness of feedback did not vary across the two rubrics.

Despite the lack of statistical evidence that rubric type affected helpfulness, we did note that authors rated domain-related feedback as 6 or 7 more often than problem-specific rubric. Additionally, problem-specific feedback was rated 3 or below more often than domain-related feedback. The higher back-review ratings could mean that the authors felt that the domain-related feedback more often contained praise. It has been observed that students rate praise as helpful[60] even though praise is not associated with authors' actually implementing reviewer feedback in a subsequent draft.[61] Interpreting the higher scores as praise is consistent with the back-review ratings definitions, Appendix A.3, according to which the higher-rated feedback "identified most key problems" in their writing, "suggested useful solutions," and "identified key strengths."

In order to probe why problem-specific feedback was sometimes unhelpful, we analyzed all 92 comments from peer authors that were paired with back-review ratings of 3 or lower. The most frequent explanations of low back-review ratings were that the reviewer's feedback was empty or almost empty (19), that the reviewer missed or misunderstood key parts of the author's argument (20), or that the reviewer's feedback was correct, but suggested no solutions (33).

Problem-specific authors chose not to give back-reviews more frequently than domain-related reviewers. We analyzed the 19 cases where an author gave a written back-review comment and omitted a back-review rating. The comments seemed to fit well with the back-review scale, but the authors chose to omit ratings nonetheless.

Implications: As noted, authors' back-review ratings for both types of rubrics were affected by a small but statistically significant amount of reviewer-author reciprocity. In addition, domain-relevant criteria elicited reviews containing more praise and resulted in more frequent high-end helpfulness ratings, while the problem-specific reviews elicited more frequent low-end helpfulness ratings. This suggests that an emotional aspect of feedback helpfulness may be especially important in this law school exam context. Authors receiving low-level ratings on problem-specific criteria are faced with the uncomfortable realization that they may have missed an important legal claim or issue implicit in the problem facts of the still-ungraded exam. Students expect that failures to identify such claims or issues will have a direct negative impact on an exam grade. This may cause

---

[58] THOMAS HILL & PAWEL LEWICKI, STATISTICS: METHODS AND APPLICATIONS 387 (StatSoft, 2006). Problem-specific reciprocity was found to be $\tau(579) = 0.27$, $p < 0.001$, and domain-related reciprocity $\tau\ (463) = 0.30$, $p < 0.001$.

[59] $F(1,830) = 2.69$, $p = 01.0$.

[60] Kwangsu Cho et al., *Commenting on Writing: Typology and Perceived Helpfulness of Comments from Novice Peer Reviewers and Subject Matter Experts*, 23 WRITTEN COMM. 260, 280 (2006).

[61] Melissa M. Nelson & Christian D. Schunn, *The Nature of Feedback: How Different Types of Peer Feedback Affect Writing Performance*, 37 INSTRUCTIONAL SCI. 375, 383 (2009).

anxiety when it is too late (in this exam context) for the author to do anything about it and may color authors' perceptions of helpfulness. Deficiencies in the domain-related criteria, by contrast, are less specific and less directly tied to substantive legal concepts tested in the exam; there is still hope that the instructor may look past such a deficiency unless it involves a gaping hole in the student's analysis.

One could eliminate reciprocity in computer-supported peer review systems by presenting authors with only a reviewer's comments but not the reviewer's ratings.[62] The ratings, however, communicate useful information. The ratings scheme communicates the structure of the criteria, and students can gauge their level of current performance within a range of performance ratings.

From an instructor's perspective, low ratings of helpfulness of problem-specific feedback, like low inbound peer ratings, inform the instructor that a particular problem-specific concept has proved challenging for students. Low back-review ratings may indicate that reviewers are struggling to give helpful feedback regarding a problem-specific concept, which may indicate that those reviewers do not understand the legal concept.

These last points may be a reason to design peer review exercises so that an opportunity always exists for authors to rewrite their drafts in response to reviewers' comments. In this way, the peer feedback serves a more formative function. This design would not be consistent with the summative role of a midterm examination as commonly seen in the law school context, and may not fit the current common plan for substantive law courses. For instance, integrating a writing assignment with a second draft opportunity would have been a radical departure from the IP class format, where there is pressure to speedily survey the range of state and federal IP law, and where instituting a midterm was already a "radical" change. Arguably, all of this should be reformed as well, but the aim of this study was more modest: to evaluate the potential contribution to legal education of one particular tool – computer-supported peer review and the effects of different types of review rubrics.

A general limitation of the helpfulness results reported here is that they are based only on the peer ratings and not on the peer comments. Since the ratings are numeric and linked to well-defined scales, they are both easier to analyze statistically and likely to be meaningful. The comments, being textual, are harder and more time-consuming to analyze. However, analysis of the comments could confirm whether they really matched rubric criteria, discussed problems as well as suggested solutions, and revealed differences between the two rubrics and the extent to which reviewers using the domain-relevant rubric addressed problem-specific concerns and vice versa. We hope to conduct such analyses in the future.

Rubric as conceptual scaffold for learning

Based on their objective pre- and posttest scores, we attempted to measure students' comprehension, both before and after students gave peer feedback to each other, of conceptual knowledge from the first third of the IP course, much of which was relevant to the exam problem. Student comprehension was measured after the students wrote their exam answers and again after submitting their reviews. As noted, the objective tests covered roughly the same legal claims, concepts, and issues that were addressed in the exam question, but involved different short factual scenarios and solely a multiple-choice format. Also, as illustrated in Appendix A.4, there were two test forms: half of the students in each condition took test form A, and half took test form B. For each test form, each student's number of questions answered correctly was tallied.

---

[62] Kwangsu Cho & Bosung Kim, *Suppressing Competition in a Computer-Supported Collaborative Learning System, in* 4553 LECTURE NOTES IN COMPUTER SCIENCE, HUMAN-COMPUTER INTERACTION: HCI APPLICATIONS AND SERVICES – 12TH INTERNATIONAL CONFERENCE HCI INTERNATIONAL 2007, BEIJING, CHINA, JULY 22-27, 2007 PROCEEDINGS, PART IV, 208, 211 (Julie A. Jacko, ed.).

We could not empirically confirm that the instructor-designed objective test was a valid measure of student understanding (where validity was measured by similarity to instructor scores of the students' essays). For each test form, every student's number of questions answered correctly was tallied. The Pearson correlations of these counts with the instructor's score of the student's papers were not significant for either test form.[63] Pre-test performance also did not correlate to peer ratings that were elicited via the domain-related rubric and problem-specific rubric. In addition, although the test forms were intended to be equivalent, statistical analysis of the students' performance indicated differences between them; the two test forms were neither internally consistent, nor consistent with each other. Thus, although the objective tests had strong face validity and pedigree (with questions adapted from published examples[64]), our efforts at assessment with them proved problematic.

Implications: From an instructional viewpoint, we hoped that the exercise of reading and making sense of the somewhat divergent information, a frequent aspect of the reviewing process, would be pedagogically fruitful.[65] Student reviewers applying especially the more detailed, problem-specific rubric might encounter different authors' approaches, leading some of them to see the problem in different ways, and learn something about the legal claims and how they apply. This was the prime motivation for the hypothesis concerning the rubrics as conceptual scaffold.[66]

Unfortunately, the learning measure built into the experiment, students' performance on the objective pre- and post-tests, turned out to lack validity, as measured against instructor's independently assigned essay scores, and both versions of the test had problems with consistency. Given the problems with the instrument, change in students' conceptual understanding could not be measured.

Summary

In this experiment we compared peer review of law school exam essays using two types of rubrics to guide reviewers and authors: a conceptual, problem-specific rubric versus a more general domain-related rubric. According to the results, both kinds of review rubrics produced valid assessments by students of their peer's writing, as measured against the instructor's independently assigned exam essay scores. Validity of student assessments with the problem-specific rubric was confirmed with those of a trained rater. The peer assessments produced by each rubric lacked uniform reliability across criteria. For most review criteria in either rubric, reviewers' ratings were fairly inconsistent as measured by the Intra-Class Coefficient. Reviewers were more responsive with respect to the problem-specific than the domain-related criteria: peer assessments with the domain-related rubric's criteria were highly correlated and lacked independence. Peer

[63] $r(28) = 0.00$, $p = 0.99$, and $r(26) = 0.20$, $p = 0.31$.

[64] STEPHEN M. MCJOHN, EXAMPLES & EXPLANATIONS: INTELLECTUAL PROPERTY (3d ed., Aspen, 2009).

[65] Jennifer Wiley & James F. Voss, *Constructing Arguments From Multiple Sources: Tasks That Promote Understanding and Not Just Memory for Text*, 91 J. EDUC. PSYCHOL. 301, 309 (1999); Danielle S. McNamara et al., *Are Good Texts Always Better? Interactions of Text Coherence, Background Knowledge, and Levels of Understanding in Learning from Text*, 14 COGNITION & INSTRUCTION 1, 34 (1996); Kwangsu Cho & Young Hoan Cho, *Learning from Ill-Structured Cases, in* PROCEEDINGS OF THE 29TH ANNUAL COGNITIVE SCIENCE SOCIETY CONFERENCE 1722 (D. S. McNamara & J. G. Trafton eds., 2007), *available at* http://csjarchive.cogsci.rpi.edu/proceedings/2007/docs/p1722.pdf.

[66] Prof. David Herring made the interesting observation that students also learn from reading the reviews they receive, suggesting that the post-test should have been scheduled sometime after reviews had been received. That also seems worth trying.

assessments with the problem-specific rubric did not show high inter-criteria correlation (i.e., were independent) as was confirmed by comparison with a trained rater's assessments. Authors found the reviews from each rubric to be usually helpful. The legal-concept-oriented multiple-choice tests were not shown to be valid measures according to the instructor's exam essay scores. Given problems with the multiple-choice tests' validity and unequal difficulty, we found no evidence that students' learned from the peer review activities with either type of rubric.

VIII. Conclusions

When it comes to assessing legal coursework performance, instructors face a quandary. Intuitively, essay questions seem to lend themselves to evaluating legal analytical problem-solving skills, but the way they have been criticized administered in law schools has been criticized for lacking validity and reliability. From a practical viewpoint, multiple-choice tests seem to be gaining favor in law schools, not least because they reduce both grading time (while still requiring considerable time to construct) and grading subjectivity, and can be shown to be reliable. The question is whether multiple-choice tests in law school are valid, especially if an instructor's goal is to assess "student expertise as legal analysts confronted with a problem-solving task."[67] In this experiment, our inability to demonstrate the validity of the multiple-choice tests is indicative of this problem. The use of explicit criteria for evaluating legal writing has been recommended to improve validity and reliability,[68] but probably at the expense of complicating both exam preparation and the already onerous process of grading essay exams.

Computer-supported peer-review with explicit criteria provides another approach to designing assessment techniques to address these tradeoffs. That is the design space we have explored with peer-review, using both domain-related and problem-specific criteria. Assessments with both rubrics were shown to be valid, as compared to the instructor's aggregate scores on the midterm exam in Intellectual Property law. While lacking consistent reliability across criteria, some criteria in each rubric demonstrated or approached effective reliability, which may be as good as can be hoped in dealing with typically ill-defined legal problems. Thus, the evidence of this study suggests that law students can perform criteria-based assessment for each other in the peer review process (and hopefully learn from it, although our study failed to establish learning from reviewing).

Computer-supported peer review could help legal instructors improve legal education in their own classrooms by providing more writing opportunities, such as a mid-semester writing exercise with feedback based on problem-specific and legal domain-related criteria. Ideally, it prepares students for the final examination, allows them to see how their peers analyze the same problems that they also have been studying, provides more feedback than students usually get in legal education, and enables them to compare their progress in mastering the course material to that of their fellow students.

A. Choosing between Review Rubrics

The results reported here have significance for the choice of reviewing rubrics and for the design, implementation, and evaluation of computer-supported peer review as a mechanism for teaching law students skills for analyzing open-ended legal problems.

The choice between rubric types depends on the instructor's goals. Some law courses focus on improving legal writing as an end in itself. For these, it is important to distinguish the writing and critiquing skills that make up a domain-related rubric, and it may be helpful to collapse the various

[67] *See* STUCKEY, *supra* note 39 (citing Wegner).

[68] *See* STUCKEY, *supra* note 55.

problem-specific conceptual issues. Most law courses, however, focus on substantive legal subject matter and view legal writing as a window into the students' analytical understanding and skills of applying targeted concepts in problem-solving. For these courses, it is helpful to tease apart problem-specific conceptual issues.

At some level, peer reviewers applying either of the two types of rubrics are evaluating the same phenomenon, the quality of legal argumentation about solving a problem; but the rubrics slice through this common phenomenon in different ways. Although each rubric places value on identifying and making reasoned arguments about conceptual issues, the problem-specific rubric focuses on the particular conceptual issues involved in the problem, namely, the relevant claims; the domain-related does not specify these concepts but lets the reviewer decide which concepts raised by the author are relevant, and treats in the aggregate the concepts and what the authors do with them. While reviewers applying the domain-related rubric might naturally consider an author's treatment of the relevant legal claims as they assessed how well the author identified issues, developed arguments, justified overall conclusions and wrote the answer, the reviewers' focus and feedback would be structured around these criteria, not necessarily on the claims. Any feedback on the relevant legal claims would likely be distributed across the domain-related criteria combined with other of the claims, and diluted by more general feedback on writing. When that feedback is evaluated, the domain-related back-review scale focuses authors on the reviewers' ability to give domain-related feedback, not concept-oriented feedback on the authors' treatment of the claims.

Thus, if the instructor's goal is conceptual analysis, a problem-specific rubric seems preferable to (or solely to) the domain-related one. On the other hand, a problem-specific rubric needs to be tailored to each new problem or even created anew. Domain-related rubrics are more widely applicable across problems and adaptable to other kinds of legal writing beside examination essays (*i.e.*, briefs, memoranda, articles, etc.)

B. Do-it-yourself for Legal Instructors

Any law school instructor who would like to administer a computer-supported peer-review writing exercise in his or her course can easily do so with SWoRD. Instructors can obtain instruction and sign up for accounts at https://sites.google.com/site/swordlrdc/. SWoRD leads an instructor through a process of completing simple web-based forms in order to:

- enter a new course and set up a course writing assignment, including a schedule of deadlines and grace periods for students to submit drafts (if any), final papers, reviews and backreviews;
- provide a clear description of the writing assignment and set the number of reviewers per paper;
- enter explanations of the review criteria on which student reviewers will provide written comments;
- for each review criterion, specify a 7-point numerical rating scale.
- preview the reviewing form that students will see, complete with prompts for comments and a 7-point ratings form for each criterion. SWoRD creates these automatically from the instructors' inputs.

Of all of the above, defining the review criteria and completing explanations of each of the seven points on the rating scale are the most difficult. SWoRD enables instructors to specify as many review criteria as they see fit. Instructors can supply their own criteria and ratings scales and/or select from a library of review criteria that other instructors have previously employed. As more legal instructors use SWoRD, its library of criteria and ratings scales will expand to include more entries adapted to legal courses.

Beyond filling out the initial web forms, instructors do not need to do anything else to make the review process work except remind students of the importance of exercising care in writing and reviewing and of meeting SWoRD deadlines. Students in a course also obtain SWoRD accounts and use them to submit their papers electronically. SWoRD automatically assigns papers to reviewers based on the number of reviewers the instructor selected in setting up the assignment. For each paper assigned to them, reviewers fill out the review forms on-line using their SWoRD accounts, including providing written comments and numerical ratings for each review criterion. SWoRD distributes the reviews anonymously to authors who are then prompted to provide backreviews. All of this proceeds automatically on-line. SWoRD does, however, provides instructors a number of tools to stay involved. A "View Students" option enables an instructor to monitor the progress of the review process. Instructors can use SWoRD to review the papers and distribute their reviews to authors along with the student peers' reviews. After the reviewing is complete, a View Stats option enables instructors to see a breakdown by student of the number of reviews completed, mean reviewer scores and standard deviations per criterion, and average scores. Legal instructors can even download to Excel files students' comments and ratings and perform their own experiments.

C. Improving Computer-Supported Peer Review

The results reported here also support the design of statistical models that may inform instructors about the state of a peer review exercise. Our broader ultimate aim is to provide legal instructors using computer-supported peer review systems with more timely information on how well students have understood the conceptual issues underlying a writing assignment. From the instructor's viewpoint, a class writing assignment is a black box. Until instructors actually read the first or final drafts, they do not have much information about how well the assignment has succeeded as a pedagogical activity, and even then, it is hard to get a complete picture. Ultimately, computer-supported peer review should be able to open up the black box (even more than SWoRD currently supports) enabling instructors to better understand how students are interpreting, applying, and learning from a writing exercise.

We have developed and evaluated several statistical models (*i.e.*, hierarchical Bayesian models) relating instructor scores of student essays in the IP midterm examination to peer scores based on the two peer assessment rubrics.[69] With additional refinement, parameter estimates from these statistical models could provide instructors with actionable information on individual pupils, the whole class, and the assessment rubric, even suggesting changes in curriculum or assessment. The outputs could help alert a legal instructor if the review criteria differ in difficulty, are poorly anchored, or are redundant. They could also help to inform legal instructors about the criteria's relative impact on approximating instructor scores, and whether the instructor needs to clarify the criteria and concepts for law students and revise the criteria for future peer review.[70]

In future work, we hope to develop and evaluate methods to provide legal instructors with a comprehensive overview of the progress of a class writing assignment in terms of how well students understand the issues based on structured reviewing rubrics like those used here, feedback students provide and receive in the peer review process, and machine learning /

---

[69] Ilya M. Goldin & Kevin D. Ashley, *Peering Inside Peer Review with Bayesian Models*, 6738 LECTURE NOTES IN ARTIFICIAL INTELLIGENCE, ARTIFICIAL INTELLIGENCE IN EDUCATION: 15TH INTERNATIONAL CONFERENCE, AIED 2011, AUCKLAND, NEW ZEALAND, JUNE/JULY 2011, 90 (Gautam Biswas et al., eds.).

[70] These parameters include estimates of a student's proficiency with regard to each of a rubric's criteria, the importance of the mean of inbound peer ratings (*i.e.*, ratings to an author per or across criteria) to estimating the instructor's score, and the per-criterion variance parameters pooled across all students.

computational linguistics analysis of the resulting texts. A computer-supported peer-review system like Comrade or SWoRD will present the instructor with an overview of the state of a peer review exercise. It could summarize salient information for the class as a whole, group students based on common features of their texts, and enable instructors to more effectively delve into particular students' writings in a guided manner. The instructor's overview, backed by a computational model of how well students understand the issues, will estimate a student's understanding of issues based on multiple reviewers' assessments of the author's understanding and the student's reviewer feedback to other authors. While these assessments would initially be based on *post hoc* models (*i.e.*, using supervised learning from papers the instructor already graded), we plan to investigate the extent to which, in subsequent applications of similar assignments, it is possible to model student understanding of the underlying issues even *before* the instructor grades the papers.

The enhanced peer review system will enable legal instructors to answer pedagogically important questions, such as: Did students understand the writing assignment in the manner that the instructor intended? What didn't they get? Were there some review criteria that reviewers did not apply in a consistent way? Were there some criteria for which the reviews were not informative? To what extent were the review comments based on substantive issues? On low-level writing issues? Constructive? What kinds of revisions are students making based on the reviews? Answers to these questions will help legal instructors ensure that law students obtain adequate formative feedback on essay exams and other kinds of legal writing.

Overall, computer-supported peer review is a teaching tool whose many targeted legal instructor-users could easily exchange teaching materials enabling them to build upon the best practices of their peers. If computer-supported peer review caught on among legal instructors, one could imagine enabling instructors to share assignments accompanied by criteria and rubrics, recommendations, and expertise. For example, rating criteria tailored to legal courses and problems can include, as part of feedback to instructors, information about previously obtained inter-student reliability data; commenting criteria can include information about the mean helpfulness of comments that were produced; and whole assignments can include information about mean number of paper revisions produced from one draft to the next, broken down by types of revisions. One could include comments from other legal instructors who have adopted or adapted the various writing assignments and artifacts in their teaching.

Finally, law schools need a way to respond to the persistent calls in higher education for objective evidence that law students learn. A statistically-based computational model of the kind mentioned above, one based on problem-specific, conceptually-oriented peer-review rubrics and applied within a particular law school course or in successive courses, could provide objective evidence of learning specific legal concepts and writing skills as a natural by-product of the computer-supported peer review activities. That is an interesting study in itself that depends on the work proposed here.

Appendix A

Ratings criteria and prompts

1. Domain-related rating prompts. Reviewers rated peer work on four criteria pertaining to legal writing:

Issue Identification ("issue")

1 - fails to identify any relevant IP issues; raises only irrelevant issues
3 - identifies few relevant IP issues, and does not explain them clearly; raises irrelevant issues
5 - identifies and explains most (but not all) relevant IP issues; does not raise irrelevant issues
7 - identifies and clearly explains all relevant IP issues; does not raise irrelevant issues

Argument Development ("argument")

1 - fails to develop any strong arguments for any important IP issues
3 - develops few strong, non-conclusory arguments, and neglects counterarguments
5 - for most IP issues, applies principles, doctrines, and precedents; considers counterarguments
7 - for all IP issues, applies principles, doctrines, and precedents; considers counterarguments

Justified Overall Conclusion ("conclusion")

1 - does not assess strengths and weaknesses of parties' legal positions; fails to propose or justify an overall conclusion
3 - neglects important strengths and weaknesses of parties' legal position; proposes but does not justify an overall conclusion
5 - assesses some strengths and weaknesses of the parties' legal positions; proposes an overall conclusion
7 - assesses strengths and weaknesses of parties' legal positions in detail; recommends and justifies an overall conclusion

Writing Quality ("writing")

1 - lacks a message and structure, with overwhelming grammatical problems
3 - makes some topical observations but most arguments are unsound
5 - makes mostly clear, sound arguments, but organization can be difficult to follow
7 - makes insightful, clear arguments in a well-organized manner

2. Problem-specific rating prompts. Reviewers rated peer work on five problem- specific writing criteria (the relevant legal claims), which all used the same scale:

Claims:

1. Smith v. Barry for breach of the nondisclosure/noncompetition agreement ("nda")
2. Smith v. Barry and VG for trade-secret misappropriation ("tsm")
3. Jack v. Smith for misappropriating Jack's idea for the I-phone-based instrument-controller interface ("idea1")
4. Barry v. Smith for misappropriating Barry's idea for the design of a Jimi-Hydrox-related look with flames for winning ("idea2")
5. Estate of Jimi Hydrox v. Smith for violating right-of-publicity ("rop")

Rating scale:

1 - does not identify this claim
3 - identifies claim, but neglects arguments pro/con and supporting facts; some irrelevant facts or arguments
5 - analyzes claim, some arguments pro/con and supporting facts; cites some relevant legal standards, statutes, or precedents
7 - analyzes claim, all arguments pro/con and supporting facts; cites relevant legal standards, statutes, or precedents

3. Back-review rating scale, grounded at 1, 3, 5, 7.

Q: To what extent did you understand what was wrong with your paper based on this feedback?
A: The feedback...

1 - does not substantively address my analysis,
3 - identifies some problems, but suggests no useful solutions,
5 - identifies most key problems, and suggests useful solutions,
7 - identifies all key strengths and problems, and suggests useful solutions to the problems

4. Sample questions from the two forms of the objective test. Salient differences are emphasized with *italics* and correct answers are bold; these were not emphasized in presentation to the students. [See McJohn, S. M. *Intellectual Property* (3d Ed.)08f Wolters Kluwer: Austin]

Form A (excerpt)

Paige, an academic researcher in computer science, wrote a program that learns to filter out spam emails based on the content of a user's Inbox and Deleted Items folders. He wrote a paper describing his method in detail, submitted the paper to a computer science conference, and posted the paper on his website. In the first footnote, the paper states, "*All are welcome to use this method* on condition that they pay me $39.50, per year, just a dime a day for no more spam!" This was an unusual thing for Paige to do; in academic computer science journals, it is assumed that the ideas and methods published there are free for the reader to use. Turner found Paige's paper on the website, read it, and *used the method described there to create a machine-learning spam filter for himself.*

Does Turner owe Paige the fee of $39.50 per year?

A. Yes, Turner used Paige's idea, which is both novel and complete, without shouldering the time and expense of coming up with the idea.
B. No, Paige's idea became public knowledge, and there was no confidential relationship between Turner and Paige.
C. Yes, Paige's footnote presented an offer which Turner accepted by using Paige's idea.
D. No, although there was an implied contract between Paige and Turner, it failed for lack of consideration.

Form B (excerpt)

Paige, an academic researcher in computer science, wrote a program that learns to filter out spam emails based on the content of a user's Inbox and Deleted Items folders. He wrote a paper describing his method in detail, submitted the paper to a computer science conference, and posted the paper on his website. In the first footnote, the paper states, "*For a ready-made computer program embodying this method, just click this link and you can download the program* on condition that you agree to pay me $39.50, per year, just a dime a day for no more spam!" This was an unusual thing for Paige to do; in academic computer science journals, it is assumed that the ideas and methods published there are free for the reader to use and, *in addition, it is assumed that one does not advertise products.* Turner found Paige's paper on the website, where the footnote contained the link, read the paper, *clicked the link, downloaded the program and used it as his own personal machine-learning spam filter.*

Does Turner owe Paige the fee of $39.50 per year?

A. Yes, Turner used Paige's idea, which is both novel and complete, without shouldering the time and expense of coming up with the idea.
B. No, Paige's idea became public knowledge, and there was no confidential relationship between Turner and Paige.
C. Yes, Paige's footnote presented an offer which Turner accepted by clicking and downloading Paige's program.
D. No, although there was an implied contract between Paige and Turner, it failed for lack of consideration *since the idea implemented in the program was publicly disclosed*.

Appendix B

Midterm Intellectual Property Exam Question

Problem

In January, 2004, Jack, an undergraduate in the Ames University CS Department, chatted with his Java programming course instructor, Professor Smith, about a possibility for his course project. Jack explained his idea for a new I-Phone musical application. The program would feature a "ukulele controller interface" in the form of an image on the I-Phone screen, resembling the neck of a ukulele with 4 simulated strings and frets. With the I-Phone on a flat surface, one could play the simulated ukulele by "plucking" strings with a finger of one hand and "pressing" strings onto a fret with fingers of the other hand. With a swish of a finger on the I-Phone screen, one could move up and down the neck reaching all 12 frets. In this way, one could "play" the ukulele on the I-Phone.

Thinking the task too hard, Smith discouraged Jack from pursuing this as his course project. After the semester ended, however, Smith realized that such an I-Phone-based instrument controller interface could make a great new I-Phone musical game application. As envisioned by Smith, instead of a ukulele, the image would be of the neck of a six-stringed guitar with 19 frets. An I-Phone user could play a song on the "guitar controller interface" just like on a real guitar. First, the game application would play a segment of a song and as the song progressed, colored markers, indicating which string to pluck and where to press a fret for each note, would travel up and down the screen in time with the music. Once the song segment finished, the player must "play" the notes on the instrument controller on his own in order to score points, plucking and pressing the simulated strings.

In June, 2004, Smith hired Barry, a computer science Master's degree candidate, as a part time programmer to help design a software module to generate and operate the guitar controller interface. Six months later, Smith insisted that Barry enter into, and Barry signed, a non-competition/nondisclosure agreement under which Barry agreed "to treat everything he learned while working for Smith as confidential information" and "not to work in the computer game programming field for three years after leaving Smith's employ." Barry worked on the task for a few months but encountered a technical programming problem involving synchronizing sounds, screen taps, and swishes up and down the simulated neck. One evening, while skipping stones on the campus pond, Barry remembered seeing a solution to a somewhat similar synchronization problem in a book on algorithms. Back at home, Barry adapted the book's solution to solving his current problem. It worked! A few months later, the guitar controller interface module was up-and-running in an I-Phone game application.

Barry graduated, left Smith's employ, and moved away, but he could not forget Smiths's idea for a dynamite I-Phone musical video game application. Barry proceeded to create his own I-Phone musical video game application. Since he had already solved the synchronization problem once, it was pretty easy even though Barry had not kept a copy of the guitar controller interface computer code he developed for Smith. Barry did have to modify the approach to deal with the faster game play he envisioned. When a player wins Barry's game, the guitar neck image spins wildly. The image looks like the neck of a Leghorn L6-s guitar like the one rock idol, Eddie Spindrift used to spin around after a set. In fact, Barry's game looks so promising that this month, VeeGames, Inc. (VG) plans to acquire all of Barry's rights for the high six figures (!) and to market the game under the name Guitar-Gyro.

Meanwhile, last month, Smith began marketing his I-Phone musical video game application under the name Guitar-Pyro. Before graduating, Barry had suggested that Smith design the game application to simulate the neck of a Giblet SG guitar, just like the one Jimi Hydrox, the famous rock singer used to play before he died. Smith did just that. When a player wins, Guitar-Pyro's guitar controller image bursts into simulated flames just like Jimi's used to do. Within a month of Guitar-Pyro's debut, musically-inclined kids in Ames City and elsewhere were tweeting their friends urging them to try out the Guitar-Pyro I-Phone app with its wicked guitar controller interface and simulated flames.

As an associate working for the law firm representing Smith's interests, you have been asked to provide advice concerning Smith's rights given the above developments. Your boss tells you to assume that she will research the extent to which Guitar-Pyro (or Guitar-Gyro) is patentable or subject to federal copyright, and she has asked you to focus on any other issues. (Ignore all problems presented by any real-world products (*e.g.*, Guitar-Hero, Gibson guitars) that the video gamers/musicians among you may recognize as similar to the above. Also, ignore any I-Phone licensing issues.)

Appendix C

Instructor's Answer Key (Excerpts)

Claim of breach of non-disclosure / non-competition agreements by Smith v. Barry (nda):

- Address the nondisclosure/noncompetition agreement's extreme breadth (3 years, no work in computer games), especially in that a student is expected to work in the area of his/her studies.
- The agreement would be unenforceable since Barry received no additional value in exchange for his agreeing six months after he started work and unless trade secrets were at stake.

Claim of trade-secret misappropriation by Smith v. Barry and VG (tsm):

- VG has not used Smith's trade secrets (as required in Metallurgical Industries v. Fourtek case.) Barry may not have used the trade secret either; he modified it to deal with faster game play.
- The guitar-related game/name/images linked to a historical rocker and the controller interface have been disclosed publicly, so any trade secret protection is at best short-lived.
- Interface synchro method and code are technical info; some factors favor protection: they have value, Smith protected the info, his employee had to learn about them to do his job. Barry has a right to use the method in subsequent employment as his general skills and knowledge. He relied on his own memory of a solution learned in class, and did not take away "specific knowledge" such as concrete code.
- Barry developed the solution to the synchro problem on his own, or adapted a book's method. Since there is no contract allocating rights to Smith, Barry owns the info. If Barry is an "R and D" employee, Smith owns what Barry developed, but Barry is merely a programmer. If Barry used Smith's equipment, Smith would have "shop rights". Smith only gets a nonexclusive right to use the method; Barry may sell to VG.

Claim of idea misappropriation by Jack v. Smith (idea1):

- Re Jack's claim for misappropriating his idea for the interface, there is no express or implied contract in this student/teacher course setting and no expectation or custom against exploiting a student's idea.

- With respect to a quasi-contract or a property theory, Jack's idea is novel and somewhat concrete, but perhaps not sufficiently concrete for the latter. There seems to be no unjust enrichment; it may be unfair to use a student's idea without compensation, but the student did not invest in development.
- Smith did not use the idea; he implemented a controller interface for a guitar, not a ukulele.

Claim of idea misappropriation by Barry v. Smith (idea2):

- Re claim re Barry's idea for the Jimi-Hydrox-related flames design, unless there is a basis in custom, Smith doesn't have to compensate his employee who volunteered the idea. Any expectation of confidentiality from the employment agreement and relationship runs in favor of the employer.
- On the other hand, Barry was a computer programmer; this idea did not have to do with his expertise or why he was hired, so, arguably, it is not covered by his employment duties.

Claim of right of publicity infringement by Estate of Jimi Hydrox v. Smith (rop):

- Assuming the right-of-publicity claim is descendible, Hydrox would have the right to recover for commercial exploitation of his celebrity identity.
- Smith does not use Jimi Hydrox's name, but Hydrox may have become so identified with the Giblet Guitar and flaming guitar act that consumers would associate him with Guitar-Pyro, perhaps assuming that the estate endorsed Guitar-Pyro (see Here's Johnny Portable Toilets, White v. Samsung.)
- Other issues are whether the exemption for entertainment/information products applies to an i-Phone application and whether there would be strong first amendment protection for Smith, since the game was not transformative of Hydrox's act, but it also did not appropriate the whole act (Zacchini).