Eliciting formative assessment in peer review

Ilya M. Goldin & Kevin D. Ashley

Carnegie Mellon University & University of Pittsburgh, PA | United States

Abstract: Computer-supported peer review systems can support reviewers and authors in many different ways, including through the use of different kinds of reviewing criteria. It has become an increasingly important empirical question to determine whether reviewers are sensitive to different criteria and whether some kinds of criteria are more effective than others. In this work, we compared the differential effects of two types of rating prompts, each focused on a different set of criteria for evaluating writing: prompts that focus on domain-relevant aspects of writing composition versus prompts that focus on issues directly pertaining to the assigned problem and to the substantive issues under analysis. We found evidence that reviewers are sensitive to the differences between the two types of prompts, that reviewers distinguish among problem-specific issues but not among domain-writing ones; that both types of ratings correlate with instructor scores; and that problem-specific ratings are more likely to be helpful and informative to peer authors in that they are less redundant.

Keywords: computer-supported peer-review, rubrics, peer assessment validity

journal of
WRITING RESEARCH

Goldin, I.M., & Ashley, K.D. (2012). Eliciting formative assessment in peer review. *Journal of Writing Research, 4(2),* 203-237. http://dx.doi.org/10.17239/jowr-2012.04.02.5

Contact: Ilya Goldin, Carnegie Mellon University, Human-Computer Interaction Institute, 5000 Forbes Avenue

Pittsburgh, PA 15213, PA | USA - goldin@cmu.edu

1. Computer-supported peer review

It is useful to offer formative feedback and assessment to students who are working on open-ended problems (Rittel & Webber, 1973; Voss & Post, 1988), such as problems that involve conflicting objectives or reasonable alternative solutions. Courses in which open-ended problems are addressed make natural candidates for Writing in the Disciplines / Writing Across the Curriculum instructional programs (Bazerman et al., 2005). These courses target both content knowledge and writing skills.

Assessment of such student works is not straightforward. First, the works often involve free-form text because they require arguments and justifications for one solution over others. Second, because students may frame the open-ended problems according to their knowledge, beliefs and attitudes, it is difficult to create a gold standard and evaluate student works against such a standard. Third, assessment of writing is very demanding of instructor time.

One technique that may enable formative assessment of student essays that address open-ended problems is computer-supported peer review. Using student reviewers allows assessment of free-form essays, including essays that may vary in how they frame an open-ended problem. Thanks to computer automation of administrative aspects of peer review, the burden on the instructor can be decreased. Peer review has been shown to provide accurate summative assessment of student work (Draaijer & van Boxel, 2006). Under the right circumstances, feedback can have a positive impact on learning outcomes with effect sizes as high as 1.10 (Hattie & Timperley, 2007), and peer feedback can be similarly effective to instructor feedback (Rijlaarsdam, 1987).

A peer review process depends on the rubric that students use to assess each other's work and to give feedback to each other. Given the variety of possible rubrics, and given that rubrics influence the experience of both reviewers and authors, it is important to determine whether some rubrics are more valuable than others. However, despite decades of research on writing assessment and recent activity in peer review scholarship (Goldin, Brusilovsky, Schunn, Ashley, & Hsiao, 2010; e.g., Strijbos & Sluijsmans, 2010), it is still unclear how a rubric should be structured to produce feedback that is valid, reliable, differentiated and helpful. This paper is an early look into how rubrics can be structured and the impact of alternative structures. The paper proposes some new conceptual tools for thinking about rubrics, and develops two new rubrics for legal writing on the basis of these concepts. It then evaluates the new rubrics in a real-world peer review exercise.

1.1 Assessment of writing

Assessment may be used for summative or formative purposes, the latter being the focus of this research. Formative assessment may be defined on the basis of two key components: a student's work needs to be evaluated with respect to criteria, and this evaluation must be made useful to the student either directly or indirectly (e.g., via an instructor). As summarized by Cizek (2010), the theory of formative assessment was originally developed by Scriven in the context of program evaluation, and Bloom and colleagues applied it in the context of student learning and distinguished it from summative assessment (Bloom, Hastings, & Madaus, 1971; Scriven, 1966). More recently, as Cizek (2010) reports, formative assessment has been described as "a tool for helping to guide student learning as well as to provide information that teachers can use to improve their own instructional practice"(Shepard, 2006). This view of formative assessment is more expansive than an alternative definition according to which only feedback that leads to an improvement in learning outcomes may be termed formative (Shute, 2008). By contrast, summative assessment involves grading or ranking a student's level of achievement. Formative assessments give feedback to help students learn, not (just) to grade or rank them.

According to assessment theory, an instructor wishing to assess student work must often make "qualitative judgments", characterized as follows:

> 1) Multiple criteria are used in appraising the quality of performances. 2) At least some of the criteria used in appraisal are fuzzy ... A fuzzy criterion is an abstract mental construct denoted by a linguistic term which has no absolute and unambiguous meaning independent of its context. 3) Of the large pool of potential criteria that could legitimately be brought to bear for a class of assessments, only a relatively small subset are typically used at any one time. 4) In assessing the quality of a student's response, there is often no independent method of confirming, at the time when a judgment is made, whether the [judgment] is correct. 5) If numbers (or marks, or scores) are used, they are assigned after the judgment has been made, not the reverse. ...
>
> It is also useful to make a distinction among end products according to the degree of design expected. (...) [In fields such as writing] design itself is an integral component of the learning task.... Wherever the design aspect is present, qualitative judgments are necessary and quite divergent student responses could, in principle and without compromise, be judged to be of equivalent quality (Sadler, 1989, pp. 124-126).

In other words, the assessment of writing is in itself an open-ended problem, which is separate from the open-ended problem that the student is analyzing.

Instructors may assess written works for many reasons, e.g., to measure students' analytical skills, knowledge, and understanding (Stiggins, 2005), which are especially relevant to writing in the content disciplines. For formative peer assessment of writing, relevant techniques include holistic scoring, primary trait scoring, and analytic scoring (O'Neill, Moore, & Huot, 2009). The scholarship on these techniques is vast, and sometimes uses conflicting language; cf. the definitions of holistic scoring according to Cooper (1977) and Wolcott & Legg (1998). The following brief summary does not aim to be a definitive statement and merely describes how these terms are used in this document.

Holistic scoring evaluates an entire essay at once; its motivation is that an essay is more than the sum of its "atomistic" (Lloyd-Jones, 1977) parts. According to O'Neill et al. (2009), its roots were at the Educational Testing Service (Godshalk, Swineford, & Coffman, 1966), which needed to develop a methodology for rapid, reliable, summative assessment of writing samples. Although a rater may choose to provide formative feedback after arriving at a holistic impression, the scoring does not facilitate this directly. In practice, holistic scoring is used normatively, i.e., to rank a written work relative to other works, rather than to evaluate the work against some fixed standard irregardless of the quality of other works.

Primary trait scoring, a counter-point to holistic scoring (Lloyd-Jones, 1977), proposes that different rhetorical modes—"explanatory, persuasive, and expressive"—deserve distinct approaches to scoring. Given a mode and a writing assignment, an assessment administrator ought to create a scoring guide that is focused on some particular primary trait, e.g., "imaginative expression of feeling through inventive elaboration of a point of view" (Lloyd-Jones, 1977). The assessment of the essay is to be based solely on the primary trait, and not other elements of writing (Wolcott & Legg, 1998). Primary trait scoring is problem-specific: "A wide open subject, such as that allowed in conventional holistic scoring, permits each writer to find a personally satisfying way to respond, but in Primary Trait Scoring a stimulus must generate writing which is situation-bound" (Lloyd-Jones, 1977). Because of the focus on a single trait and on one assignment, primary trait scoring can be used to provide detailed formative feedback to students.

The analytic scoring method rejects the "essay as a whole" holistic approach as not providing sufficiently justified judgments of writing quality, and aims to evaluate written works based on multiple well-articulated elements of writing. Where holistic scoring provides a single score that sums up all the qualities of an essay, analytic scoring provides one score for each element of interest. An early factor analysis of the comments of independent readers of a 300 essay corpus determined the following elements: ideas, form, flavor, mechanics, and wording (Diederich, French, & Carlton, 1961). As Wolcott & Legg (1998) point out, although large-scale standardized assessments require consistency in scoring across essays, in classroom use, instructors may adapt the scoring guide to the assignment at hand and supplement the score for each element with individualized feedback.

Assessment instruments are traditionally evaluated in terms of validity, which examines whether the instrument really measures what it is purported to measure, and reliability, which looks at whether the instrument produces a consistent result, e.g., when used by different assessors or on different occasions. The relative importance of validity and reliability to educators has not remained constant over time (Huot, 1990; O'Neill et al., 2009; Yancey, 1999), due in part to the inherent tension between these concepts: "The greater the reliability of an assessment procedure, the less interesting a description it provides of writing" (Williamson, 1994). Or, more bluntly: "The concepts of theoretical interest (in psychology and education) tend to lack empirical meaning, whereas the corresponding concepts with precise empirical meaning often lack theoretical importance" (Torgerson, Theory, & Methods, 1958), cited in (Williamson, 1994).[1]

For example, an oft-cited modern analytic scoring rubric distinguishes among Six Traits of writing: ideas/content, organization, voice, word choice, sentence fluency, and conventions (Spandel & Stiggins, 1996). One evaluation of the Six Trait rubric found that it has high inter-trait (inter-dimension) correlation, which suggests that its dimensions are not measuring distinct aspects of writing, and that it suffers from low test-retest reliability (Gansle, VanDerHeyden, Noell, Resetar, & Williams, 2006). That study also found that an alternative, Curriculum-Based Measures, has less inter-dimension correlation and higher reliability. Curriculum-Based Measures assesses writing aspects such as the number of correctly spelled words, and the number of correctly capitalized words (Jewell & Malecki, 2005), but such measures seem likely to show a ceiling effect at advanced levels of writing, such as one would expect from college students. Further, it is unclear that an instructor could act on the information provided by Curriculum-Based Measures, or that Curriculum-Based Measures could shed light on the quality of a student's solution of an open-ended problem.

As Deane & Quinlan (2010) note, multiple studies have found that distinct traits of essay quality are highly correlated (Lee, Gentile, & Kantor, 2008; McNamara, 1990), and such inter-trait correlation may motivate the use of holistic scoring over analytic scoring.

To put these analytic and holistic rubrics in context, consider essay assessments that are typical of American law school examinations. These are usually fictional but plausible fact situations, carefully designed by the instructor to raise various legal issues that need to be analyzed in terms of strengths and weaknesses for the parties to the case. Such essay assessments approximate assessment of authentic performance in that practicing lawyers do need to explain what legal issues arise in a novel legal 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. Each legal claim involves different legal interests and requirements and presents a somewhat different framework for viewing the problem. Each claim is defined in terms of elements; if the student finds a basis for each element in the problems facts, the legal claim arguably applies and would support a party's demand in a lawsuit for compensation or other relief. The assertions that elements are satisfied are often equivocal, requiring students to consider arguments and counterarguments. In addition, various defenses may apply in which case the claim would fail even though the elements were satisfied.

In general law school courses, those that focus on teaching a particular area of law, such essay assessments focusing on fictional scenarios are used instead of performance assessments (Stiggins, 2005). Since such courses are often quite large, performance assessment is usually limited to clinical courses involving small numbers of students in representing actual clients under the close supervision of instructors. Fictional scenarios are preferred because real-world cases may not present a pedagogically ideal collection of legal issues and factual circumstances. Students are encouraged, however, to cite relevant real-world cases from the course casebook to support their analyses.

It is not clear that traditional writing-oriented rubrics can adequately assess such law school essays. Some aspects of writing assessment (e.g., whether the essay expresses the writer's individual voice) are, at best, irrelevant. Aspects of writing assessment that do seem relevant are those that deal with clarity and rigor of argumentation, but a focus on argumentation in general may miss nuances of legal argumentation.

Thus, despite a variety of theoretical and practical approaches to writing assessment, open questions remain. In particular, it is unclear how assessment techniques such as those above may transfer from writing courses to more domain- or discipline-specific contexts as in Writing Across the Curriculum. As will be seen below, the study reported here is rooted in a novel classification of assessment rubrics in terms of generality (applying to a specific problem vs. application to writing in the domain) and orientation (oriented toward concepts vs. toward process). This classification explains some aspects of rubric design for writing courses in the disciplines.

1.2 Student-involved Assessment

Possible sources of assessment include not only the instructor, but also the students themselves. Student-involved assessment for learning may include peer assessment and self-assessment, and both can be administered for summative or formative reasons.

Students and instructors may hold different views of peer assessment. One cause of this may be that "the instructor has access to grades for all papers, whereas the students only see grades on their own papers (and perhaps one or two more by social comparisons with friends)" (Kwangsu Cho, Schunn, & Wilson, 2006).

More broadly, students may not share the instructor's expectation that the purpose of assessment is formative. The different impacts of summative and formative assessment on students are vividly illustrated by two recent studies. In an experiment in which students engaged in formative self-assessment, researchers report that the students "had positive attitudes toward self-assessment after extended practice; felt they can effectively self-assess when they know their teachers expectations; claimed to use self-assessment to check their work and guide revision; and believed the benefits of self-assessment include improvements in grades, quality of work, motivation and learning" (H. Andrade & Du, 2007). By contrast, in another study where students engaged in summative self-assessment and peer assessment, they "felt it impossible to be objective when considering their own work. In peer-assessment, the students found it difficult to be critical when assessing the essay of a peer. The students found it easier to assess technical aspects of the essays when compared to aspects related to content"

(Lindblom-Ylänne, Pihlajamaki, & Kotkas, 2006). Students tend to be skeptical of summative peer assessment even if it is accurate (Draaijer & van Boxel, 2006).

These findings are consistent with empirical research on feedback. As Shute (2008) summarizes, "features of feedback that tend to impede learning include: providing grades or overall scores indicating the students standing relative to peers, and coupling such normative feedback with low levels of specificity (i.e., vagueness)."

These findings are also consistent with the theory of formative assessment. Students need to understand what performance they should aim for, their level of current performance, and how to improve their performance (Hattie & Timperley, 2007; D.Royce Sadler, 1983). Criteria-referenced peer assessment is one way that students may receive information on all three of these elements.

Peer assessment may also help students develop meta-cognitive skills, such as self-monitoring: "The most readily available material for students to work on for evaluative and remedial experience is that of fellow students. ... [Peer review is important because] it is clear that to build explicit provision for evaluative experience into an instructional system enables learners to develop self-assessment skills and gap-closing strategies simultaneously, and therefore to move towards self-monitoring" (D.R. Sadler, 1989). Improvement of self-monitoring skills through per review correlates with improvement in writing quality of second drafts (Kwangsu Cho, Cho, & Hacker, 2010).

In writing in particular, theories of writing and revision (Alamargot & Chanquoy, 2001) note that authors need to be able to understand the quality of their works in progress and how to improve them. For instance, as motivation for a theory of revision, Hayes, Flower, Schriver, Stratman, & Carey (1987) note that "writers have difficulty detecting faults in their own text", and that "finding and fixing problems appear to be separate skills". The 1996 Hayes model of writing considers that the task environment has a social component that involves the audience and collaborators (Hayes, 1996). At the same time, expert-novice studies show that novice writers perform poorly at detecting and diagnosing composition defects, which suggests that peer feedback on composition may similarly fail to detect and diagnose defects (Hayes et al., 1987). Formative peer feedback is as effective as instructor feedback in terms of (1) post-feedback scores on writing achievement variables, (2) levels of writing apprehension, (3) quantity of planning activities such as self-instruction and generation of ideas, (4) quantity of interruptions of the writing process by a choice process, and (5) quantity of evaluation and revision in the revision process (Rijlaarsdam, 1987).

Instructors both generate and make use of assessment judgments. "Broadly speaking, feedback provides for two main audiences, the teacher and the student. Teachers use feedback to make programmatic decisions with respect to readiness, diagnosis and remediation. Students use it to monitor the strengths and weaknesses of their performances, so that aspects associated with success or high quality can be recognized and reinforced, and unsatisfactory aspects modified or improved" (D.R. Sadler, 1989). Instructors need to make decisions that affect individual assessees (e.g., whether to recommend additional exercises) as well as groups of students (e.g., whether to re-teach subject matter that is challenging to some students).

Thus, while students may be invited to engage in summative peer assessment, they may mistrust it, and it will not help them develop self-monitoring ability. On the contrary, formative peer assessment is motivated theoretically and empirically.

1.3 Computer-supported Peer Review in Education

Since the early days of computer-supported peer review in education (Neuwirth, Chandhok, Charney, Wojahn, & Kim, 1994; Patterson, 1996; Scardamalia & Bereiter, 1994; Zeller, 2000), the research on peer review systems, their use cases, their interfaces, and implications for the learning sciences has increased dramatically (Goldin, Ashley, & Schunn, 2012). Peer review is used in instruction in many and perhaps all academic disciplines, with especially significant research communities in writing, computer science, nursing, and second language education. Whether or not it is aimed at convergence with instructor assessment, or as an independent perspective that may be valid and informative on its own, peer review can facilitate opportunities for formative feedback beyond what instructors alone can provide. For example, it can enable instructors to assign exercises in which students work on multiple drafts of analyzing and writing about an open-ended problem. When student authors receive feedback from peer reviewers, they practice skills such as seeing their own work from other perspectives, revising with the reader in mind, responding to criticism, and integrating information from multiple sources. In addition, the task of reviewing others' work lets students practice cognitive skills including evaluation and critiquing, as well as social skills such as framing their feedback so that it is useful to the author.

Issues of validity and reliability are central to peer review insofar as peer review is an assessment technique. However, these issues take on different forms in different peer review settings; we cannot determine once and for all whether or not peer review valid or reliable. For example, 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" (Miller, 2003). An instructor's view of the validity of an exercise may differ from an individual student's view, because an instructor evaluates validity based on how well peer assessment converges to the instructor's own aggregating across all student works, but an individual student author evaluates validity based on whether the peer ratings that only he or she receives happen to deviate from the instructor's grade (Kwangsu Cho et al., 2006). One alternative to holding instructor assessment as the standard is to define validity as convergence of peer assessment to self-assessment (Miller, 2003). Reliability of peer assessment is most easily addressed by inviting multiple peer assessors, e.g., because the aggregated summative assessment of several reviewers is more reliable than a single reviewer's assessment (Kwangsu Cho & Schunn, 2007).

Elicitation of assessment is guided by the user interface that mediates the relationship between assessor and the object being assessed, i.e., the peer reviewer and the work of the peer author. Even a completely free-form assessment interface accompanied by an instruction to "tell the author what you think" mediates the relationship; in this example, the interface leaves it to the reviewers to choose and define their own assessment criteria. Criteria are abstract ideals to which students (ought to) aspire, and against which one hopes to assess student performance[2]. A rubric (H. G. Andrade, 2000) is an operational definition of the criteria of interest. Each of a rubric's dimensions defines a single criterion, and each dimension spans a range of performance levels, e.g., from poor to proficient.

One aspect of peer review that has become standard is to elicit peer assessments by the use of prompts. In psychological research, prompts have been used to elicit specific kinds of information and to stimulate particular cognitive or metacognitive activity. For example, prompts have been used to elicit self-explanations (Chi, de Leeuw, Chiu, & LaVancher, 1994), to encourage monitoring and reflection in individual writing (Hübner, Nückles, & Renkl, 2006) and in online conversation (Baker & Lund, 1997), to develop arguments (Bereiter & Scardamalia, 1987), and to stimulate explanation and elaboration between peers (King, 1997). Prompts fit well with rubrics, because they allow collection of feedback in many forms, including numeric ratings and written comments. For example, ratings on a grounded scale can indicate the current performance levels. However, ratings are likely to function normatively, and normative feedback that is unexplained and unelaborated may impede learning (Shute, 2008). Rubrics may contain checklists of typical errors (Sanders & Thomas, 2007), but such rubrics may not fit with assessment of open-ended problems. Thus, prompts that request ratings should also request explanatory comments (Wooley, Was, Schunn, & Dalton, 2008). In addition to explaining a rating, comments can suggest ways for the assessee to improve performance. The resulting peer feedback should allow students to monitor and self-regulate their writing (Hacker, Keener, & Kircher, 2009).

1.4 New Concepts for Rubrics

Rubrics may be used within peer review to support assessment, but few studies examine rubrics *per se*. As the literature review in a recent dissertation notes, "while there seems to be a general consensus that rubrics are important and that they improve the peer review activity, there is not as much agreement on how they should be implemented" (Turner, 2009). The choice of rubric influences the experience of both reviewers and authors. It is desirable to define a rubric that stimulates reviewers to produce formative feedback and accurate assessment, and that reflects the true range of performance in student work, and to avoid a rubric in which some dimensions are redundant or uninformative. More subtly, presenting a rubric to the students is a teaching act in itself, because it communicates what assessment criteria the instructor considers to be important, what constitutes high and low quality performance in terms of normative standards, and how an expert may assess work in this domain.

Insofar as rubrics are instruments that can be employed for various purposes, one might wish to have a way of choosing the right instrument, and of creating new instruments where old ones do not suffice. We have arrived at a set of concepts that help in that regard. These concepts are intuitive notions rather than formal mathematical definitions, because the goal of this work is to provide ways for thinking about and comparing rubrics rather than to enable rubric generation, verification, or other operations that demand formal precision. (By way of self-reflection, the authors acknowledge that their view of the world, and the terminology introduced below, is influenced by the authors' training in computer science. Aside from how a computer-scientific perspective may be problematized, the authors are hopeful that the terminology is clear and useful to audiences in the humanities and social sciences.) These concepts are *support*, *generality*, *porting*, and *orientation*.

The *support* of a rubric is the set of the exercises to which the rubric can be applied. One can speak about whether a particular exercise falls within or outside of the support of a rubric. Without defining support formally, we can say that the support of a rubric that applies only to essays about *Hamlet* is smaller than the support of a rubric that applies to essays about *Hamlet* and *Macbeth*. The latter rubric is more *general*, i.e., *generality* is a rubric property that is a function of the size of the support: the larger the support, the more general the rubric.

Rubric generality is a continuum, and key points along that continuum are domain-independent, domain-relevant, and problem-specific rubrics. This distinction is particularly salient for Writing in the Disciplines / Writing Across the Curriculum courses in that general writing instruction does not address issues of domains and domain problems.

A domain-independent rubric provides operational definitions of criteria such that the rubric could apply to any domain. For instance, early versions of the SWoRD system (Kwangsu Cho & Schunn, 2007) suggested insight, logic and flow as default criteria to assess writing in any discipline.

By comparison, a domain-relevant rubric contextualizes general criteria within a domain, and is less general (i.e., necessarily has smaller support) than a domain-independent rubric. For example, the general assessment criterion of logic pertains to whether or not the paper presents a well-reasoned argument, backed by evidence. This may be operationalized within engineering ethics case analysis with reference to domain-relevant argument structures such as general ethical issues and unknown morally relevant facts (Harris, Pritchard, & Rabins, 2000). Domain-relevant criteria need not be specific to argumentation as a rhetorical mode. The key distinction from domain-independent rubrics is that domain-relevant rubrics are grounded in the ideas and terminology of the domain.

A problem-specific rubric is least general in that it incorporates elements of the problem explicitly. For instance, students in a zoology course were asked to produce a summary of a research paper and to assess each other's summaries. One prompt in that rubric was Does the summary state that the study subject was the great tit (Parus major) or the Wytham population of birds? AND does the summary further state that the sample size was 1,104 (egg) clutches, 654 female moms, or 863 identified clutches? (Walvoord, Hoefnagels, Gaffin, Chumchal, & Long, 2008)

Dimensions of different generality are sometimes combined in a single rubric for one assignment. The same zoology rubric also contained a domain-independent prompt, "How would you rate this text?"

One study found that rubric generality interacted with the students' aptitude for following directions (Lin, Liu, & Yuan, 2001). Specifically, students in a writing exercise in a computer science class used either a holistic rubric, where they "gave a total score and offered a general feedback for an entire assignment," or a domain-independent writing rubric. The domain-independent rubric included the following dimensions: "(1) relevance of the project to the course contents (2) thoroughness of the assignment (3) sufficiency of the references (4) perspective or theoretical clarity (5) clarity of discussion, and (6) significance of the conclusion." The type of rubric was found to have no effect on the quality of feedback as rated by an expert. Students who were less inclined to follow directions benefited from domain-independent reviewer support and were hurt by holistic reviewer support (in terms of higher second draft quality as assessed by peers); on the contrary, students who were more inclined to follow directions benefited from holistic reviewer support.

Validity and reliability of peer review depends critically on the rubric. Peer assessment of oral presentations converged to self-assessment when peer reviewers used a rubric composed of twenty five domain-relevant criteria that were distinct and domain-relevant, but not when they used a rubric of six domain-independent traits (Miller, 2003). Peer assessment via a holistic rubric converged to tutor assessment, but assessment via 16 domain-relevant criteria did not (Chalk & Adeboye, 2005). The correlation of summative tutor and peer assessments, although statistically significant, was low, r=0.27,df=62, p<0.05. (The latter study is complicated in that the assessment instrument with specific criteria lacked free-form commenting, while the holistic instrument required it.)

Porting is the act of changing a rubric's support. Suppose that a rubric applies to essays written on *Hamlet* and an instructor wishes to use a similar rubric for essays on *Macbeth*. To port the rubric from *Hamlet* essays to *Macbeth* essays, the instructor can make a new rubric in which all references to *Hamlet* need to be supplanted by references to *Macbeth*. The new rubric is as general as the original because each rubric applies to just one kind of essay. Alternatively, the instructor can make a rubric that applies to both *Hamlet* and *Macbeth* by restating the rubric to generalize across the references to the particulars of either *Hamlet* or *Macbeth*. This new rubric is more general than the original because it applies to essays on both plays.

Finally, the *orientation* of a rubric is the broader theme that ties together rubric dimensions. For instance, in one early study, a web-based Group Support System was developed to support articulation and application of criteria by instructor and students (Kwok & Ma, 1999). The system was used by students for the duration of a semester (13 weeks) in a course where students worked as a group on a large information systems project. Peer assessment was performed within groups of 20 students, and each group devised its own rubric. The rubrics were *oriented* towards aspects of both process (e.g., collaboration within the group) and product (e.g., software reliability). The system allowed students to assess themselves and each other with respect to the rubric. By comparison with students who engaged in similar activities face-to-face, the Group Support System led to two outcomes with small but statistically significant differences: student final projects were of higher quality and students focused more on deep features of the domain. As the paper notes, it is unclear whether these differences are due to the process support that was provided by the software or to the rubrics that the students devised.

Orientation can affect how a rubric fits analysis of open-ended problems. An analysis of an open-ended problem often needs to be an argument. This may be addressed via criteria focused on the mechanics of argument *per se*. For example, a domain-independent argument-oriented rubric could be based on rhetorical elements such as claims, warrants and evidence (Toulmin, 2003), and a domain-relevant argument-oriented rubric could contextualize these rhetorical elements in argumentation skills often practiced in the domain (e.g., citing precedents). Alternatively, a rubric may focus on the content of the argument. A key structure in analysis of an open-ended problem is the set of conceptual issues that tie together relevant facts and that facilitate evaluation of alternative solutions.

For example, criteria for evaluating solutions to a computer programming assignment may focus on concepts of object-oriented software design such as abstraction, decomposition and encapsulation (Turner, 2009); such a rubric is said to be concept-oriented. A rubric investigated in an introductory computer science course contained these three conceptually oriented dimensions (abstraction, decomposition, and encapsulation); two additional dimensions were functionality and style, which are general notions of computer programming. One finding was that student reviewers who used this rubric to assess expert-created examples significantly improved in their understanding of decomposition, which was demonstrated well in these examples. Learning was measured by having students create concept maps of abstraction, decomposition, and encapsulation before, during and after the intervention, which took place over ten weeks and four programming assignments. In a different condition, students who provided formative feedback to their peers (rather than assessing expert-created examples) showed an improvement in understanding of decomposition during the intervention, but not on the posttest.

Optimizing for rubric generality versus for fit to open-ended problems is a trade-off. On the one hand, some assessment experts recommend that instructors devise generally applicable rubrics. One guide to teachers on formative assessment states that a rubric that is practical "is of general value; that is, it is not task specific; it can be used to evaluate performance in response to a number of different tasks" (Stiggins, 2005, p. 160). Domain-relevant rubrics can be reused across many open-ended problems in a domain, and domain-independent rubrics can be reused even more broadly than domain-relevant ones. Developing separate problem-specific rubrics for each problem may be a burden on the instructor.

On the other hand, it is much easier to structure a problem-specific rubric in terms of concept-oriented criteria than a domain-relevant rubric. This is because given a single problem, even an open-ended one, it may be possible to arrive at a short list of conceptual issues that could or should be addressed in a student's analysis. Enumerating all concepts that could be relevant to a domain is an enormous undertaking, and a rubric that does so would have too many dimensions to be usable in practice. By making explicit the deep features of a problem, a concept-oriented rubric focuses reviewer attention on what the author had to analyze, and provides a context for the analytical and writing activities. Notably, the levels of performance measured by a concept-oriented criterion may still focus on logical rigor and written expression of argument so that these valuable aspects of domain-independent and domain-relevant criteria are retained.

As far as known, rubrics of any generality and with an orientation to either domain-relevant argumentation skills or to problem-specific concepts can elicit feedback regarding the key feedback questions Where am I going? How am I going? and Where to next? (Hattie & Timperley, 2007) That is, the feedback must help assessees understand what performance they should aim for, their level of current performance, and how to improve their performance. Given the growth in popularity of and potential impact of peer assessment for analysis of open-ended problems, it is important to investigate the trade-off of rubric generality and fit to open-ended problems.

The work described here compared the effects of two analytic rubrics for peer assessment of writing: rubrics that focus on domain-relevant aspects of writing composition versus rubrics that are specific to aspects of the assigned problem and to the substantive conceptual issues under analysis. The fact that a domain-relevant rubric can be used broadly means that it is more likely to be validated. Because evaluating a rubric can be challenging and time-consuming, instructors would benefit if they could reuse rubrics validated by third-party instructors and researchers. However, when rubrics are used for formative assessment, the primary outcomes are whether the feedback helps students improve their performance and whether the assessment is accurate.

1.5 Hypotheses

Specifically, this study addressed the following research questions considering the effects of supporting reviewers with problem-specific and domain-relevant rubrics: Is peer assessment valid and reliable? Are peer reviewers responsive to the analytic design of the two rubrics, or do they treat the rubrics holistically? Finally, do authors consider feedback elicited via these rubrics to be helpful?

Peer assessment validity. Both types of rubrics were expected to encourage peer reviewers to produce valid feedback on written works. Operationally, rubric validity was defined as the validity of peer ratings elicited by the rubric. Validity was measured as correlation between aggregated inbound peer ratings and summative instructor scores of the written works. When students are meant to learn writing in the domain as well specific subject-matter ideas, instructors need to evaluate student work according to both sets of criteria. Thus, when peers evaluate each other's work with respect to problem-specific and domain-relevant rubrics, they explore two important but distinct aspects of the class material, and both ought to correspond to summative instructor scores. Additionally, given the novelty of the problem-specific rubric, peer ratings of the papers from the problem-specific condition were validated at the level of separate dimensions by correlating them against the ratings of a trained rater.

Peer assessment reliability. The problem-specific rubric was expected to elicit more reliable peer ratings than the domain-relevant rubric, because problem-specific criteria may be easier to apply objectively than domain-relevant criteria. If an essay is missing a key concept, reviewers are likely to agree. By comparison, domain-relevant criteria may be more subjective. For instance, reviewers may disagree in terms of what constitutes good issue identification or good document organization even if they are supported with a rubric.

Reviewer responsiveness to rubric. Peer reviewers were expected to be responsive to the rubrics before them, i.e., to give their ratings according to the dimensions of the rubrics and not holistically. Reviewer responsiveness to analytic rubrics is not a foregone conclusion. A rubric may be constructed in such a way that different criteria evaluate the output of the same underlying cause (Diederich et al., 1961; Gansle et al., 2006). Even if the criteria address what can hypothetically be different skills (e.g., argumentation vs. issue identification), students may acquire these skills together, and the skills may also manifest themselves together. Another consideration is that students may not differentiate among criteria (i.e., even if there are substantive distinctions, they may be too subtle) or they may interpret the criteria not in the way that the instructor intended. Reviewer responsiveness was evaluated by asking if ratings received by authors are linked across rubric dimensions, or if the dimensions are independent.

Furthermore, given the novelty of the problem-specific rubric, its conceptual distinctions were validated by comparing the student peer ratings of written works against the ratings of a trained rater.

Feedback helpfulness. Even if feedback is valid and reliable, and even if reviewers pay attention to the dimensions of a rubric, the feedback they produce may not be formative. This is difficult to define operationally; for instance, as noted above, researchers do not agree on what is formative, let alone peer reviewers. Nonetheless, student authors can be asked directly whether or not feedback was helpful to them, i.e., to give a back-review of the feedback they received. It was expected that peer reviewers would produce helpful feedback according to both rubrics because both the problem-specific and domain-relevant rubrics explore important but distinct aspects of class material. Before addressing feedback helpfulness, however, it is important to take into account author-reviewer reciprocity, as explained below.

Author-reviewer reciprocity. Peer reviewers and peer authors may at times engage in tit-for-tat reciprocal behavior (Kwangsu Cho & Kim, 2007). Authors receiving high inbound peer ratings may respond with high back-review ratings, while low inbound peer ratings may elicit low back-review ratings. Since problem-specific criteria may be easier to apply objectively than domain-relevant criteria, authors may find it easier to evaluate such objective feedback on its own merits. If so, there may be a decrease in reciprocal behavior among authors receiving problem-specific feedback.

2. Methods

2.1 Participants

All 58 participants were second or third year students at a major US law school, enrolled in a course on Intellectual Property law. Students were required to take an essay-type midterm examination (Goldin, 2011, Appendix A) and to participate in the subsequent peer-review exercise. Students were asked to perform a good-faith job of reviewing. The syllabus indicated, "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."

2.2 Apparatus

The study was conducted via Comrade, a web-based application for peer review. For purposes of this study, Comrade can be seen as similar to other peer-review applications, including SWoRD and Aropa (Kwangsu Cho & Schunn, 2007; Hamer, Kell, & Spence, 2007). Comrade was configured to conduct peer review in the following manner:

1. Students wrote essays and uploaded them into Comrade.
2. Essays were distributed to a group of 4 student peers for review.
3. The peer reviewers submitted their feedback to the essay authors.
4. The authors gave back-reviews to the peer reviewers.

Students were free to choose their word processing software in step 1, but they were required to save their essays in a digital file format that other students could read. In step 2, student authors uploaded their essays into Comrade for distribution to reviewers, and Comrade enforced a check on acceptable file formats. To facilitate anonymity in peer review, each student was able to choose a nickname directly in Comrade, and was identified to other students only by that nickname (Lu & Bol, 2007). After the deadline passed for uploads, Comrade randomly assigned students to review each other's work using an algorithm that ensures that the reviewing workload is distributed fairly, and that all authors receive a fair number of reviews (Zhi-Feng Liu, San-Ju Lin, & Yuan, 2002). At this point, students were able to download each other's papers from Comrade, read them and enter their feedback (step 3). Reviewer feedback was elicited according to either the domain-relevant or the problem-specific rubric, which are described below. In step 4, reviewer feedback was delivered to student authors, and the authors were asked to evaluate feedback helpfulness.

In addition, students were asked a series of multiple choice questions about the legal concepts that they were studying between steps 1 and 2 and again between steps 3 and 4, also online via Comrade.[3] After step 4, all students were invited to fill out an optional survey.

2.3 Research Design

Just prior to the peer-review exercise, participants completed writing a mid-term, open-book, take-home examination. It comprised one essay-type question, and student answers were limited to no more than four double-spaced or 1.5-spaced typed pages. Students had 3 days to answer the exam question. The question presented a fairly complex (2-page, 1.5-spaced) factual scenario involving a computer science instructor, Professor Smith, who created a product that was somewhat similar to an idea revealed to him by a former pupil. Students were asked to provide advice concerning a particular party's rights and liabilities given the scenario. The instructor designed the facts of the problem to raise issues involving some of the legal claims and concepts that were discussed in the first part of the course. Specifically, the instructor focused on plausible legal 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 a provision of the federal trademark law. In addition, various other parties had plausible legal claims against Professor Smith for idea misappropriation and violating the right of publicity. Students were expected to analyze the facts, identify the claims and issues raised, make arguments pro and con resolution of the issue in terms of the concepts, rules, and cases discussed in class, and make recommendations accordingly. Because the instructor was careful to include factual weaknesses as well as strengths for each claim, the problem was open-ended; strong arguments could be made for and against each party's claims.

The experiment was administered as a between-subjects treatment. Students used one of two rubrics to review the work of their peers, either the domain-relevant rubric (domain-relevant condition) or the problem-specific rubric (problem-specific condition). Students only received feedback from reviewers within the same condition. There was no training of students in evaluating peer works.

Independently of the peer review activities, the instructor assigned an overall score to each student's essay. This single score reflected those criteria explicitly called out in the rubrics as well as any others that the instructor felt were relevant to his assessment of the essays. The scores were numeric; the letter grades for which the scores served as a basis were not used in the study reported here.

For each dimension within their assigned rubric, students were asked to give a rating of the peer author's work and to comment on that rating. In other words, these were analytic rubrics where each rating and comment focused on a specific dimension of the work rather than merely contributing to a holistic impression. Although the comments were not analyzed formally in this research due to time constraints, they were collected to fulfill the three functions of feedback: the ratings were grounded with respect to a scale so that the peer authors could see how their peers evaluate their current level of performance, the same scale also showed the target level of performance, and the comment was intended to help peer authors understand how to reach the target level of performance, all with respect to distinct dimensions. The rating scales had 7 points, grounded at 1,3,5,7 (Figure 1, Figure 2).

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

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

Legal Claims:

Smith v. Barry for breach of the nondisclosure/noncompetition agreement **(nda)**

Smith v. Barry and VG for trade-secret misappropriation **(tsm)**

Jack v. Smith for misappropriating Jacks idea for the I-phone-based instrument-controller interface **(idea1)**

Barry v. Smith for misappropriating Barrys idea for the design of a Jimi-Hydrox-related look with flames for winning **(idea2)**

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

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

The rubrics differ in their generality and orientation (see New Concepts for Rubrics). Because they were devised for the same exercise, some overlap is inevitable. Domain-relevant "issues" are very similar to the five legal claims that are the dimensions of the problem-specific rubric, but only the problem-specific rubric calls them out explicitly. Out of the many concepts that make up intellectual property law, only a few pertain to this exercise. As a consequence, it was possible to orient the problem-specific rubric to legal concepts. By contrast, the domain-relevant rubric, while potentially applicable to other exercises, cannot be concept-oriented because there would be too many concepts to list. Instead, the domain-relevant rubric is oriented to rhetorical skills, such as identifying issues and making arguments. The problem-specific rubric does not omit these rhetorical and analytical skills, because it can echo them in the rating scale used for each dimension. Thus, the chief difference between the rubrics is in how they represent or decompose the underlying criteria. Additionally, the domain-relevant dimensions on justifying an overall conclusion and on mechanics of writing do not overlap with the problem-specific rubric.

After receiving feedback, each author was asked to give a back-review of the feedback on each dimension to each reviewer. The same condition-neutral back-review scale was administered for all dimensions as appropriate to the experimental treatment so that feedback from each problem-specific and domain-relevant criterion was evaluated on its own merits (Figure 3).

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

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

In a posthoc analysis to validate the problem-specific rubric, a trained rater used the rubric to rate all papers in the problem-specific condition. The rater was a former student that had previously excelled in the same course. To help the rater emulate the instructor's scoring, the training was that, first, the former student was given an answer key to the exam question that had been prepared by the instructor. Second, the student rated four midterm essays that were chosen as representing a variety of levels of performance of each concept, using the answer key, and the instructor and the student discussed any differences of opinion. After this training, the student rated the remaining papers using the problem-specific criteria and the answer key. (Validation was particularly important for the problem-specific rubric given its novelty. Limited resources did not allow us to validate both rubrics or to invite multiple raters whose agreement could be measured.)

Thus, the manipulation consisted of assigning students to give and receive feedback according to either the problem-specific rubric or the domain-relevant rubric. The students' feedback to peer authors and back-reviews were collected as dependent variables. The participants Law School Admission Test (LSAT) scores were also collected (48 of 58 students opted to allow their LSAT scores to be used), as well as the midterm papers themselves, and the instructor-assigned scores on the papers. Finally, a trained rater used the problem-specific rubric to rate all papers in the problem-specific condition.

2.4 Procedure

On Day 1, 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. Then participants were randomly assigned to one of the two conditions in a manner balanced with respect to their LSAT scores. Students then completed a multiple choice, conceptually oriented test ("pretest") in one of two test forms. Half of the students in each condition received form A and half received form B. From Day 3 to 7, students logged in to review the papers of the other students. Each student received four papers to review, and each review was anticipated to take about 2 hours. After reviewing but before receiving feedback from other students, each student completed a second multiple choice test ("posttest"); those students who had earlier completed test form A now completed test form B, and vice versa. On Day 8, students logged in to receive reviews from their classmates. On Day 10, students provided the reviewers with back-reviews explaining whether the feedback was helpful. Students also took a brief survey on their peer review experience.

3. Results

Table 1. Characteristics of rubric dimensions: mean and standard deviation of inbound peer ratings (IPR); correlation to trained rater; single-rater reliability (SRR); effective reliability (EFR); reciprocity; helpfulness (inbound back-review ratings or IBR)

Domain-relevant Dimension	Mean IPR (SD)	Peer vs. Trained *r* [95% CI]	SRR [95% CI]	EFR [95% CI]	Reciprocity (τ) H_0:τ=0, α=0.05	Mean IBR (SD)
argument	5.37 (1.20)	N/A	0.32 [0.13, 0.54]	0.65 [0.37, 0.82]	0.28*	5.49 (1.38)
conclusion	5.48 (1.09)	N/A	0.12 [-0.04, 0.34]	0.34 [-0.2, 0.68]	0.26*	5.64 (1.47)
issue	5.37 (1.12)	N/A	0.5 [0.31, 0.69]	0.8 [0.64, 0.9]	0.25*	5.34 (1.56)
writing	5.74 (1.36)	N/A	0.18 [0.01, 0.42]	0.48 [0.03, 0.75]	0.38*	5.82 (1.31)

Problem-specific Dimension	Mean IPR (SD)	Peer vs. Trained r [95% CI]	SRR [95% CI]	EFR [95% CI]	Reciprocity (τ) H_0:τ=0, α=0.05	Mean IBR (SD)
idea1	4.82 (1.37)	0.43 [0.07, 0.69]	0.21 [0.02, 0.45]	0.51 [0.09, 0.77]	0.27*	5.23 (1.65)
idea2	1.98 (1.59)	0.33 [-0.04, 0.63]	0.1 [-0.07, 0.34]	0.3 [-0.33, 0.67]	0.21*	4.23 (1.93)
nda	4.53 (1.66)	0.71 [0.45, 0.85]	0.62 [0.42, 0.79]	0.86 [0.74, 0.94]	0.23*	4.99 (1.73)
rop	2.79 (2.15) ($\sigma^2_{\mathrm{n[IPR]}}$ = 0.96)	0.81 [0.62, 0.91]	0.83 [0.71, 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]	0.4 [0.2, 0.63]	0.73 [0.49, 0.87]	0.24*	4.95 (1.79)

The following presentation of results addresses the effect of two different rubrics on validity and reliability of peer assessment, on reviewer responsiveness to analytic rubrics, and on feedback helpfulness. Peer feedback on student essays was gathered via one of two rubrics, either domain-relevant (n=29) or problem-specific (*n*=28). Each essay was reviewed by four peer reviewers. The instructor also scored the essays. Student authors rated the feedback that they received for helpfulness.

As an exploratory look at the dataset, the mean inbound peer rating within each rubric dimension was computed for each student author. For example, each paper receiving feedback via the domain-relevant rubric was described by four mean scores, one for each dimension of that rubric. Mean inbound peer ratings ranged from a low of 1.86 (problem-specific condition, *idea2* prompt regarding the second idea misappropriation claim) to a high of 5.54 (domain-relevant condition, *writing* prompt regarding the effect of organization on writing quality) on a 7-point rating scale (Table 1). In addition, other parameters were computed, as described below. Standard deviations were plausible for both rubrics, suggesting that there was a range of performance within each dimension. Bayesian estimates of the per-dimension standard deviation of all ratings (Goldin & Ashley, 2011) were similar to the classical point estimates, except in the case of the "right of publicity" dimension (labeled in Table 1 as $\sigma^2_{n[IPR]}$).

3.1 Assessment Validity

The validity of rubric-supported peer assessment was evaluated by correlating the peer ratings of students' essay-form midterm exam answers to summative instructor scores of the same exam answers. (But note that the peer reviewers were focused on providing formative feedback, not summative assessment.) To understand this correlation in context, both peer and instructor scores were compared against LSAT scores.

3.1.1 Peer Assessment Validity

Within each experimental treatment, every peer author's mean inbound peer rating was computed across all rating dimensions and across all reviewers. For example, for a paper by a student in the domain-relevant condition, this was the mean of 4 rating criteria * 4 reviewers = 16 inbound ratings. (The instructor only gave overall, not per-dimension, ratings, so a more fine-grained comparison was not possible.) The Pearson correlations of these means with the instructor's score of the same papers were significant for both the problem-specific condition, *r*(26)=0.73, *p*<0.001, 95% CI [0.49, 0.87], and the domain-relevant condition, *r*(27)=0.46, *p*=0.011, 95% CI [0.12, 0.71]. Both types of rubrics may be seen as valid analytical rubrics (where validity is determined by similarity to instructor scores). The problem-specific relationship between mean peer ratings and instructor scores was not significantly stronger than the domain-relevant relationship, z=1.51 using a Fisher transformation test at =0.05, i.e., the problem-specific rubric is not "more valid" than the domain-relevant one. Although some evidence points in favor of the problem-specific rubric (i.e., the narrowness of the confidence interval, the strength of the correlation), this one comparison of two rubrics is too small a sample to endorse the use of a problem-specific rubric over a domain-relevant one.

Given the novelty of the problem-specific rubric, peer ratings of the papers from the problem-specific condition were additionally validated at the level of separate dimensions by correlating them against the ratings of a trained rater. First, the mean inbound peer rating within each problem-specific dimension was computed for each paper, such that every paper was described by five mean scores, one for each dimension of that rubric. Second, a former student that had previously excelled in the same course was trained to rate papers with the problem-specific rubric. Finally, within each dimension, this trained student's rating of each paper was correlated against the mean inbound peer rating. Peer ratings in all but one problem-specific dimension correlated significantly with the ratings of the trained rater (Table 1). The lone exception was the second idea misappropriation claim for which peer rating reliability was particularly low, as shown below.

While instructor scores and peer ratings were related to each other, neither was related to students' LSAT scores. 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. The lack of correlation to LSAT performance is somewhat troubling. However, the LSAT is conventionally validated against "the average grade earned by the student in the first year of law school" (Thornton, Stilwell, & Reese, 2006), not the grades of second- and third-year students, as in this population. Furthermore, bar exam performance is better predicted by law school Grade Point Average than by the LSAT (Wightman & Ramsey, 1998).

3.1.2 Peer Assessment Reliability

Ratings produced via both domain-relevant and problem-specific rubrics were evaluated for reliability. Following Kwangsu Cho et al. (2006), reliability was computed in terms of the Intra-Class Coefficient (ICC) (McGraw & Wong, 1996). According to this formulation, reliability of ratings is defined as the proportion of variance that is due to the signal of the paper's true expression of some rating criterion rather than the noise of reviewer differences. The ICC assumes that there is a common population variance across the reviewers. Again following Kwangsu Cho et al. (2006), two versions of the ICC are particularly relevant to peer review: single-rater reliability (SRR) and effective reliability (EFR). Both treat reviewers as random (i.e., interchangeable), and both focus on reviewer consistency rather than exact agreement. SRR and EFR differ in that SRR estimates the reliability of a single, typical reviewer, while the EFR estimates the reliability of the average combined ratings given by multiple reviewers. By definition, EFR and SRR range from 0 to 1, and EFR is always greater than SRR. In the terminology of McGraw & Wong (1996), SRR is ICC(C,1), and Effective Reliability (EFR) is

ICC(C,k=4). The ICC serves as a check on the level of noise in each dimension of both rubrics.

Both rubrics had some dimensions that were not reliable (Table 1). Effective reliability for the problem-solving criteria ranged from 0.3 to 0.95, and for the domain-relevant criteria from 0.34 to 0.8. While there is no hard rule that distinguishes "good" and "bad" ICC values, effective reliability was relatively low for two of the five problem-specific dimensions, both of which pertained to idea misappropriation. Among the domain-relevant dimensions, effective reliability was relatively high only for the dimension pertaining to issue identification. It was expected that the problem-specific rubric may be easier to apply objectively than a domain-relevant rubric, and thus that the problem-specific rubric would be more reliable. The results supported that hypothesis, but once again, this comparison of only two rubrics is too small a sample to draw a definitive conclusion.

The problem-specific rubric elicited ratings at both high and low ends of the rating scale that was used for all dimensions (Table 1). The two problem-specific concepts that had the lowest mean inbound peer ratings, namely the second idea misappropriation claim and the right of publicity claim were, respectively, the least and most reliable problem-specific concepts. Notably, low peer ratings did not lead to low reliability among reviewers.

3.2 Reviewer Responsiveness to Rubric

Even if both types of rubrics elicit valid peer ratings, as established in terms of the correlation of authors' inbound peer ratings with an instructor's score, the rubrics may elicit ratings in a holistic manner, rather than an analytic manner. Each rubric was evaluated for whether the dimensions within the rubric were distinguished from each other in terms of peer ratings. In addition, given the novelty of the concept-oriented distinctions made in the problem-specific rubric, student essays were scored according to the same rubric by a trained rater.

3.2.1 Distinctions among Dimensions within Each Rubric

It is desirable for dimensions within an analytic rubric to be distinct from one another. For example, an instructor implementing a rubric likely wants peer authors to receive formative feedback that is grounded and explained in terms of each respective criterion. Further, it is a misuse of reviewer effort and author attention to give and receive feedback that turns out to be redundant and hence relatively uninformative.

To check whether the rubrics elicit differentiated ratings, first, the mean inbound peer rating within each rubric dimension was computed for each student author. For example, each paper receiving feedback via the domain-relevant rubric was described by four mean scores, one for each dimension of that rubric. Second, within each rubric, these mean scores across all papers were correlated, resulting in 6 pairwise correlations for the domain-relevant rubric (Table 3) and 10 pairwise correlations for the problem-specific rubric (Table 2). Correlations between mean inbound ratings in the domain-relevant condition were *all* strong and statistically significant. In the problem-specific condition, ratings between only two pairs of criteria were highly correlated (the first idea misappropriation claim *idea1* vs. the trade-secret misappropriation claim *tsm*, and the claim for breach of non-disclosure *nda* vs. the trade-secret misappropriation claim *tsm*). This suggests that peer reviewers treated the domain-relevant rubric as a single construct, but distinguished among multiple constructs when they used the problem-specific rubric.

Table 2. Pairwise correlations between mean inbound peer ratings among dimensions of the problem-specific rubric. In parentheses, correlations among the ratings of author work by a trained rater. Asterisk indicates correlations significantly different from zero at α=0.05.

	Problem-specific Dimensions (*r*)			
	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)

Table 3. Pairwise correlations between mean inbound peer ratings among dimensions of the domain-relevant rubric. Asterisk indicates correlations significantly different from zero at α=0.05.

	Domain-relevant Dimensions (*r*)		
	conclusion	issue	writing
argument	0.72*	0.69*	0.65*
conclusion		0.61*	0.77*
issue			0.67*

The apparent difference between rubrics was not due to the slightly different number of dimensions (4 in the domain-relevant rubric, 5 in the problem-specific). The null hypothesis is that rubric dimensions ought not to be related to each other. Simply due to chance, a significant pairwise correlation between dimensions is more probable for a rubric that has more dimensions rather than fewer. For the problem-specific rubric, there were 10 possibilities of a significant pairwise correlation, but its dimensions were actually less inter-related than domain-relevant dimensions.

The extent to which each rubric represented a unitary construct, i.e., internal consistency, was measured using McDonald's ω_h over per-dimension mean inbound peer ratings. After a factor analysis, McDonald's ω_h is "based upon the sum of the squared loadings on the general factor", i.e., it is "an index of how much the test measures one common factor" (Revelle & Zinbarg, 2008). McDonald's ω_h provides a stronger lower bound estimate of internal consistency than other measures, including Cronbach's (Zinbarg, Revelle, Yovel, & Li, 2005). As with , internal consistency is guaranteed to be restricted to the range $[\omega_h, 1]$. For the domain-relevant rubric, ω_h=0.85, and for the problem-specific rubric, ω_h. =0.56[4] In other words, as applied by the peer reviewers, the dimensions of the domain-relevant rubric represented a single unitary construct, while the dimensions of the problem-specific rubric likely differentiated among multiple constructs.

There may be several possible explanations for inter-criteria correlation in either condition. First, although peer reviewers could have rated each other inaccurately, this is unlikely given that both types of ratings are valid with respect to instructor scores. While novice writers may struggle with making revisions, detecting errors and fixing them are separate skills (Hayes et al., 1987), and rating on an anchored scale is only a recognition task, not a recall one. Further, these were second and third year law students, who must be familiar with legal argumentation, that is, with the domain-relevant rubric. Nonetheless, the following section investigates whether they missed important relationships among the problem-specific criteria because of their inexperience with Intellectual Property, which would have led to the low inter-dimension correlations.

Second, as pointed out by an anonymous reviewer of this paper, "it might also be possible that the students essays were quite homogeneous with regard to rhetorical aspects of writing quality." That is, within any one rubric dimension, the range of performance across all students could have been narrow. However, variances of ratings show that this is not the case (Table 1).

Third, it could be that some criteria are intrinsically interdependent. For example, among the domain-relevant criteria, it could be that rigorous argument structure *(argument)* is necessarily dependent on identifying the key issues in a problem *(issue)*. Analogously, among the problem-specific criteria, legal claims of trade secret misappropriation *(tsm)* do often arise in the context of breach of non-disclosure and non-competition agreements *(nda)*, which was one of the two significant correlations in that condition.

Fourth, it could be that the criteria are simply correlated in terms of how the behavior they describe is expressed by students. For example, if a student employs good grammar *(writing)*, it is likely that this student will also write good conclusions *(conclusion)*, even if one does not directly cause the other.

3.2.2 Validity of Problem-Specific Conceptual Distinctions by Peer Reviewers

Mean inbound peer ratings according to problem-specific criteria were mostly uncorrelated with each other. One explanation could be that peer reviewers missed important relationships among these criteria, which could happen if the conceptual issues were too difficult for peer reviewers to assess. To check on this, a former student that had previously excelled in the same course was trained to rate papers with the problem-specific rubric, and the correlations among these ratings were computed for each pair of criteria in the same manner as for the mean inbound peer ratings.

There were no significant pairwise correlations according to the trained rater that were missed by the peer reviewers (Table 2). Of the two significant pairwise correlations that were present according to the peer reviewers, one was also significant according to the trained rater (the first idea misappropriation claim *idea1* vs. the trade-secret misappropriation claim *tsm*), and one was not significant according to the trained rater (the claim for breach of non-disclosure *nda* vs. the trade-secret misappropriation claim *tsm*). In the aggregate, peer reviewers distinguished among problem-specific concepts similarly to the trained rater.

3.3 Feedback Helpfulness

After receiving feedback, each author was asked to give a back-review of the feedback on each dimension to each reviewer. The same condition-neutral back-review scale was administered for all dimensions as appropriate to the experimental treatment so that feedback from each problem-specific and domain-relevant criterion was evaluated on its own merits (Figure 1).

3.3.1 Author-Reviewer Reciprocity

Peer reviewers may engage in reciprocal behavior, i.e., authors may be tempted to give high back-review ratings to reviewers that give the authors' works high peer ratings, and low back-review ratings in response to low ratings from reviewers.

Reciprocity was defined operationally as the correlation between the peer ratings given by reviewers and back-review ratings given by authors in response. Owing to the ordinal nature of the ratings, correlations were computed as Kendall's τ, which is "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"(Hill & Lewicki, 2006). Reciprocity aggregated across all dimensions of the problem-specific rubric was found to be $\tau(579)=0.27$, $p<0.001$, and domain-relevant reciprocity was $\tau(463)=0.30$, $p<0.001$. Reciprocity varied little when breaking out rating dimensions (Table 1)[5]. Thus, there is a small but statistically significant amount of reviewer-author reciprocity using both rubrics. Contrary to expectations, problem-specific criteria did not make it easier for authors to evaluate feedback objectively.

3.3.2 Helpfulness of Feedback via Domain-Relevant and Problem-Specific Support

Feedback helpfulness was compared between the two conditions. Both rubrics elicited helpful feedback most of the time, as indicated by the mean helpfulness ratings in each dimension (Table 1). A straightforward test of whether helpfulness differs by rubric would be a one-way ANOVA of back-review ratings with rubric as a factor of two levels, problem-specific and domain-general. However, because there was a statistically significant level of reciprocity in the ratings elicited by both rubrics, it is necessary to adjust for inbound peer ratings in evaluating whether helpfulness differs by rubric. An ANCOVA of back-review ratings with inbound peer rating as covariate and rubric type as factor showed that the inbound peer rating was a significant predictor of the back-review rating, $F(1,830)=129.18$, $p<0.001$, which was expected due to the reciprocity finding. Having adjusted for reciprocity, rubric type was not a significant predictor of the back-review rating, $F(1,830)=2.69$, $p=0.10$, i.e., feedback helpfulness did not differ by rubric.

It is possible to interpret the raw back-review ratings by referring to the anchors of the rating scale provided to the authors (Figure 1), although this forgoes the adjustment for reciprocity. Domain-relevant feedback was rated 6 or 7 more often than problem-specific feedback. As defined by the rating scale, such ratings indicated that authors felt that the feedback not only "identified most key problems" in their writing and "suggested useful solutions", but that the feedback also "identified key strengths". Feedback that "identifies key strengths" could be considered as praising the author's work; thus the authors apparently felt that the domain-relevant feedback often contained praise.

Additionally, problem-specific feedback was rated 3 or below more often than domain-relevant feedback. To understand why problem-specific feedback was sometimes considered unhelpful, all 92 comments from peer authors that were paired with back-review ratings of 3 or lower were analyzed. In these comments, 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 often than domain-relevant reviewers. In 19 cases, authors omitted back-review ratings but left written comments, which were analyzed. The comments seemed to fit well with the back-review scale, but the authors chose to omit ratings nonetheless.

4. Discussion

This experiment compared formative assessment in peer review via problem-specific, concept-oriented support for reviewers and authors versus domain-relevant, argumentation-oriented support. The results showed that both kinds of reviewing rubrics led to valid peer assessment of student work. Examining the rubrics' analytic dimensions separately showed some differences, but given the small sample of the comparison (just one of each type of rubric), this comparison is tentative and further study is required. Dimensions of the problem-specific rubric were reliable more often than dimensions of the domain-relevant rubric. The domain-relevant rubric showed high inter-dimension correlation. The problem-specific rubric did not show high inter-dimension correlation according to peer reviewers, and this was confirmed by a trained rater. When evaluating feedback helpfulness, authors "reciprocated" by giving low back-review ratings in response to low peer ratings. Adjusting for reciprocity showed that peer authors judged feedback elicited by both rubrics to be similarly helpful, but the domain-relevant rubric elicited praise more often than the problem-specific rubric. Some considerations on choosing between a domain-relevant rubric and a problem-specific rubric follow.

Both domain-relevant and problem-specific mean inbound peer ratings correlated strongly with an instructor's aggregate scores of a midterm exam in Intellectual Property law. The validity of the problem-specific ratings within each dimension was further confirmed against the ratings of a trained rater. This is an especially important finding for a course in law, a domain of open-ended problems, where it is difficult to achieve reproducible assessment and to do so with plausibly valid criteria.

The high inter-dimension correlation of the domain-relevant rubric is a strike against the domain-relevant rubric. The most likely explanation is that the domain-relevant dimensions were inherently correlated in this corpus. Even if a rubric addresses what can hypothetically be different skills (e.g., argumentation vs. issue identification), students may acquire these skills together, and the skills may also manifest themselves together. Having found redundancy in peer ratings, we cannot know if comments were similarly redundant. That said, an instructor who cannot anticipate whether or not student essays will be correlated in terms of domain-relevant criteria may reasonably choose a problem-specific rubric. Because problem-specific support to reviewers leads to ratings that do not correlate with each other, such ratings are not redundant, and more likely to be informative.

While neither rubric was reliable across all dimensions, lack of reliability is not necessarily a cause for concern; indeed, the importance of reliability in peer review may be overstated (N. F. Liu & Carless, 2006). In particular, this may not be a concern when rubrics are applied to an open-ended problem. In this study, the problem-specific rubric emphasized legal claims, each of which provides a separate analytical framework for the open-ended problem. Further, reviewer disagreements with respect to conceptual issues could be legitimate because open-ended problems may be framed in multiple ways. The different problem-specific reviews may thus lead authors to see the problem in different ways.

Reliability may be important in some cases, especially summative assessment: "For 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" (Kwangsu Cho & Schunn, 2007). If so, reliability may be improved by increasing the number of peer reviewers (Kwangsu Cho & Schunn, 2007), and by calibrating their rating techniques (Russell, Cunningham, & George, 2004). Improving reliability may be easier with the problem-specific rubric than the domain-relevant one, because problem-specific criteria can be clear and specific (e.g., what factors may support a legal claim of trade-secret misappropriation, what factors may constitute a good response to such a claim), while domain-relevant criteria are more open-ended (e.g., what constitutes a good legal argument in general). Notably, "fixing" the problem of reliability for problem-specific criteria leads the instructor to teach material that is very appropriate to the topic of the course. If reliability of feedback is less important than multiple perspectives, and if there is a mechanism to encourage diversity in feedback, the number of reviewers can be reduced, which would also reduce the burden of reviewing for the students.

Back-review ratings for both types of rubrics were affected by a small but statistically significant amount of reviewer-author reciprocity. While it is possible to eliminate reciprocity by concealing peer ratings, i.e., by only presenting comments to peer authors (Kwangsu Cho & Kim, 2007), this may be undesirable. The ratings may communicate formative feedback to students, including level of current performance and the target level of performance, and the structure of the criteria. Additionally, it is awkward to collect ratings without passing them on.

Adjusting for reciprocity showed that peer authors judged feedback elicited by both rubrics to be similarly helpful. Domain-relevant reviews earned back-review ratings that noted praise in the review more often than the problem-specific rubric, but students are known to rate praise as helpful (K. Cho, Schunn, & Charney, 2006), and praise is not associated with implementation of feedback in a subsequent draft (Nelson, 2008). By contrast, problem-specific reviews earned more back-review ratings indicating that the feedback identified problems but lacked solutions. From an instructor's perspective, low ratings of helpfulness of problem-specific feedback, if not overwhelming in number, are a *positive* aspect of a peer review exercise. Much as low inbound peer ratings inform the instructor that a particular problem-specific concept has proved challenging for students, low back-review ratings inform the instructor that students struggle with giving helpful feedback for a problem-specific concept, which may indicate that students do not understand the concept. In future research, a fair comparative evaluation of helpfulness would entail delivering both problem-specific and domain-relevant feedback to each author to see which type the authors prefer when they can choose among them.

The work sheds some light on how criterion-based formative writing assessment may be implemented in domain-specific writing contexts. Specifically, the classification of assessment rubrics in terms of generality and orientation may aid in designing rubrics for open-ended problems in writing courses in the disciplines. It could be that for some courses it is important to distinguish the writing and critiquing skills that make up a domain-relevant rubric, and to collapse the various problem-specific concepts. These are likely to be courses focused on writing as a subject in itself. However, for courses with substantive subject matter apart from (or in addition to) writing, review rubrics that are more concept-oriented and that address aspects of specific open-textured problems, may be of greater pedagogical value.

The two rubrics evaluated here share some underlying criteria, but present them from different perspectives. For example, both rubrics place value on identifying and making reasoned arguments about conceptual issues. It would be natural for reviewers using domain-relevant support to discuss problem-specific concepts. However, any problem-specific concept-oriented information would be distributed across the domain-relevant dimensions, attenuating the concept-oriented signal, and leading to interference from multiple concepts and non-concept-oriented feedback. Moreover, if that feedback is to be evaluated according to the domain-relevant back-review scale, that would lead back-reviews to pertain to the reviewers' ability to give domain-relevant feedback, not concept-oriented feedback. Thus, instructors who value conceptual analysis should choose the problem-specific rubric over the domain-relevant one.

The domain-relevant rubric is more general than the problem-specific rubric; its support encompasses more exercises. However, evidence that points in favor of the problem-specific rubric includes the strength of the validity correlation, the narrowness of the correlation confidence interval, the larger number of dimensions with high effective reliability, and the low inter-dimension correlation. These strengths are due to the fact that the problem-specific rubric is oriented to legal concepts, not to skills. Further, it is possible to increase the support of the problem-specific rubric without hurting its orientation to concepts. The problem-specific rubric could be ported to a new exercise simply by adding dimensions for those legal concepts that are relevant to the new exercise, and omitting irrelevant dimensions. The new dimensions can be defined using the same rating scale as the existing ones.

While this one comparison of two rubrics is too small a sample to endorse the use of one rubric type over another, it should serve as a starting point for further principled exploration of rubric design and its impact on assessment of writing in the disciplines.

5. Conclusions

This research makes theoretical and applied contributions. First, it introduces support, generality, porting and orientation of review criteria as concepts that are useful for thinking about rubrics. Second, it describes two new rubrics for legal writing derived on the basis of these concepts: a domain-relevant, skill-oriented rubric and a problem-specific, concept-oriented rubric. Third, it evaluates and compares the new rubrics in terms of validity, reliability, reviewer responsiveness, and feedback helpfulness. Fourth, it provides an example of how computer-supported peer review helps to study rubrics.

The limitations of the research are that it only considers peer ratings, leaving peer comments to be examined in future work. While the rubric comparison is informative, it is not determinative in that covers only one example of each type of rubric. In hindsight, it is apparent that the evaluation of feedback helpfulness would have been more robust if each author had received feedback via both rubrics.

The lesson we take away is this: Rubrics affect the experience of students and instructors; they are not neutral. Given that rubrics are widely used and endorsed (H. G.

Andrade, 2000; Stiggins, 2005), they deserve more critical attention than is usually accorded to them either by instructors or researchers. Our experience highlights a synergy between rubrics and peer review: rubrics provide critical support and valuable insights for peer review in educational settings, and peer review provides an excellent laboratory for evaluating rubrics.

Notes

1. The duopoly of validity and reliability has itself been criticized for ignoring other characteristics (e.g., Baartman, Bastiaens, Kirschner, & Vandervleuten, 2006), including in peer assessment settings (Ploegh, Tillema, & Segers, 2009).
2. This simplified discussion does not distinguish 'criterion' and 'standard'; cf. (D.Royce Sadler, 1987).
3. Test results were inconclusive and discussion is omitted here. Further details are available in (Goldin, 2011).
4. Computed using the psych package ver. 1.2.1.
5. In prior work, reciprocity was defined as the Pearson correlation (Kwangsu Cho & Kim, 2007), which produces similar results for the problem-specific and domain-relevant ratings.

Acknowledgments

This research was supported in part by the Provost's Advisory Council on Instructional Excellence at the University of Pittsburgh. We thank Christian D. Schunn, Peter Brusilovsky, and Louis Gomez for serving as Ilya Goldin's doctoral dissertation committee.

References

Alamargot, D., & Chanquoy, L. (2001). *Through the models of writing.* Boston, MA: Kluwer Academic Publishers. doi: 10.1007/978-94-010-0804-4

Andrade, H. G. (2000). Using rubrics to promote thinking and learning. *Educational Leadership, 57,* 13–19.

Andrade, H., & Du, Y. (2007). Student responses to criteria-referenced self-assessment. *Assessment & Evaluation in Higher Education, 32,* 159–181. doi:10.1080/02602930600801928

Baartman, L., Bastiaens, T., Kirschner, P., & Vandervleuten, C. (2006). The wheel of competency assessment: Presenting quality criteria for competency assessment programs. *Studies In Educational Evaluation, 32,* 153–170. doi:10.1016/j.stueduc.2006.04.006

Baker, M., & Lund, K. (1997). Promoting reflective interactions in a computer-supported collaborative learning environment. *Journal of Computer Assisted Learning, 13,* 175–193.

Bazerman, C., Little, J., Bethel, L., Chavkin, T., Fouquette, D., & Garufis, J. (2005). Reference guide to writing across the curriculum. West Lafayette, IN: Parlor Press.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale N.J.: L. Erlbaum Associates.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook On Formative and Summative Evaluation of Student Learning.* McGraw-Hill Customer Service.

Chalk, B., & Adeboye, K. (2005). Peer Assessment Of Program Code: a comparison of two feedback instruments. In *Proceedings of the 6th Annual Conference for the Higher Education Academy Subject Network for Information and Computer Science (HEA-ICS)*. University of York, UK.

Chi, M. T. H., de Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18,* 439–477. doi:10.1016/0364-0213(94)90016-7

Cho, K., & Kim, B. (2007). Suppressing competition in a computer-supported collaborative learning system. In *Proceedings of the 12th International Conference on Human-Computer Interaction: Applications and Services* (pp. 208–214). Beijing, China: Springer-Verlag.

Cho, K., & Schunn, C. D. (2007). Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers and Education, 48,* 409–426. doi:10.1016/j.compedu.2005.02.004

Cho, K., Cho, M.-H., & Hacker, D. J. (2010). Self-monitoring support for learning to write. *Interactive Learning Environments, 18*, 101–113. doi:10.1080/10494820802292386

Cho, K., Schunn, C. D., & Charney, D. (2006). Commenting on Writing. Typology and Perceived Helpfulness of Comments from Novice Peer Reviewers and Subject Matter Experts. *Written Communication, 23,* 260–294.

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

Cizek, G. J. (2010). An Introduction to Formative Assessment. In H. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 3–17). New York: Routledge.

Cooper, C. R. (1977). Holistic evaluation of writing. In C. R. Cooper & L. Odell (Eds.), *Evaluating Writing: Describing, Measuring, Judgin*g (pp. 3–32). National Council of Teachers of English.

Deane, P., & Quinlan, T. (2010). What automated analyses of corpora can tell us about students' writing skills. *Journal of Writing Research, 2,* 151–177.

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). *Factors in judgements of writing ability* (Research Bulletin No. RB 61-15). Princeton, NJ: Educational Testing Services.

Draaijer, S., & van Boxel, P. (2006). Summative peer assessment using "Turnitin" and a large cohort of students: a case study. In M. Danson (Ed.), *Proceedings of 10th International Computer Assisted Assessment Conference* (pp. 167–180). Loughborough University, UK: Professional Development Loughborough University. Retrieved from http://hdl.handle.net/2134/4559

Gansle, K. A., VanDerHeyden, A. M., Noell, G. H., Resetar, J. L., & Williams, K. L. (2006). The Technical Adequacy of Curriculum-Based and Rating-Based Measures of Written Expression for Elementary School Students. *School Psychology Review, 35*, 435–450.

Godshalk, F. I., Swineford, F., & Coffman, W. E. (1966). *The Measurement of Writing Ability* (No. CEEB RM No. 6.). Princeton, NJ: College Entrance Examination Board.

Goldin, I. M. (2011). *A Focus on Content: The Use of Rubrics in Peer Review to Guide Students and Instructors.* Retrieved from http://etd.library.pitt.edu/ETD/available/etd-07142011-004329/

Goldin, I. M., & Ashley, K. D. (2011). Peering Inside Peer Review with Bayesian Models. In G. Biswas, S. Bull, J. Kay, & A. Mitrovi (Eds.), *Artificial Intelligence in Educatio*n (Vol. 6738, pp. 90–97). Auckland, New Zealand: Springer.

Goldin, I. M., Ashley, K. D., & Schunn, C. D. (2012). Redesigning Educational Peer Review Interactions Using Computer Tools: An Introduction. *Journal of Writing Research,* 4(2), 111-119. doi: 10.1007/978-3-642-21869-9_14

Goldin, I. M., Brusilovsky, P., Schunn, C., Ashley, K. D., & Hsiao, I.-H. (Eds.). (2010). *Workshop on Computer-Supported Peer Review in Education,* 10th International Conference on Intelligent Tutoring Systems. Pittsburgh, PA. Retrieved from http://cspred.org

Hacker, D. J., Keener, M. C., & Kircher, J. C. (2009). Writing is Applied Metacognition. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of Metacognition in Education* (pp. 154–172). Routledge.

Hamer, J., Kell, C., & Spence, F. (2007). Peer assessment using Aropä. In *Proceedings of the 9th Australasian Conference on Computing Education* (Vol. 66, pp. 43–54). Ballarat, Victoria, Australia: Australian Computer Society, Inc.

Harris, C. E., Pritchard, M. S., & Rabins, M. J. (2000). *Engineering Ethics: Concepts and Cases* (Vol. 2). Belmont, CA: Wadsworth.

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research, 77*, 81–112. doi:10.3102/003465430298487

Hayes, J. R. (1996). A New Framework for Understanding Cognition and Affect in Writing. In C. M. Levy & S. Ransdell (Eds.), *The Science of Writing: Theories, Methods, Individual Differences, and Applications* (pp. 1–28). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Hayes, J. R., Flower, L., Schriver, K. A., Stratman, J. F., & Carey, L. (1987). Cognitive processes in revision. In S. Rosenberg (Ed.), *Advances in applied psycholinguistics* (Vol. 2, pp. 176–240). Cambridge University Press.

Hill, T., & Lewicki, P. (2006). *Statistics: methods and applications: a comprehensive reference for science, industry, and data mining.* StatSoft, Inc.

Huot, B. (1990). Reliability, Validity, and Holistic Scoring: What We Know and What We Need to Know. *College Composition and Communication, 41*, 201–213. doi:10.2307/358160

Hübner, S., Nückles, M., & Renkl, A. (2006). Prompting cognitive and metacognitive processing in writing-to-learn enhances learning outcomes. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 357–362). Vancouver, Canada: Lawrence Erlbaum Associates.

Jewell, J., & Malecki, C. K. (2005). The Utility of CBM Written Language Indices: An Investigation of Production-Dependent, Production-Independent, and Accurate-Production Scores. *School Psychology Review, 34*, 27–44.

King, A. (1997). ASK to THINK-TEL WHY: A model of transactive peer tutoring for scaffolding higher level complex learning. *Educational Psychologist, 32*, 221–235. doi: 10.1207/s15326985ep3204_3

Kwok, R. C. W., & Ma, J. (1999). Use of a group support system for collaborative assessment. Computers & Education, 32, 109–125. doi:10.1016/S0360-1315(98)00059-1

Lee, Y. W., Gentile, C., & Kantor, R. (2008). *Analytic scoring of TOEFL CBT essays: Scores from humans and e-rater. (TOEFL Research Report No. RR–81) (p. 84). Princeton, NJ: Educational Testing Service*

Lin, S. S. J., Liu, E. Z. F., & Yuan, S. M. (2001). Web-based peer assessment: feedback for students with various thinking-styles. *Journal of Computer Assisted Learning, 17,* 420–432. doi:10.1046/j.0266-4909.2001.00198.x

Lindblom-Ylänne, S., Pihlajamaki, H., & Kotkas, T. (2006). Self-, peer- and teacher-assessment of student essays. *Active Learning in Higher Education, 7,* 51–62. doi:10.1177/1469787406061148

Liu, N. F., & Carless, D. (2006). Peer feedback: the learning element of peer assessment. *Teaching in Higher Education, 11,* 279–290.

Lloyd-Jones, R. (1977). Primary Trait Scoring. In C. R. Cooper & L. Odell (Eds.), *Evaluating Writing: Describing, Measuring, Judging* (pp. 33–68). National Council of Teachers of English.

Lu, R., & Bol, L. (2007). A comparison of anonymous versus identifiable e-peer review on college student writing performance and the extent of critical feedback. *Journal of Interactive Online Learning, 6*, 100–115.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1,* 30–46. doi:10.1037/1082-989X.1.1.30

McNamara, T. F. (1990). Item Response Theory and the validation of an ESP test for health professionals. *Language Testing, 7,* 52–76. doi:10.1177/026553229000700105

Miller, P. J. (2003). The effect of scoring criteria specificity on peer and self-assessment. *Assessment & Evaluation in Higher Education, 28,* 383–94. doi:10.1080/0260293032000066218

Nelson, M. (2008). *The nature of feedback: how different types of peer feedback affect writing performance*. Retrieved from http://etd.library.pitt.edu/ETD/available/etd-12072007-100802/

Neuwirth, C. M., Chandhok, R., Charney, D., Wojahn, P., & Kim, L. (1994). Distributed collaborative writing: a comparison of spoken and written modalities for reviewing and revising documents. In *Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence* (pp. 51–57). New York, NY, USA: ACM.

O'Neill, P., Moore, C., & Huot, B. (2009). *A guide to college writing assessment*. Logan, Utah: Utah State University Press.

Patterson, E. (1996). The analysis and application of peer assessment in nurse education, like beauty, is in the eye of the beholder. *Nurse Education Today, 16,* 49–55. doi:10.1016/S0260-6917(96)80093-1

Ploegh, K., Tillema, H. H., & Segers, M. S. R. (2009). In search of quality criteria in peer assessment practices. *Studies In Educational Evaluation, 35*, 102–109. doi:10.1016/j.stueduc.2009.05.001

Revelle, W., & Zinbarg, R. E. (2008). Coefficients Alpha, Beta, Omega, and the glb: Comments on Sijtsma. *Psychometrika, 74,* 145–154. doi:10.1007/s11336-008-9102-z

Rijlaarsdam, G. (1987). Effects of Peer Evaluation on Writing Performance, Writing Processes, and Psychological Variables. In *Proceedings of the 38th Annual Meeting of the Conference on College Composition and Communication*. Atlanta, Georgia.

Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Sciences, 4,* 155–169. doi:10.1007/BF01405730

Russell, A. A., Cunningham, S., & George, Y. S. (2004). 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*. American Association for the Advancement of Science.

Sadler, D. R. (1983). Evaluation and the improvement of academic learning. *The Journal of Higher Education, 54*, 60–79.

Sadler, D. R. (1987). Specifying and Promulgating Achievement Standards. *Oxford Review of Education, 13*, 191–209. doi:10.1080/0305498870130207

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

Sanders, K., & Thomas, L. (2007). Checklists for grading object-oriented CS1 programs: concepts and misconceptions. *ACM SIGCSE Bulletin, 39*, 166–170. doi:10.1145/1269900.1268834

Scardamalia, M., & Bereiter, C. (1994). Computer Support for Knowledge-Building Communities. *Journal of the Learning Sciences, 3,* 265–283. doi:10.1207/s15327809jls0303_3

Scriven, M. (1966, mar). *Social Science Education Consortium*. Publication 110, the Methodology of Evaluation.

Shepard, L. A. (2006). *Classroom Assessment*. (R. L. Brennan, Ed.)ACE/Praeger Series on Higher Education. Praeger.

Shute, V. J. (2008). Focus on Formative Feedback. *Review of Educational Research, 78,* 153–189. doi:10.3102/0034654307313795

Spandel, V., & Stiggins, R. J. (1996). *Creating Writers: Linking Writing Assessment and Instruction*. Addison Wesley Publishing Company.

Stiggins, R. J. (2005). *Student-involved assessment for learning*. Pearson/Merrill Prentice Hall.

Strijbos, J.-W., & Sluijsmans, D. (Eds.). (2010). Special Issue on Unravelling Peer Assessment. Learning and Instruction, 20 (4), 265–348.

Thornton, A. E., Stilwell, L. A., & Reese, L. M. (2006). *The Validity of Law School Admission Test Scores for Repeaters: 2001 Through 2004 Entering Law School Classes (LSAT Technical Report No. TR 06-02) (p. 21). Newtown, PA: Law School Admission Council.*

Torgerson, W. S., Theory, S. S. R. C. (. C. on S., & Methods. (1958). *Theory and methods of scaling* (Vol. 1967). Wiley New York.

Toulmin, S. E. (2003). *The Uses of Argument*. Cambridge University Press.

Turner, S. A. (2009). *Peer Review in CS2: the Effects on Attitudes, Engagement, and Conceptual Learning.* Retrieved from http://scholar.lib.vt.edu/theses/available/etd-08272009-003738/

Voss, J. F., & Post, T. A. (1988). On the solving of ill-structured problems. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise* (pp. 261–285). Hillsdale, NJ: Lawrence Erlbaum.

Walvoord, M. E., Hoefnagels, M. H., Gaffin, D. D., Chumchal, M. M., & Long, D. A. (2008). An analysis of Calibrated Peer Review (CPR) in a science lecture classroom. *Journal of College Science Teaching, 37,* 66–73.

Wightman, L. F., & Ramsey, H., Jr. (1998). *LSAC National longitudinal bar passage study (p. 112). Newtown, PA: Law School Admission Council.*

Williamson, M. (1994). The Worship of Efficiency: Untangling Theoretical and Practical Considerations in Writing Assessment. *Assessing Writing, 1,* 147–73.

Wolcott, W., & Legg, S. M. (1998). *An Overview of Writing Assessment: Theory, Research, and Practice*. Urbana, Illinois: National Council of Teachers of English.

Wooley, R., Was, C. A., Schunn, C. D., & Dalton, D. W. (2008). The effects of feedback elaboration on the giver of feedback. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 2375–2380). Washington, DC: Cognitive Science Society.

Yancey, K. B. (1999). Looking Back as We Look Forward: Historicizing Writing Assessment. *College Composition and Communication, 50,* 483–503. doi:10.2307/358862

Zeller, A. (2000). Making students read and review code. *SIGCSE Bull., 32*, 89–92. doi:10.1145/353519.343090

Zhi-Feng Liu, E., San-Ju Lin, S., & Yuan, S.-M. (2002). To Propose a Reviewer Dispatching Algorithm for Networked Peer Assessment System. In L. Kinsthuk, K. Akahori, R. Kemp, T. Okamoto, L. Henderson, & C. Lee (Eds.), *Proceedings of the International Conference on Computers in Education.* Auckland, New Zealand: IEEE Computer Society.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's Alpha, Revelle's Beta, and Mcdonald's Omega h: their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70,* 123–133. doi:10.1007/s11336-003-0974-7