A FOCUS ON CONTENT:
THE USE OF RUBRICS IN PEER REVIEW TO GUIDE STUDENTS AND INSTRUCTORS

by

Ilya M. Goldin

B.S., Stevenson University (formerly Villa Julie College), 1999

M.S., University of Pittsburgh, 2007

Submitted to the Graduate Faculty of

Arts & Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2011

UNIVERSITY OF PITTSBURGH

ARTS & SCIENCES

This dissertation was presented

by

Ilya M. Goldin

It was defended on

April 29, 2011

and approved by

Kevin D. Ashley, Professor, School of Law, and Intelligent Systems Program

Peter Brusilovsky, Associate Professor, School of Information Sciences, and Intelligent Systems Program

Louis Gomez, Professor, School of Education

Christian D. Schunn, Associate Professor, Psychology, Learning Sciences and Policy Program, and Intelligent Systems Program

Dissertation Director: Kevin D. Ashley, Professor, School of Law, and Intelligent Systems Program

A FOCUS ON CONTENT:
THE USE OF RUBRICS IN PEER REVIEW
TO GUIDE STUDENTS AND INSTRUCTORS

Ilya M. Goldin, PhD

University of Pittsburgh, 2011

Students who are solving open-ended problems would benefit from formative assessment, i.e., from receiving helpful feedback and from having an instructor who is informed about their level of performance. Open-ended problems challenge existing assessment techniques. For example, such problems may have reasonable alternative solutions, or conflicting objectives. Analyses of open-ended problems are often presented as free-form text since they require arguments and justifications for one solution over others, and students may differ in how they frame the problems according to their knowledge, beliefs and attitudes.

This dissertation investigates how peer review may be used for formative assessment. Computer-Supported Peer Review in Education, a technology whose use is growing, has been shown to provide accurate summative assessment of student work, and peer feedback can indeed be helpful to students. A peer review process depends on the rubric that students use to assess and give feedback to each other. However, it is unclear how a rubric should be structured to produce feedback that is helpful to the student and at the same time to yield information that could be summarized for the instructor.

The dissertation reports a study in which students wrote individual analyses of an open-ended legal problem, and then exchanged feedback using Comrade, a web application for peer review. The study compared two conditions: some students used a rubric that was relevant to legal argument in general (the domain-relevant rubric), while others used a rubric that addressed the conceptual issues embedded in the open-ended problem (the problem-specific rubric).

While both rubric types yield peer ratings of student work that approximate the instructor's scores, ratings elicited by the domain-relevant rubric was redundant across its dimensions. On the contrary, peer ratings elicited by the problem-specific rubric distinguished among its dimensions. Hierarchical Bayesian models showed that ratings from both rubrics can be fit by pooling information across students, but only problem-specific ratings are fit better given information about distinct rubric dimensions.

Contents

List of Tables

List of Figures

Preface

Dissertation committee member Chris Schunn once heard me comment that "I was dissertating... or being dissertated." He clarified, "no, you're dissertating yourself."

Throughout the dissertating, my advisor and committee were gentle, critically constructive and constructively critical, and consistently supportive. Chris Schunn's research on peer review served as an impetus and foil to my own. Chris freely provided guidance, encouragement, and lucid critique. The inspiration of having Peter Brusilovsky and Louis Gomez on the committee (and the prospect of having to answer to them at the defense) played no small role in forcing me to think rigorously about my research.

Kevin Ashley, as graduate advisor and dissertation director, guided and supported me in innumerable ways, small and large, professional and personal, from the moment I enrolled at Pitt. His research is meaningful, interdisciplinary, and timeless rather than trendy. I am deeply honored to have worked with him. My only regret at concluding the dissertation is that our collaboration may also be at its end.

The research benefited from external advice and assistance. Matthew Klepacz, from the Statistics Department consulting service directed by Alan Sampson, provided clear-headed perspective on a problem that risked turning into a time-sink. Carnegie Mellon faculty were always surprisingly welcoming to a Pitt person. There, Howard Seltman had patience with my statistical naivete, and shared his wisdom and his software for Bayesian modeling. Anonymous reviewers from the Intelligent Tutoring Systems 2010 and Artificial Intelligence in Education 2011 conferences, and from the Journal of Writing Research helped with insightful and demanding feedback. Joshua Powell and Daniel Levine provided assistance at crucial junctures.

During a part of my graduate career I was employed at JustSystems Evans Research, Inc. I am grateful to David Evans and colleagues for helping me grow intellectually and professionally, and for encouraging my academic pursuits.

The dissertation would not exist without the love and support of my family. I dedicate this work to my grandmother, Faina Isaakovna Purizhanskaya, who encouraged me every step of the way, even as I was dissertating myself.

It must be stressed that the peer review process we propose is not to be used for promotion or salary increases, but is directed at:

1) programmer education
2) improving cooperation and communication in a programming team

3) self-evaluation .

(Anderson & Shneiderman, 1977)

As yesterday's positive report card shows, childrens do learn when standards are high and results are measured.
—President George W. Bush, speaking in support of the No Child Left Behind Act, New York, Sept. 26, 2007. (Weisberg, 2007)

Although the students may accept a teacher's judgment without demur, they need more than summary grades if they are to develop expertise intelligently.
(Sadler, 1989)

1.0 Introduction

It is desirable to offer feedback to students who are working on open-ended problems, and to inform instructors about the progress that the students are making. Open-ended problems may have reasonable alternative solutions, or conflicting objectives. When solving such problems, students may frame the problems according to their knowledge, beliefs and attitudes. Analyses of open-ended problems are often presented as free-form text since they require arguments and justifications for one solution over others.

This dissertation investigates how peer review may help students learn from each other, and how it may further inform the instructor. Computer-Supported Peer Review in Education, a technology whose use is growing, has been shown to provide accurate summative assessment of student work, and peer feedback can indeed be helpful to students. A peer review process depends on the rubric that students use to assess and give feedback to each other. However, it is unclear how a rubric should be structured to produce feedback that is helpful to the student and at the same time yields information that could be summarized for the instructor.

The dissertation describes a study in which students in a course on Intellectual Property law wrote individual analyses of an open-ended problem, and then gave feedback to each other using Comrade, a web application for peer review. The study compared two conditions: some students used a rubric that was relevant to legal argument in general (the domain-relevant rubric), while other students used a rubric that was focused on legal arguments about the specific conceptual issues embedded in the open-ended problem (the problem-specific rubric). The study evaluated aspects of validity, reliability and helpfulness of peer feedback, and compared how different hierarchical Bayesian models of peer feedback may help in using peer assessment to inform the instructor.

1.1 Motivation

Learning and assessment are difficult in domains such as ethics and the law that involve analysis of open-ended problems. These problems, sometimes also called ill-defined or wicked problems (Rittel & Webber, 1973; Voss, Post, Chi, Glaser, & Farr, 1988; Aleven, Lynch, Pinkwart, & Ashley, 2009), are usually distinguished by a goal that can be perceived only through analysis and refinement, and by allowing multiple acceptable solution paths. Solvers may frame open-ended problems differently according to their knowledge, beliefs, and attitudes, thereby yielding different representations for the problem in terms of relevant facts and applicable operators. Analyses of open-ended problems are often presented as free-form text since they require arguments and justifications for one solution over others.

Traditionally, instructors have been called on to assess student work and to provide formative feedback. Under the right circumstances, feedback can have a positive impact on learning outcomes with effect sizes as high as 1.10. (Hattie & Timperley, 2007) However, feedback is not readily available for those learning to solve open-ended problems. For practical reasons, instructors cannot give detailed, formative feedback to large numbers of students on a frequent basis. The fields of Intelligent Tutoring Systems (ITS) and Automated Essay Scoring (AES) have excelled in generating feedback for some types of exercises, but these techniques cannot easily assess open-ended problem solving expressed in free-form text.

Some features that make ITS and AES applications so successful are inapplicable to analysis of open-ended problems. Intelligent tutors can be said to have an inner loop, over the steps of a solution to a problem, and an outer loop, over a set of problems on some topic. (VanLehn, 2006) Because the inner loop contains a fine-grained model of a task, e.g., an enumeration of the known correct and incorrect solution paths through a physics problem (VanLehn et al., 2005), a student's actions can be interpreted probabilistically to identify the student's likely solution path, even if the actions are expressed in brief snippets of natural language. (A. Graesser et al., 2000; Aleven, Ogan, Popescu, Torrey, & Koedinger, 2004; Jordan, Makatchev, Pappuswamy, VanLehn, & Albacete, 2006) If the solution path is incorrect, the ITS can provide immediate feedback on student errors. To solve an open-ended problem, a student may start from many different initial states, proceed along different solution paths, and arrive at an end state that is good because it is argued to be so by comparison with other solutions. Accordingly, it can be hard to build a task model for this kind of problem solving, to detect the student's solution path, to evaluate the student's chosen end state and justification for the end state, and to give feedback at various intermediate points.

Without an inner loop, some systems can still assess a student's complete solution and provide feedback on the product, if not the process. In some domains, solutions may be subject to objective constraints. (Mitrovic, Martin, & Mayo, 2002) In textual domains, Automatic Essay Scoring can be used to evaluate features of discourse structure such as whether or not an essay has a thesis statement or whether the body of a paragraph matches its topic sentence. (Burstein, Marcu, & Knight, 2003) AES can also detect a student essay's similarity to standard essays at known levels of quality and to check whether or not a specific topic has been discussed. (Foltz, Gilliam, & Kendall, 2000; Landauer, Laham, Foltz, Shermis, & Burstein, 2003) However, solving some open-ended problems may require feedback with respect to conceptual content in one or more defensible solutions, or from multiple perspectives, or there may be no way to define *a priori* what constitutes an acceptable response. So long as an automated system cannot assess a student's answer, it cannot update its representation of what the student does and does not know, the so-called student model. This precludes it from tutoring the student through guidance, feedback and selection of new problems.

1.2 Approach

To deliver feedback, at a minimum, one must assess a student's work and formulate feedback on the basis of this assessment. Although an instructor or an intelligent tutor conventionally performs both of these tasks, the responsibility for them could be shared by distinct actors, such as the student's peers in the same class. The practice of assessing peer works and exchanging feedback with one's peers is sometimes called peer review.

First, a student's peers constitute an alternative source of formative feedback. Even if student peers are novices by comparison with the instructor, they are likely more knowledgeable than similarly situated students who are not enrolled in the class, and by virtue of shared classroom experience they understand the context for the works that their classmates are producing. Second, the peers are also a source of assessment, which can serve as a basis for feedback or other response by an instructor or an ITS. For example, if an instructor could know that students had mastered some knowledge components but not others, then the instructor could deliver individualized feedback or class-wide instruction that took these levels of mastery into account. Arguably, assessment is required even when it is desirable to give only correct-answer feedback. (Shute, 2008) Nonetheless, because student peer reviewers are novices, scaffolding may be necessary to ensure that they assess each other accurately and provide high quality feedback.

At present, despite a great deal of peer review research, e.g., (Strijbos & Sluijsmans, 2010; I. M. Goldin, Brusilovsky, Schunn, Ashley, & Hsiao, 2010), it is unclear how to share the responsibilities of assessment and feedback formulation with students in a way that encourages accurate assessment and formative feedback. This is the research goal of this dissertation.

A peer review experience that provides students with formative feedback and allows them and their instructor to gauge their performance requires a structured way of eliciting assessments and feedback from peer reviewers. This structure is assembled from several components, namely prompts, criteria, and rubrics, explained below.

1.2.1 Peer Review

Peer review, including computer-supported peer review, has a long history in education research. (Patterson, 1996; Zeller, 2000) Peer review can be administered in many ways and aimed at improving many distinct outcomes. (Topping, 1998; Falchikov & Goldfinch, 2000; Gielen, Peeters, Dochy, Onghena, & Struyven, 2010)

The most basic peer review process is that students produce works, exchange these works, and finally review (i.e., read and assess) each other's works. All other steps may be omitted, depending on the context. For example, in many implementations peers not only review works, they also record their assessments, these assessments may further be delivered to the peer authors, and the authors may be asked to produce a subsequent draft.

Some of the advantages of peer review include the following. Students benefit in that receiving feedback from multiple peers' on the first draft of an assignment can lead them to improve the quality of their second drafts even more than receiving feedback from an expert. (Cho & Cho, 2007) Student authors receive an extra channel of feedback in addition to and distinct from assessment by the instructor or self-assessment (Sluijsmans, 1998), and, in playing both roles of author and reviewer, students may learn from engaging in an authentic activity in the many professional domains that institutionalize peer review.

Peer review may be used to provide formative and summative assessment. (Topping, 2010) 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.

The purpose of feedback is to help assessees understand what performance they should aim for, their level of current performance, and how to improve their performance. (Hattie & Timperley, 2007) To elicit this feedback, we can use prompts. In psychological research, prompts have historically 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, 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). Feedback may be gathered in many forms, including numeric ratings and written comments. For example, ratings on a grounded Likert 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) 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)

1.2.2 Criteria and Rubrics

Criteria are abstract ideals to which students (ought to) aspire, and against which one hopes to assess student performance.[1] 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.

The choice of rubric influences the experience of both reviewers and authors. The first research question of this dissertation is how the choice of rubric can stimulate reviewers to produce formative feedback and accurate assessment. Additionally, it is desirable to

[1] While this explanation is sufficient for the present discussion, note that it collapses the notions of 'criterion' and 'standard'; cf. (Sadler, 1987)

define a rubric 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.

This dissertation examines two approaches to rubric design: optimizing for generality versus optimizing for fit to open-ended problems.

A rubric's generality may be domain-independent, domain-relevant, or problem-specific. A domain-independent rubric refers to criteria that could apply to any domain. For instance, in the SWoRD system, criteria such as insight, logic and flow (Cho & Schunn, 2007) may be used to assess writing in an engineering ethics class as well as in a class on research methods in cognitive psychology.

By comparison, a domain-relevant rubric contextualizes general criteria within a domain, and is is less generally applicable 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)

These differently-scoped criteria 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?"

How a rubric fits analysis of open-ended problems may be influenced by the rubric's underlying philosophy. 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 argument 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 programming such as abstraction, decomposition and encapsulation. (Turner, 2009)

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 criteria. 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 scope and with a focus on either domain-relevant skills or 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) 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 rubric fit.

1.2.3 Artifacts of Peer Review

Peer review yields a rich variety of artifacts that can serve as raw data for characterizing student behaviors in ways that shed light on outcomes of interest. The artifact generation is a secondary benefit of authentic activities of peer such as generating formative feedback and normative assessment, and helping students practice writing, and cognitive and social skills. Artifact generation is amplified by the multi-participant nature of peer review, and by sub-processes such as peer assessment of student-authored works, and peer assessment of the peer reviews themselves. Computer technology facilitates the data collection and enables efficient administration of peer review.

The second research question of this dissertation, elaborated below, is whether these artifacts can be used to derive accurate measures of student performance that may be informative to an instructor or an ITS. Conventionally, tutoring systems assess a student's knowledge by eliciting fine-grained information from the student, and by interpreting this information with reference to a domain model. (VanLehn, 2006) Such fine-grained information is not available for students engaged in analyzing open-ended problems. Nonetheless, because peer review elicits assessments of student knowledge from peer reviewers, if such assessments could be mined and synthesized, they may fit the role of a student model for an intelligent tutoring system. Similarly, educators could find information on students' proficiencies and weaknesses to be helpful, such as for planning future instruction. (In fact, it could be that such information can be mined in sufficient quantity that interpretation may require guidance from intelligent instructor support.) However, it is unclear how to structure rubrics such that they both provide formative feedback to students and information to be synthesized by an automated process, and even well-structured rubrics need to be validated. It is also unclear how the synthesizing process should work, and whether the synthesis of an individual student's assessments could incorporate information from outside of the peer review setting or from patterns of behavior that are generalized from other students.

As a basis for this synthesis, artifacts of peer review may be modeled statistically based on a conception of peer review as a social network or graph with every student as a node. When the student acts as a reviewer, there are outbound edges from this student to the peer authors whose work she is reviewing. When the student acts as an author, there are inbound edges to this student from the other students reviewing her work. Thus, the feedback received by a student is that student's inbound feedback, and the feedback given by a student is that student's outbound feedback. In peer review exercises where authors "back-review" their inbound feedback, the back-reviews received by a reviewer are the reviewer's inbound back-reviews, and the back-reviews given by an author are outbound back-reviews. One way to derive performance measures based on this model is by aggregating inbound feedback (i.e., the feedback received by a single author from multiple reviewers). Inbound ratings of a single work can be averaged for reliable summative assessment. (Cho & Schunn, 2007; Paré & Joordens, 2008) Inbound comments on an analysis of an open-ended problem may suggest different perspectives on the problem and different ways of improving the analysis.

1.3 Research Questions

In the work described here, we compared the differential effects of problem-specific, concept-oriented criteria and domain-relevant, rhetorically oriented assessment criteria. We chose these two because, as explained above, they represent an important trade-off.

This trade-off bears on eliciting accurate assessment and formative feedback in peer review. Since student peers are novices, there ought to be value in prompting them to focus on the conceptual issues that are the subject of class instruction and that are relevant to the open-ended problem under analysis. This should lead to feedback that is grounded in the conceptual issues and is thus more helpful to the authors. Similarly, peer assessment ought to be more accurate if it is focused on appropriate deep features.

This trade-off also bears on exploiting artifacts of peer review to characterize student behaviors. Predictive statistics and machine learning value models that fit the data and that do not overfit the data. Not coincidentally, the search for the right model to characterize student behaviors via artifacts of peer review is analogous to the search for the right rubric to assess student work. A peer review process that is structured in terms of useful abstractions should yield peer review artifacts that are structured in terms of the same abstractions, which may facilitate data mining and model fitting.

Specifically, the dissertation addresses two sets of research questions. First, we considered the effects of supporting reviewers with problem-specific and domain-relevant rubrics. To check the validity of the two types of reviewer support, the ratings elicited by these rubrics were correlated with instructor scores, and this was compared against the correlation of performance on an objective test with instructor scores. We also examined the reliability of both types of ratings. In light of these reliability and validity findings, we inquired whether or not reviewers are responsive to the analytic design of the two rubrics, or if they treat the rubrics holistically, with a special focus on the validity of problem-specific conceptual distinctions given the novelty of that rubric. Finally, the ratings from each rubric were evaluated for helpfulness to peer authors, with checks on whether such helpfulness evaluation is valid and affected by reviewer-author reciprocity. This is covered in chapter 3.0 .

Second, we looked at how data generated in the process of peer review can be modeled and mined to inform an instructor or an ITS. We inquired whether the peer review process contains useful latent information, and we examined how this latent information helps in using peer assessment as a proxy for instructor assessment. An important aspect of this investigation considers whether the additional complexity required for sophisticated modeling to extract latent information is a worthwhile trade-off for the inferences supported by the models. Specifically, we built several hierarchical Bayesian models of peer ratings. These models different in their parametrizations, such as by assuming the mutual independence of the students enrolled in a given course or by relaxing that assumption, and by representing rating dimensions separately or by collapsing them together. We evaluated and compared the models. This is discussed in chapter 4.0 .

Finally, chapter 2.0 reviews related work, and chapter 5.0 concludes by discussing the contributions of this research and directions for future work.

2.0 Related Work

As previewed in the introduction, this dissertation brings together several threads. The chief aim of the thesis is to explore how formative assessment on student analyses of open-ended problems may be provided by student-to-student peer review. The related work discussed below explains what open-ended problems are and how student solutions may be assessed. Since open-ended problems are often analyzed in writing, the related work includes assessment of writing, especially with rubrics; the involvement of students in assessment; the use of computers to support peer review and the role that rubrics can play in that; and why some alternative educational technologies struggle to provide formative assessment for open-ended problems.

2.1 Open-ended Problems

The notion of open-ended problems has been been explored under many names, including open-ended, ill-defined, ill-structured, and wicked, and in many domains, including architecture, law, design, and urban planning. (Lynch, Ashley, Aleven, & Pinkwart, 2006; Buchanan, 1992; Rittel & Webber, 1973) There is no single element that makes problems open-ended. Instead, open-endedness is a continuum, and just how open-ended a problem is depends on an accumulation of characteristics. (Voss et al., 1988) The more of the following characteristics a problem has, the farther away it is from well-defined problems.

> (1) The goal is vaguely stated, and requires analysis and refinement in order to make the particular issue tractable. (2) The constraints of the problem typically are not in the problem statement; instead, the solver needs to retrieve and examine the constraints… (3) Different solvers may vary considerably in the nature and contents of each of the [representation and solution] phases. This is because ill-structured problems may be approached in different ways, according to the solver's knowledge, beliefs, and attitudes. (4) Solutions…usually are regarded in terms of some level of plausibility or acceptability. Further-

> more, solution evaluation may be a function of the evaluator's knowledge and beliefs… (5) When a solution is stated, it usually is justified by verbal argument… (6) [T]o know if [a solution] would ''really work'' would require implementation and subsequent evaluation. When to terminate discussion of the solution is thus somewhat arbitrary. (7) The size of the database required for most ill-structured problems and the difficulties in accessing it make simulation difficult. (Voss, Hitchcock, & Verheij, 2006)

This description of open-ended problems implies that these problems often require framing (e.g., items 2 and 3), that their solutions are often expressed as free-form text (items 5 and 6), and that no deterministic algorithms exist for solving an open-ended problem, only heuristic techniques to guide the analysis (items 1, 6, and 7). (Maner, 2002) These aspects of open-ended problems present a challenge both to the students learning to solve them, and to the instructors who need to assess student learning.

For instance, an engineering ethicist may assign students to analyze a case such as the explosion of the Challenger shuttle (Pinkus, 1997), and to explain what aspects of professional engineering practice may have contributed to the explosion. The case analysis method taught in a popular textbook (Harris et al., 2000) asks students to consider the morally relevant facts of the case, both known and unknown; to structure the analysis via the conceptual issues that can relate the facts to each other; to use their moral imagination to propose and compare alternative resolutions to the dilemma; and finally to justify a particular resolution. Although a whole class of students may follow this procedure in analyzing a particular case, their answers may vary greatly because students will differ in how they frame the case. To frame a case is to state one's perspective on what issues are salient in a case, and which are less important. Framing also invites discussion of how professional codes of ethics bear on the case, and whether previous cases constitute relevant analogies. Framing further necessitates mapping these considerations to the facts of the case at hand.

One area of research into open-ended problems has been expert-novice studies. These have shown that when novices approach a problem, they may recognize both deep and superficial types of features, and they tend not to distinguish the two types. On the contrary, experts look past the surface to focus on a problem's deep features. For example, one way to determine deep features in physics is to ask experts to group problems together and to describe the rationale for the grouping: "If 'deep structure' is defined as the underlying physics law applicable to a problem; then, clearly, this deep structure is the basis by which experts group the problems." (Chi, Feltovich, & Glaser, 1981) Noting deep similarities between the problem at hand and other known problems allows an expert to reuse the known problems' solutions and to make an argument for one solution over another. The ability to map between problems based on their deep features has been used in experiments as a measure of understanding. (Wiley & Voss, 1999)

Continuing the engineering ethics example, the deep features likely translate to what may be called a "conceptual issue." (Harris et al., 2000) In the Challenger case, one conceptual issue was the chain of command in terms of deciding whether or not to launch the shuttle; the morally relevant facts include the individual and organizational actors in the case and their authority relationships. That this is a useful abstraction that can be brought to bear in analyzing this problem is not obvious to a novice. Part of education in engineering ethics is to explain how this abstraction arises in this and other cases. Students' written analyses of a problem can be assessed in terms of whether the abstraction is labeled by its formal name, whether it is defined from first principles, and whether applied to the facts of the problem. (I. Goldin, Pinkus, & Ashley, n.d.) Thus, "relationship between engineer and manager" a basis for an assessment criterion for the Challenger case, and it may naturally be used as a dimension of a problem-specific rubric.

2.2 Assessment of Writing

Assessment may be used for summative or formative purposes, the latter being the focus of this dissertation. Formative assessment may be seen as based on two key components: a student's work needs to be evaluated with respect to a rubric, and this evaluation must be made useful to the student either directly or indirectly (e.g., via an instructor). This interpretation of formative assessment is rooted in recent literature on formative assessment. (Cizek, 2010; Shute, 2008) As summarized by Cizek (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. (Scriven, 1966; Bloom, Hastings, & Madaus, 1971) More recently, in the words of (Cizek, 2010), 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 fits with this dissertation's vision for peer review, and it is more expansive than an alternative definition that only feedback that leads to an improvement in learning outcomes may be termed formative. (Shute, 2008)

Theory on assessment has noted that an instructor wishing to assess student work must often make “qualitative judgments”, which are 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)

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. In particular, they may assess writing to measure students’ analytical skills, knowledge, and understanding (Stiggins, 2005), which are especially relevant to writing in the content disciplines. For the purpose of formative peer assessment, relevant writing assessment techniques include holistic scoring, primary trait scoring, and analytic scoring. (O’Neill, 2009) The scholarship on these techniques is vast, and sometimes uses conflicting language; cf. the definitions of holistic scoring in (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, 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—expressive, persuasive, and explanatory—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 pointed out in (Wolcott & Legg, 1998), while 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; Yancey, 1999; O'Neill, 2009), 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, Scaling Theory, & Methods, 1958), cited in (Lord, Novick, & Birnbaum, 1968), itself 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 do not seem applicable to 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 noted in (Deane & Quinlan, 2010), multiple studies have found that distinct traits of essay quality are highly correlated (T. McNamara, 1990; Lee, Gentile, & Kantor, 2008), and such inter-trait correlation may motivate the use of holistic scoring over analytic scoring.

Thus, despite a variety of theoretical and practical approaches to writing assessment, the problem is far from solved.

2.3 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 arrive at different views of peer assessment. One cause of this may be that "the instructor has access to grades for all papers, whereas the students

[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).

only see grades on their own papers (and perhaps one or two more by social comparisons with friends).” (Cho, Schunn, & Wilson, 2006)

More broadly, however, students may view peer assessment differently from the instructor if the students do 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 teacher's 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-ylanne, Pihlajamaki, & Kotkas, 2006) Students tend to be skeptical of summative peer assessment even if it is accurate. (Draaijer & Boxel, 2006)

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

These findings are also consistent with 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. (Sadler, 1983; Hattie & Timperley, 2007) Criteria-referenced peer assessment is one way that students may receive information on all three of these elements.

> 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. (Sadler, 1989)

In writing in particular, theories of writing and revision (Alamargot & Chanquoy, 2001) also note that authors need to be able to understand the quality of their works in progress and how to improve them. For example, this is implied in the Compare and Diagnose steps of the Compare-Diagnose-Operate procedure. ("Does learning to write have to be so difficult", 1983)

By way of illustration, one instructional technique that may enhance formative assessment is joint student-instructor articulation of assessment criteria. (Stiggins, 2005) This may be employed in peer assessment settings:

> By involving students in the design of instruction and assessment, they become aware of how and on what knowledge and skills they are assessed. Peer assessment can be conceived as an evaluative device, but in our approach it is also a powerful learning activity. [If a test is] kept under lock and key...there is virtually no way that students can "learn by doing" as happens through engaging in a performance-based assessment in which they are involved as one of the assessors. (Sluijsmans & Prins, 2006)

Similarly, in self-assessment, student articulation of criteria is theorized to complement and catalyze self-regulated learning. (H. L. Andrade, 2010)

Instructors both generate and receive 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." (Sadler, 1989) Instructors need to make decisions that affect individual assessees (e.g., whether to recommend additional exercises) as well as large groups of students (e.g., whether to re-teach subject matter that is challenging to all the students in the class).

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.

2.4 Computer-supported Peer Review in Education

Peer review may be implemented in many different ways in service of a variety of outcomes in many instructional settings. (Gielen, Dochy, & Onghena, 2010) Because of such diversity, the nomenclature surrounding peer review is not standardized, which may lead to confusion among both practitioners and researchers. Topping's review (Topping, 1998) addresses literature on "peer assessment, peer marking, peer correction, peer rating, peer feedback, peer review and peer appraisal." As (Armstrong & Paulson, 2008) points out, there are distinctions among peer review, peer response, peer editing, peer evaluation, and peer criticism or critique, all of which may take place in writing-oriented courses. We have also encountered the term "workshopping" (i.e., critiquing each other's work in a writers' workshop). Only some of these practices involve formative assessment. For instance, "peer marking" asks student peers to evaluate each other's work summatively, e.g., for a grade. (Paré & Joordens, 2008)

An early example of software for peer review in education, Peer Grader, was used as early as 2000 to review student-written research papers, to support students in compiling bibliographies relevant to class lectures, to annotate lecture notes, to make up original problems, to review other students' designs, and to do weekly reviews in independent-study courses. (E. Gehringer, 2000) Since then, the research on peer review systems, their use cases, their interfaces, and implications for the learning sciences has increased dramatically, including a recent workshop (I. M. Goldin et al., 2010), a special issue of a journal (Strijbos & Sluijsmans, 2010), and another special issue in progress (Schunn, Ashley, & Goldin, n.d.).

Depending on how broadly one construes peer review, as of this writing, dozens of systems for computer-supported peer review in education have been implemented by educators, researchers, and commercial vendors, and more or less formal evaluations of many have been published in the academic literature. The chief benefit of these systems, aside from any theoretical perspective on peer review, is that they enable peer review in instructional settings where its manual implementation would be practically impossible. This is achieved through automation of key processes, such as collection of student assignments, distribution of these to peers for review, collection of reviews with regard to a rubric, delivery of this structured feedback to peer authors, blinding reviewers and authors for anonymous communication, assignment of reviewers to authors, and back-evaluation of reviews from authors to reviewers. The number of these systems speaks to how easy it is to implement a basic system with high utility. Indeed, before dedicated software was available, general-purpose software, e.g., conferencing systems, were used to conduct peer review exercises. (Cunningham, 1994)

Computer-supported peer review is used in instruction in virtually all academic disciplines, with especially significant research communities in writing, computer science, nursing, and second language education. Some promising research directions include the following. Domain-specific educational practices may be exploited to integrate peer review for maximal pedagogical usefulness; in computer science, students may get feedback on their programs both from peer reviewers and from automated testing tools, and the generation of the assignments themselves may be parametrized to ensure that reviewers and authors have distinct assignments. (Zeller, 2000) Peer assessment, self-assessment, and collaborative assessment may be combined within a single system to enrich the space of instructional activities. (Gouli, Gogoulou, & Grigoriadou, 2008) Criteria-based self-assessment (Li & Kay, 2005) can be used to generate a "scrutable" student model (Weber & Brusilovsky, 2001), i.e., one that a student can examine and modify. Rather than assigning students to review works selected randomly or letting students choose works to review, peer works may be assigned or recommended to individual reviewers based on characteristics of the reviewer, author, and the work itself (Crespo García, Pardo, & Delgado Kloos, 2006; Masters, Madhyastha, & Shakouri, 2008), but the literature on effective group composition is not definitive, cf. (Webb, Nemer, & Zuniga, 2002; Hsiao & Brusilovsky, 2008). Students who peer review papers that score low in terms of peer assessment may produce better second drafts than students who peer review high-scoring papers. (Cho & Cho, 2007) Reviews, whether numeric or textual, may be evaluated automatically with machine learning techniques, which can serve as a basis for formative or summative assessment. (Cho, 2008; Xiong, Litman, & Schunn, 2010; Ramachandran & Gehringer, 2010) Preventing authors from knowing reviewer identity increases the number of critical reviewer comments and improves writing performance on a transfer task. (Lu & Bol, 2007)

The finding that summative peer assessment is very similar to summative assessment by an instructor has been noted multiple times. Combining the opinions of multiple reviewers for each essay provides a more reliable estimate of the quality of the essay than a single reviewer's opinion; for example, if the correlation of reviewer and instructor scores is 0.6, an effective reliability of combined reviewer scores of 0.9 requires about 6 reviewers. (Cho & Schunn, 2007) Reviewers may be evaluated via the numeric ratings they produce, e.g., in terms of metrics such as systematic difference, consistency, and spread. (Cho & Schunn, 2007) These evaluations may be factored into a summative peer assessment of an author's work as a differential weight on the scores given by the reviewers, and these evaluations may be computed at the same time as the quality of the peer author works under review. (Hamer, Ma, & Kwong, 2005; Lauw, Lim, & Wang, 2007) Evaluations of reviewers may also be used to grade reviewer effort (E. Gehringer, 2000), and communicated to the reviewers to help them monitor and improve their performance, privately or to the whole class as public praise of good performance. (E. F. Gehringer, Gummadi, Kadanjoth, & Andrés, 2010) By calibrating reviewers, it is possible to ensure a minimum reviewer accuracy before reviewing begins. (Russell, Cunningham, & George, 2004) Summative computer-supported peer assessment that was substituted for instructor assessment violated a contract with a labor union representing teaching assistants; the lawsuit was settled by the University of Toronto. (Sequeira, 2010)

As with any assessment technique, validity and reliability are key issues of interest with regard to peer review. If validity of summative peer assessment is defined as the convergence of peer assessment to instructor assessment, the instructor may have a different view of the validity of an exercise than the individual student. This is because the instructor's impression of validity is a kind of average that takes into account all the papers in the class, while an individual student author's impression depends on whether the peer ratings received by that author deviate from the instructor's grade. (Cho & Schunn, 2007) The general tension between validity and reliability has been noted in peer assessment: peer review may demonstrate a "convergence of different raters on a 'single truth"', or it may "uncover the presence of multiple perspectives about the performance being assessed, which do not necessarily have to agree." (Miller, 2003)

2.5 Rubrics in Peer Review

Rubrics are often used within peer review to support formative feedback and 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)

One early paper (Kwok & Ma, 1999) describes a web-based Group Support System that was developed to support articulation and application of criteria by instructor and students. There are several notable aspects. Peer assessment was performed within groups of 20 students, and each group chose its own set of assessment criteria. Criteria included 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 criteria. 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. 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 due to the software or to the criteria that the students choose.

Another study (Lin, Liu, & Yuan, 2001) compared two types of rubrics in a writing exercise in a computer science class: students in the holistic feedback condition "gave a total score and offered a general feedback for an entire assignment," while reviewers in the specific feedback condition used 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. Interestingly, there was an interaction of rubric type with the students' aptitude for following directions: 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.

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 does not. (Chalk & Adeboye, 2005) The correlation of summative tutor and peer assessments, although statistically significant, was low, $r = 0.27, df = 62, p < 0.05$. This comparison is complicated by the fact that the assessment instrument with specific criteria lacked free-form commenting, while the holistic instrument required it.

Rubrics may contain checklists of typical errors (Sanders & Thomas, 2007), but such rubrics may not fit with assessment of open-ended problems.

None of the studies above examined rubrics that were problem-specific or concept-oriented. A rubric investigated in an introductory computer science course contained three conceptually oriented dimensions (abstraction, decomposition, and encapsulation, which are concepts of object-oriented software design); two additional dimensions were functionality and style, which are general notions of computer programming. (Turner, 2009) One finding in that study 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 showed an improvement in understanding of decomposition during the intervention, but not on the posttest.

In sum, although there is theoretical justification for the use of rubrics and a great deal of research on rubric use in writing assessment, especially for summative purposes such as placement (section 2.2), the research base on application of different rubrics in formative peer assessment is small and inconclusive.

2.6 Educational Technology for Open-Ended Problems

Educational technology has been developed to provide feedback to students engaged in some kinds of open-ended tasks, including some that involve textual and diagrammatic input from students.

It is possible to design an ITS to deal with free text student essays. Examples include Select-a-Kibitzer (A. C. Graesser & Wiemer-Hastings, 2000), Summary Street (Steinhart, 2001), AutoTutor (A. Graesser et al., 2000), Apex (Lemaire & Dessus, 2001), e-Rater (Burstein et al., 2003), and Criterion (Higgins, Burstein, Marcu, & Gentile, 2004). These systems not only evaluate student essays on the fly, but they also provide feedback and encourage students to correct and rewrite their essays and resubmit them for new feedback. So far, however, tutoring systems for essay writing can detect only fairly general features. For instance, Criterion and e-Rater learn to detect the 'discourse segments' like thesis, main idea, supporting idea, and conclusion, but they focus on essays that are meant to have a rigid structure (an introduction, three supporting paragraphs, and a conclusion), and seem to be about 300 words long. Alternatively, systems like the Intelligent Essay Assessor (Landauer, Laham, & Foltz, 2003) that can address term-paper length essays detect general features, like coverage or absence of broad topics based on comparisons to past graded “anchor” papers.

In dealing with open-ended problems, however, it would be desirable for an ITS to provide formative feedback on the features one expects to see in an analysis of such problems, including an author’s refinement of the problem-solving goal in order to define an issue, an elaboration of constraints that the author has inferred an adequate solution needs to satisfy, arguments in favor of the proposed solution and consideration of plausible rebuttals and counterarguments. (Voss et al., 2006) At a finer grain size, one might hope to detect statements of fact, reasons, possible outcomes, comparisons of alternatives, and conclusions. (Voss, Greene, Post, & Penner, 1983) We may even wish for tutors that can discuss whether the student has framed the problem appropriately, and with rigor.

In the educational setting, text mining techniques at the level of propositions (rather than essays) have generally been used to test student awareness of particular facts, and possibly to compare this awareness against a representation of an ideal cognitive state for a problem’s solution path. (Jordan et al., 2006; Litman & Purandare, 2008; Popescu, Aleven, & Koedinger, 2005) Given the need for natural-language engineering that targets not only statements of fact, but additional features such as refinements, definitions, framing, arguments, and justifications, this is a poor fit to analyses of open-ended. These existing techniques have also used as input not essays, but dialogue or dialogue-like utterances. If a system can expect that a student’s input at a given time is likely to correspond to a specific question posed by the system, it has additional information beyond the content of the textual proposition to attempt to classify its meaning.

To collect yet more features for training classifiers, some systems use “sentence openers”, i.e., require a student to begin input by pressing a button that enters canned sentence-initial text. For example, dedicated buttons may enter text such as “I think that…” or “I agree, because…”. Sentence openers were originally developed for chat-oriented systems. (McManus & Aiken, 1995; Baker & Lund, 1997; Robertson, Good, & Pain, 1998) Such buttons not only encourage constructive dialogue among students, they may also hint at the speaker’s dialogue strategy. (Soller, 2004)

Text classification approaches face challenges in recognizing fine-grained features in open-ended problem-solving. Training and evaluation corpora in education have historically been very small, while recent algorithm development in machine learning and information retrieval has tended to Web-scale tasks. Small corpora necessitate system-building that depends on manual development of coding schemes, manual labeling of training data, and sophisticated but task-specific algorithm design that may not easily transfer to other systems. A related challenge to development of automatically trained classifiers is to go beyond optimizing accuracy and to generate transparent explanations of predictions that can be incorporated into feedback to students.

Since it is difficult for an ITS to detect content features in a reliable, automated manner, designers of ITSs for open-ended problems have explored non-textual methods for gathering information about student answers to ill-structured problems. For instance, in CSAV, law students use computer-supported Toulmin-based diagrams to represent their own developing legal arguments (Carr, 2003); in Belvedere, students construct Toulmin-style diagrams of their own scientific arguments (Suthers, Hundhausen, Dillenbourg, Eurelings, & Hakkarainen, 2001). More recently, in the LARGO program, students read transcripts of oral arguments concerning ill-structured legal problems and diagram them in terms of an underlying argument model that the designers seek to teach. (Pinkwart, Lynch, Ashley, & Aleven, 2008; Ashley, Pinkwart, Lynch, & Aleven, 2007) Programs such as Belvedere and LARGO can analyze the diagrams and provide some guidance and feedback. For instance, LARGO identifies areas of the argument text where students' have missed important information to diagram; it also highlights and gives hints on related elements of the diagram that are worth reflecting about in light of the underlying model of argumentation.

While this kind of feedback can help students understand a general model of analyzing open-ended problems, it cannot guide a student's analysis of the specific issues that an open-ended problem presents. At the same time, model-tracing ITSs for (comparatively) well-structured problems, such as Andes (VanLehn et al., 2005) or the ACT Programming Tutor (Corbett & Anderson, 1995), can make inferences about the student's mental model and prior knowledge, map student actions to solution paths, evaluate the quality of a solution, and provide hints in the case of errors. Open-ended problems impede these techniques because student actions are often represented as free-form writing, and because student actions can (and often must) redefine the problems such that modeling problems and solutions a priori is impossible. Constraint-based tutoring systems also face difficulties with essay-length free-form text. For example, while an algebraic equation tutor might contain constraints such as 'either the left-hand side or the right-hand side of the equation contains a constant' (Suraweera, Mitrovic, & Martin, 2005), an ITS for legal argument would be hard-pressed to evaluate constraints such as 'cites only relevant precedents' or 'considers the perspectives of all stakeholders', which depend on how the problem is framed. Thus, there is a critical gap between ITSs for open-ended and well-defined problems.

While formative assessment for open-ended problems is not solved in the general case, Intelligent Tutoring Systems, Automated Essay Scoring, and other educational technology can clearly be brought to bear on some aspects of this problem. Future research will likely pursue further integration between these technologies and Computer-Supported Peer Review.

3.0 Eliciting Formative Feedback via Conceptually Focused Rubrics in Peer Review

3.1 Motivation

This chapter describes an experiment that compared formative assessment in peer review via conceptual, problem-specific support for reviewers and authors versus domain-relevant support.[1] The following chapter focuses on how these methods of supporting reviewers and authors may be used to inform the instructor.

Given the variety of possible strategies that can be employed in supporting peer reviewers, and given that reviewer support influences the experience of both reviewers and authors, it is important to determine whether some kinds of reviewer support are more valuable than others. For example, it is desirable to prompt reviewers in ways that lead to helpful, formative feedback and to accurate assessment. It is also desirable to avoid prompting reviewers in ways that yield redundant information. When an instructor gives a student a rubric to assess another's paper, interacting with this rubric can cause the student to focus on those issues that are made prominent in the rubric. For example, if the rubric looks at domain-independent issues of writing composition, that communicates to the reviewer that the instructor sees various discourse features as important.

The work described here compares the effects of two analytic rubrics for evaluating 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 concep-

[1] Some results from this chapter have been reported in:

Goldin, I. M., & Ashley, K. D. (2010). Eliciting informative feedback in peer review: importance of problem-specific scaffolding. In V. Aleven, J. Kay, & J. Mostow (Eds.), 10th International Conference on Intelligent Tutoring Systems. Pittsburgh, PA.

Goldin, I. M., & Ashley, K. D. (under review). Eliciting formative feedback via conceptually focused rubrics in peer review. (C. D. Schunn, K. D. Ashley, & I. M. Goldin, Eds.) Journal of Writing Research, Special Issue: Redesigning Peer Review Interactions Using Computer Tools.

tual issues under analysis. As described earlier, domain-relevant rubrics and problem-specific rubrics represent a trade-off. A domain-relevant rubric can be used more broadly than a problem-specific one, but a problem-specific one may lead to more helpful and more accurate formative assessment.

The fact that a domain-relevant rubric can be used broadly means that it is more likely to be validated. Since 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.

Specifically, this chapter addresses the following research questions considering the effects of supporting reviewers with problem-specific and domain-relevant rubrics. To check the validity of the two types of reviewer support, the ratings elicited by these rubrics were correlated with instructor scores, and this was compared against the correlation of performance on an objective test with instructor scores. We also examined the reliability of both types of ratings. In light of these reliability and validity findings, we inquired whether or not reviewers are responsive to the analytic design of the two rubrics, or if they treat the rubrics holistically, with a special focus on the validity of problem-specific conceptual distinctions given the novelty of that rubric. Finally, the feedback from each rubric was evaluated for helpfulness to peer authors, with checks on whether such helpfulness evaluation is valid and affected by reviewer-author reciprocity.

3.1.1 Hypotheses

Regarding the different rubrics, domain-relevant versus problem-specific, we evaluated hypotheses concerning validity and reliability of the peer assessment process, reviewer responsiveness, and feedback helpfulness.

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.

From the perspective of assessment, student performance in a written work may be viewed as a proxy measure for student understanding of the subject, and other proxy measures are also possible. In particular, as discussed in (O'Neill, Moore, & Huot, 2009), objective assessment has historically been endorsed as a more reliable measure than essay assessment (Godshalk et al., 1966), even if it does not address issues of validity. In addition to having their essays reviewed by their peers, students took an objective test that addressed their understanding of relevant domain issues, and their objective test scores were also correlated against the summative instructor scores. Thus, to examine the issue of assessment from different perspectives, three different constructs (peer ratings, instructor scores, and test performance) were used to measure student understanding of relevant concepts, and these three were expected to converge.

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, researchers do not agree on what is formative (section 2.2), 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, since 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 validate ratings of helpfulness, and to take into account author-reviewer reciprocity, as explained below.

Validity of feedback helpfulness. It was expected that those reviewers who understand the domain well would be able to to provide more helpful feedback than those reviewers who understand the domain poorly. This is a gross measure of convergent validity of feedback back-review ratings, and it was expected to apply to reviewers using either rubric. If a reviewer gives helpful feedback with regard to some problem-specific concept, that suggests that the reviewer understands the concept sufficiently to assess a peer author's work, and to suggest ways of improving it with regard to that concept. Insofar as problem-specific inbound peer ratings and inbound back-review ratings both reflect conceptual understanding, i.e., share a common cause, they will be related to each other. Similarly, a reviewer's understanding of domain-relevant dimensions will be reflected in the quality of the feedback given with the domain-relevant rubric.

Author-reviewer reciprocity. Peer reviewers and peer authors may at times engage in "tit-for-tat" reciprocal behavior. (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.

3.2 Methods

3.2.1 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 (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."

3.2.2 Apparatus

The study was conducted via Comrade, a web-based application for peer review. For purposes of this study, 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 chose a nickname directly in Comrade, and was identified to other students only by that nickname. 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. (E. Z. Liu, 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 answered 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. This objective test of was also administered online via Comrade. After step 4, all students were invited to fill out an optional survey.

3.2.3 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 and asked students "to provide advice concerning [a particular party's] rights and liabilities given the above developments." The instructor designed the facts of the problem to raise issues involving many of the legal claims and concepts (e.g., trade secret law, shop rights to inventions, right of publicity, passing off) that were discussed in the first part of the course. Each claim involved different legal interests and requirements and presented a different framework for viewing the problem. 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. Since 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.

This type of essay assessment is typical of American law school examinations. It approximates 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. Performance assessment *per se* (Stiggins, 2005) is usually unfeasible in general law school courses except in clinical courses involving small numbers of students in representing actual clients under the close supervision of instructors. Real-world cases are only rarely used as problem scenarios for assessment, since real-world cases may not present a pedagogically ideal collection of legal issues and factual circumstances, but students are encouraged to cite relevant real-world cases from the course casebook in their answers.

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.

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 Likert scale so that the peer authors could see how their peers evaluate their current level of performance, the same Likert 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 Likert scales had 7 points, grounded at 1,3,5,7 (Figure 3.2 .1, Figure 3.2 .2).

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 3.2 .3)

Based roughly on the legal claims, concepts, and issues addressed in the exam question, the instructor designed a multiple choice test in two equivalent forms (A and B), each with 15 questions. (Figure 3.2 .4) The test was intended to assess whether student reviewers understood the legal concepts. The questions presented legal claims and concepts related to the exam, but not in the same way as the exam, involving new hypothetical situations with completely different facts, and in a multiple choice format rather than in essay form. The test was designed to address conceptual understanding rather than shallow knowledge. As seen in the sample question, the hypothetical fact situations required students to pick out legally salient facts, which they could only do by framing the scenario via the relevant claims and concepts, thus making these questions open-ended like the exam question. After preparing the tests, the instructor invited several particularly strong students who had taken the same course in prior years to take the test. The instructor then revised the test based on these students' answers to multiple choice questions and other feedback.

In a posthoc analysis to validate the problem-specific rubric, a trained rater used the

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 3.2 .1: Domain-relevant rating prompts. Reviewers rated peer work on four criteria pertaining to legal writing.

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 Jack's idea for the I-phone-based instrument-controller interface ("idea1")

Barry v. Smith for misappropriating Barry's 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 3.2 .2: Problem-specific rating prompts. Reviewers rated peer work on five problem-specific writing criteria (the claims), which all used the same scale.

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 3.2 .3: Back-review Likert scale, grounded at 1, 3, 5, 7.

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

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

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

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

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

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

Figure 3.2 .4: Sample questions from the two forms of the objective test. Salient differences are emphasized with italics and correct answers are bold; these were not emphasized in presentation to the students.

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 ensure similarity to the instructor's scoring method, 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 the papers using the problem-specific criteria and the answer key.

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, back-reviews, performance on the objective test and survey responses 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 students 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.

3.2.4 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.

3.3 Results

The following presentation of results addresses the effect of two different rubrics on validity and reliability of peer assessment and assessment via objective test, 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 Likert scale (Table 3.3 .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 (see chapter 4.0) were similar to the classical point estimates, except in the case of the “right of publicity” dimension (labeled in Table 3.3 .1 as $\sigma^2_{n[IPR]}$).

3.3.1 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. Additionally, scores from an objective, multiple-choice test that students took after writing their essays but before reviewing each other’s work were also compared against instructor scores.

3.3.1.1 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 *

Table 3.3 .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)	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)	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) ($\sigma^2_{n[IPR]} = 0.96$)	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)

4 reviewers = 16 inbound ratings. (The instructor only gave summative, not give 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 $\alpha = 0.05$, 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 student author. Thus, each paper receiving feedback via the problem-specific rubric 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 were significantly correlated to the ratings of the trained rater. (Table 3.3 .1) The lone exception was the second idea misappropriation claim for which peer rating reliability was particularly low. (See subsection 3.3.2)

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)

3.3.1.2 Objective Test Validity: Relationship of Objective Test Performance to Instructor Scores

Student comprehension of domain knowledge was measured after the students wrote their exam answers, concentrating roughly on the same legal claims, concepts, and issues that were addressed in the exam question. This "pretest" was administered via two test forms: half of the students in each condition took test form A, and half took test form B. The Pearson correlations of the quantity of questions answered correctly by students with the instructor's score of the students' papers were not significant for either test form, $r(28) = 0.00, p = 0.99$, and $r(26) = 0.20, p = 0.31$. Assessment via this instructor-designed objective test did not converge on assessment via instructor scores of midterm exam essays.

The test difficulty was reliable across two administrations, the first before students gave feedback to each other (pretest), and the second after they gave feedback but before they received feedback (posttest). A difficulty score for each question in each test form was computed as a tally of how often the question was answered correctly by the students. The Pearson correlation of the difficulty scores across two administrations was reasonably strong and statistically significant for both test form A, *r*(13) = 0.71, 95% CI [0.30, 0.89], *p* = 0.003, and test form B, *r*(13) = 0.78, 95% CI [0.45, 0.92], *p* < 0.001. Each test was reliable across administrations to different samples of students enrolled in the same course.

The tests were not internally consistent. Point-biserial correlations between performance on individual test items and performance on all other test items were low and not statistically significant for almost all questions on all four test forms. Some item-test correlations were negative, but omitting items with negative correlations to the test did not improve other point-biserial correlations. The test forms were also evaluated using McDonald's ω_t, which is a stronger lower bound estimate of internal consistency than other measures, including Cronbach's α. (Zinbarg, Revelle, Yovel, & Li, 2005; Revelle & Zinbarg, 2008) Unlike α, ω_t allows for loadings on multiple factors, which is appropriate for an objective test that addresses multiple distinct concepts. By this measure, internal consistency is guaranteed to be restricted to the range $[\omega_t, 1]$. Across the two test forms and two test administrations, ω_t values were 0.20 and 0.40 (pretests), and 0.30 and 0.39 (posttests), i.e., one cannot be confident that each test form was internally consistent. Test-retest reliability (computed assuming that the intervention of using one of the two rubrics had no differential effect) was absent, $r(54) = -0.03, p = 0.80$.

Although the tests were designed to be equivalent, they were apparently dissimilar in terms of student performance. First, at the level of individual questions, the Pearson correlation of the difficulty scores of the two test forms within each administration was not significant either for the pretest, $r(13) = 0.15, p = 0.59$, or for the posttest, $r(13) = 0.11, p < 0.70$. Second, at the level of the tests in the aggregate, the number of questions answered correctly by a student on the pretest was not correlated with the number answered correctly on the posttest, $r(52) = -0.06, p = 0.64$.

In sum, in this experiment, assessment via objective tests was problematic. On the one hand, the tests have strong face validity and pedigree. On the other, test scores did not converge on instructor assessment of midterm exams, and the two test forms were neither internally consistent nor consistent with each other.

3.3.2 Peer Assessment Reliability

Ratings produced via both domain-relevant and problem-specific rubrics were evaluated for reliability. Following (Cho, Schunn, & Wilson, 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 (Cho, Schunn, & Wilson, 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 3.3 .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 convincing conclusions.

The problem-specific rubric elicited ratings at both high and low ends of the Likert scale that was used for all dimensions. (Table 3.3 .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. Since reviewers provide important information when they give low rating to an author's work, it is encouraging that low ratings do not cause low reliability among reviewers.

3.3.3 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.

3.3.3.1 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. (Table 3.3 .1) 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 and 10 pairwise correlations for the problem-specific rubric. (Table 3.3 .2) Correlations between mean inbound ratings in the domain-relevant condi-

Table 3.3 .2: Pairwise correlations between mean inbound peer ratings among dimensions of each rubric. For problem-specific dimensions, correlations among the ratings of author work by a trained rater, in parentheses. Asterisk indicates correlations significantly different from zero at $\alpha = 0.05$.

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

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

tion 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. The extent to which each rubric represented a unitary construct, i.e., internal consistency, was measured using McDonald’s ω_t over per-dimension mean inbound peer ratings. (See subsubsection 3.3.1.2.) For the domain-relevant rubric, $\omega_t = 0.94$, and for the problem-specific rubric, $\omega_t = 0.75$. 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. Further, these were second and third year law students, who must be familiar with legal argumentation, that is, with the domain-relevant criteria. Nonetheless, since they were novices in the subject matter of Intellectual Property, the following section investigates whether they missed important relationships among the problem-specific criteria, which would have led to the low inter-dimension correlations.

Second, 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, 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.

Third, 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.

3.3.3.2 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 3.3 .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.

3.3.4 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 3.2 .3)

3.3.4.1 Validity of Helpfulness Ratings

It was hypothesized that those reviewers who understand the domain well, as measured by essay quality, would be able to to provide more helpful feedback than those reviewers who understand the domain poorly. Such a relationship between essay quality and feedback helpfulness would attest to convergent validity of feedback helpfulness.

Within each rating dimension, correlations were computed between each student's mean per-dimension inbound back-review score (i.e., mean helpfulness score from having acted as a reviewer) and mean per-dimension inbound peer rating. Correlation values were all close to zero, and no correlation was significant. Thus, giving helpful feedback with respect to a rubric dimension was not related to analytical or writing performance on that dimension. This does not make the helpfulness ratings invalid, but it does point to the complexity of the relationship between understanding an aspect of the domain, problem-specific or otherwise domain-relevant, and being able to provide helpful feedback about that aspect.

3.3.4.2 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. Given 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 3.3 .1) [2] Thus, there is a small but statistically significant amount

[2] In prior work, reciprocity was defined as the Pearson correlation (Cho & Kim, 2007), which produces similar results for the problem-specific and domain-relevant ratings.

of reviewer-author reciprocity using both rubrics. Contrary to expectations, problem-specific criteria did not make it easier for authors to evaluate feedback objectively.

3.3.4.3 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 3.3 .1) However, an ANOVA comparing all problem-specific back-review ratings versus all domain-relevant back-review ratings showed that authors receiving domain-relevant feedback found it to be helpful more often than authors receiving problem-specific feedback, F(1,853) = 36.82; $p < 0.001$.

Domain-relevant feedback was rated 6 or 7 more often than problem-specific rubric. As defined by the rating scale (Figure 3.2 .3), 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”. In other words, authors felt that the domain-relevant feedback contained praise more often.

Additionally, problem-specific feedback was rated 3 or below more often than domain-relevant feedback. To understand why problem-specific feedback was sometimes 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.

3.4 Discussion

This experiment compared formative assessment in peer review via conceptual, problem-specific support for reviewers and authors versus domain-relevant support. The results showed that both kinds of reviewing rubrics led to valid peer assessment of student work, which compared favorably to assessment of conceptual understanding via an objective test. Examining the rubrics’ analytic dimensions separately showed some differences, but

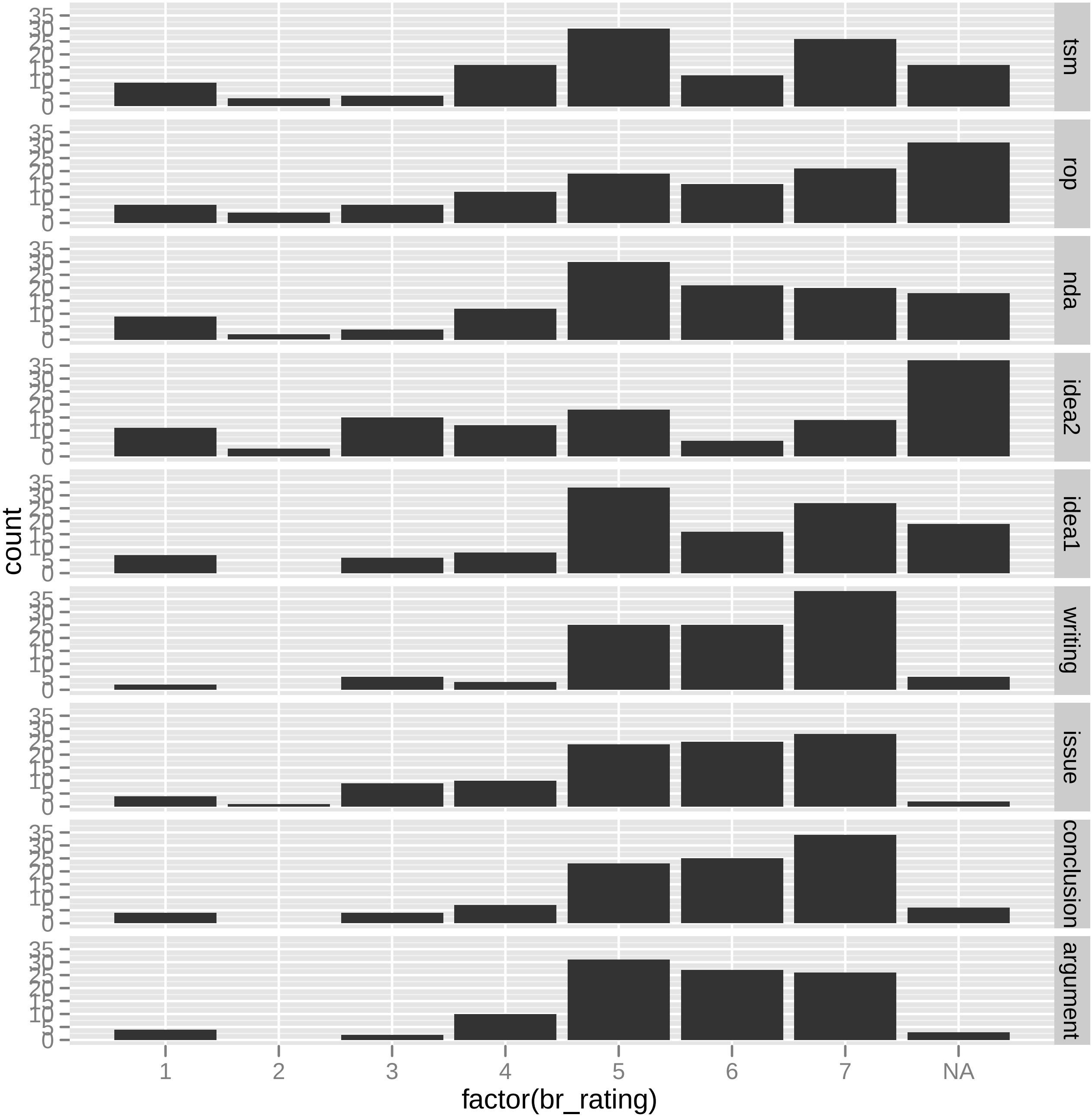

Figure 3.3 .5: Frequency of back-review ratings from dimensions of the domain-relevant and problem-specific rubrics.

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. Both rubrics usually elicited helpful feedback. Peer authors judged feedback elicited by the problem-specific rubric as helpful less often than feedback elicited by the domain-relevant rubric, but in part that was due to greater amounts of praise elicited by the domain-relevant 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. This difficulty is only emphasized by the internal inconsistency of the objective tests of conceptual understanding and by the lack of correlation between the tests and the instructor scores. Although some validity and reliability evidence points in favor of the problem-specific rubric (i.e., the strength of the validity correlation, the narrowness of the correlation confidence interval, the larger number of dimensions with high effective reliability), 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.

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. Since 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 related to legal claims, each of which provides a separate analytical framework for the open-ended problem. Further, since open-ended problems may be framed in multiple ways, reviewer disagreements with respect to conceptual issues could be legitimate. The different problem-specific reviews may thus lead authors to see the problem in different ways, and the exercise of reading and making sense of the somewhat divergent information may be pedagogically fruitful. (Wiley & Voss, 1999; D. McNamara, Kintsch, Songer, & Kintsch, 1996; Cho & Cho, 2007)

Reliability may be important in some cases, especially summative assessment: "[F]or students to take the feedback seriously, the ratings need to count for actual grades, and the validity and reliability of the grades depends upon there being ratings from multiple reviewers." (Cho & Schunn, 2007) If so, reliability may be improved by increasing the number of peer reviewers (Cho & Schunn, 2007), and by calibrating their rating techniques (Russell et al., 2004). It may be easier to teach students how to apply problem-specific criteria (e.g., what factors may support a claim of trade-secret misappropriation, what factors may constitute a good response to such a claim), rather than domain-relevant criteria (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.

While authors rated problem-specific feedback as helpful significantly less often than domain-relevant feedback, this may be an artifact of how the peer review process was designed and how helpfulness was measured than how helpful the feedback was actually. 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 (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 work, a fair and rich evaluation of helpfulness would entail delivering both problem-specific and domain-relevant feedback to each author to see which the authors prefer when they can draw on either.

Back-review ratings for both types of rubrics are 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 (Cho & Kim, 2007), this may be undesirable. The ratings may communicate formative information 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.

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 conceptual issues. 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, there is value in teasing apart problem-specific conceptual issues.

The two rubrics evaluated here share 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 include problem-specific conceptual content in the feedback they give. However, any problem-specific conceptual information would be distributed across the domain-relevant dimensions, attenuating the conceptual signal, and leading to interference from multiple concepts and non-conceptual 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.

These results have significance for the design, implementation, and evaluation of peer review as a mechanism for teaching skills for analysis of open-ended problems. They further support the design of several statistical models that may inform instructors about the state of a peer review exercise and establish connections between peer review and intelligent tutoring systems, as discussed in the following chapter.

4.0 Hierarchical Bayesian Models of Peer Review

> "Relevant evidence" means evidence having any tendency to make the existence of any fact that is of consequence to the determination of the action more probable or less probable than it would be without the evidence.
> (US Federal Rule of Evidence 401)

4.1 Motivation

The preceding chapter (chapter 3.0) examined the differential effects of problem-specific and domain-relevant support on peer ratings in peer review. The present chapter builds on those findings to model the assessment of student performance based on these rating criteria.[1][2]

The ratings received by peer authors, i.e., the inbound peer ratings that assess the authors' works, are but one artifact of peer review. As discussed below, additional observable and latent information that is a by-product of peer review could be relevant to assessment. If this information could be mined, it could be delivered to an instructor or an Intelligent Tutoring System.

First, the chapter discusses sources of observable and latent information in peer review, and how these may be incorporated into a statistical model.

[1] Some results from this chapter have been reported in Goldin, I. M., & Ashley, K. D. (to be presented). Peering into peer review with Bayesian models. In S. Bull, G. Biswas, & J. Kay (Eds.), 15th International Conference on Artificial Intelligence in Education. Auckland, New Zealand.

[2] This research was supported in part by the University of Pittsburgh Provost's Advisory Council on Instructional Excellence. The grant was "A Peer-Review-Based Student Model for Ill-Defined Problem-solving", Kevin D. Ashley, PI.

Second, several alternative statistical models are proposed that differ in their representation of the domain of peer review. The models use an expert's scores of the students' essays as the response variable; they differ in the explanatory variables that they use and in the hierarchical structure. In essence, the models aim to approximate the instructor measure of student performance by using the latent and observable artifacts of peer review. Using a regression methodology permits inferences regarding which explanatory variables are important predictors of the essay score, and how they could be combined to apprise the instructor of the outcome of a peer review exercise. The statistical modeling technique is Bayesian data analysis. (Gelman & Hill, 2006)

Third, the models are compared empirically to understand whether the additional complexity required for sophisticated modeling is a worthwhile trade-off for the inferences that the models support. The data used for the comparison come from the experiment on problem-specific and domain-relevant rubrics in chapter 3.0 . As will be seen shortly, the data from the two rubrics is treated separately, which enables a further comparison of the rubrics.

Finally, the discussion considers the lessons learned from this modeling and explains how a Bayesian model may be used by an instructor or an Intelligent Tutoring System.

4.2 Methods

4.2.1 Hierarchical Bayesian Models of Peer Review

The defining characteristic of Bayesian models is that they can incorporate prior beliefs about the parameters; for example, aggregate peer ratings may be said to be normally distributed. By combining prior beliefs with data and with formulations of likelihood, a Bayesian model yields posterior estimates for the parameters of interest and describes each estimate in terms of a probability distribution rather than just a point value.

While Bayesian modeling has long been applied in educational research, this is a novel contribution to peer review in education. From the perspective of statistical analysis, peer review is fairly complex. It involves repeated measures (multiple reviews of every paper), sparse data (any student reviews only a handful of papers), and hierarchy (ratings are generated according to a multidimensional rubric). By using Bayesian data analysis, we can enter these relationships among the data into our model in a straightforward way, and we can compare different models based on our intuitions about model structure. Furthermore, a single Bayesian computation estimates all the quantities of interest at once, bringing to bear all the available data. This means that the different parameters help estimate each other according to the expression of likelihood we enter.

Given two models that fit the data equally well, one may choose the simpler one (e.g., complex models can be prone to overfitting) or the more complex one (e.g., it may embody knowledge about domain structure). Within each condition, models may be compared in terms of Deviance Information Criterion (DIC), a metric that rewards well-fitting models, and penalizes models for complexity. Model fit is defined as deviance, similar to generalized linear models. Model complexity is determined by the effective number of parameters in the model. This is computed at model "run time" as a function of how information is pooled across groups in a multilevel model, rather than at "compile time" from the mathematical model structure. Lower DIC is better. DIC values may be compared on one dataset but not across datasets.

The models use an expert's scores of the students' essays as the response variable; they differ in the explanatory variables that they use and in the hierarchical structure. The peer ratings and the explanatory variables are, respectively, the observed and latent artifacts of peer review.

The baseline model 5.1a uses the simplest representation that could be said to map from peer ratings to an instructor's score, essentially averaging all the ratings received by a peer author. This model treats all of a pupil's inbound peer ratings as exchangeable with each other, ignoring the distinct rating dimensions, and it treats all authors as independent of each other.

Subsequent models aim to leverage latent information to improve data fit. First, in model 5.1b, student authors are no longer considered to be independent. Model 5.1b pools key parameters so that what is known about authors as a group helps us understand individual authors, and vice versa. Second, model 5.2a evaluates the utility of the ratings dimensions by representing them separately rather than collapsing them together, as in the models above.

For comparison and reference, the parameters of each model are summarized in Table 4.2 .1 and Table 4.2 .2. Each model is defined in the following sections. Models were implemented using the WinBUGS software (Lunn, Thomas, Best, & Spiegelhalter, 2000); the source code of our models is included in Appendix B.

Model were run separately for the two experimental treatments, because it would not be sensible to compute the contribution of problem-specific information for students in the domain-relevant condition and vice versa. Each model was fit 3 times, i.e., there were

Table 4.2 .1: The latent artifacts that were represented in various models. NP: no pooling, PP: partial pooling, CP: complete pooling

Model	Description	α_p, α	β_1	β_n	γ_n	X_{1p}	X_{np}	Z_{np}	$\sigma^2_{p[IPR]}$	$\sigma^2_{n[IPR]}$	σ^2
5.1a	no pooling regression: individual intercept, variance per pupil	NP	NP			NP			NP		CP
5.1b	individual intercepts pooled across pupils; individual means of inbound peer rating pooled across pupils; variance per pupil	PP	NP			PP			NP		CP
5.2a	individual intercepts pooled across pupils; individual means of inbound peer rating per-dimension across pupils; variances per dimension	PP		NP			PP			NP	CP

Table 4.2 .2: Interpretation of different parameters

Parameter	Definition	Interpretation
α_p, α	intercept, per-pupil or pooled across all pupils	the expected value of a pupil's midterm score when other predictors for this pupil are equal to zero
β_1, β_n	coefficient for mean inbound peer ratings	the weight given to the aggregate assessment of the work of all authors in estimating the instructor score of the authors' midterm exams, collapsing across rating dimensions (β_1) or for a particular dimension (β_n)
X_{1p}, X_{np}	mean inbound peer rating, per-pupil	aggregate assessment of the work of author p according to peer reviewers, collapsing across rating dimensions (X_{1p}) or for with respect to a particular dimension (X_{np})
$\sigma^2_{p[IPR]}$	variance of inbound peer ratings, per-pupil	variation among the assessments of the work of author p according to peer reviewers
σ^2	variance between pupils	variation among pupils not otherwise captured by the model
$\mu_{[IPR]}$, σ^2_{IPR}	mean, variance across authors of aggregate inbound peer ratings	pooling hyperparameters to share information across students w.r.t. authors' aggregate inbound peer ratings X_{1p}
$\mu_{n[IPR]}$	mean across authors of aggregate inbound peer ratings, per-dimension	pooling hyperparameter to share information across students w.r.t. authors' per-dimension aggregate inbound peer ratings X_{np}
$\sigma^2_{n[IPR]}$	variance of inbound peer ratings, per-dimension	variation among the assessments of all authors' works with respect to rating dimension n according to peer reviewers

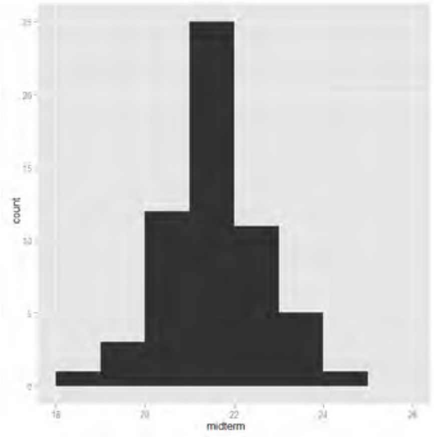

Figure 4.2 .1: Histogram of midterm scores

3 Markov Chain Monte Carlo simulations (MCMC chains), with a different randomly determined starting value for each chain. The chains were examined to ensure that they converged in their estimates of parameter values. Each fit was allowed 6000 iterations, with 1000 initial iterations discarded ("burn-in") to avoid bias due to starting values. For ease of interpretation, peer ratings were centered by subtracting the mean of all ratings from each rating. Thus, a student that is estimated to have a mean inbound peer rating of -1.25 actually has a rating that is 1.25 Likert scale units less than the mean of the actual ratings in the dataset.

4.2.2 Model 5.1a: Inbound Peer Ratings with No Pooling and Uninformative Priors

Model 5.1a is a regression of the midterm scores as a function of the pupils' inbound peer ratings only. Authors are treated as randomly drawn from a single population. The multiple ratings that each author receives are exchangeable with each other (i.e., not tied to particular reviewers), and constitute repeated measures of each student given according to a single rubric dimension.

Midterm scores and ratings are treated as normally distributed. It is reasonable to treat midterm as normally distributed, because test scores in general ought to be normal, and because midterm scores in this particular dataset seem to behave normally. (Figure 4.2 .1) It is also reasonable to treat mean inbound peer ratings as normally distributed, because the mechanism according to which they are generated—the 7-point Likert scale—ought to be designed such that ratings are approximately normal. Moreover, the mean of repeated ratings (i.e., from multiple reviewers) ought to be normal according to the Central Limit Theorem.

The inbound peer ratings are said to be normally distributed and sufficiently described by a per-pupil mean and variance. In a Bayesian model, rather than computing a mean and variance of each pupil's inbound ratings manually and using their point estimates, the means and variances are estimated "on the fly" during MCMC sampling. This yields not just point estimates, but posterior observations with accompanying credible intervals that indicate the model's certainty in the parameter estimate.

The model estimates a regression coefficient β_1 that corresponds to a weight on the aggregate peer rating X_{1p}, and an intercept α_p that varies per pupil, accommodating variability within students. These are combined linearly in a hierarchical structure defined by a per-student latent factor μ_p, which represents overall latent knowledge or ability of an individual student in the domain.

Formally, we define the model as follows. The per-pupil response variable (midterm) Y_p is normally distributed, with mean μ_p that is the per-pupil knowledge estimate and overall variance estimate σ^2 .

$$Y_p \sim N(\mu_p, \sigma^2)$$

The model includes a per-pupil intercept (i.e., coefficient for the constant term) α_p, and the mean of the inbound peer ratings X_{1p} with weight β_1. The means of ratings are denoted as X_{1p} rather than X_p since other models will add other variables indexed $2..n$ to the matrix of predictors X. Similarly, regression coefficients are denoted $\beta_{1..n}$ with a per-predictor (not per-pupil) subscript, and the regression procedure estimates a vector $\boldsymbol{\beta}$ of regression coefficients.

$$\mu_p = \alpha_p + \beta_1 * X_{1p}$$

Finally, a pupil's inbound peer ratings are treated as normally distributed according to the pupil's individual mean X_{1p} and individual ratings variance $\sigma^2_{p[IPR]}$.

$$IPR_p \sim N(X_{1p}, \sigma^2_{p[IPR]})$$

The individual means X_{1p} are treated as normally distributed with an "uninformative" prior. That is, we express our prior belief about *parameters* such as the per-pupil means of the peer ratings by saying that they are themselves distributed according to distributions defined by *hyperparameters*. If our prior beliefs are not strong, then we make them "uninformative" to the model (i.e., not leading to undue bias), such as by allowing the hyperparameters to vary a great deal. In practice, the hyperparameter mean for X_{1p} was set to 0, as is appropriate for an uninformative prior since the peer ratings were centered about their mean, and the hyperparameter variance was set to 1000.

In hierarchical modeling (Gelman & Hill, 2006), this model is called a "no pooling" regression based on the fact that it does not pool (i.e., share) information across pupils. Specifically, each pupil is described via an individual intercept α_p, an overall (between pupils) variance σ^2, and an individual mean peer rating. Alternative, as discussed below, it is possible to consider that students could be grouped together in some way, and what is learned about one student could help model a different student.

Model 5.1a is a plausible first attempt to establish if the ratings that peers give each other approximate instructor assessment. It asks if the cumulative opinion of the reviewers (i.e., the mean inbound peer rating) corresponds to an instructor's grade, and if the peer reviewers tend to agree (i.e., measuring the ratings' variance). Additionally, it incorporates prior distributions for the response, the ratings, the mean and variance parameters. Whether or not this baseline differs from the alternative models, its evaluation should still be helpful in understanding peer review.

4.2.3 Model 5.1b: Information Pooling

In model 5.1b, information is shared across pupils so that what the model learns about one student informs the estimation of parameters of other students. This is accomplished in two ways.

First, the model stipulates that all individual intercepts α_p are not independent, but drawn from a common distribution. This distribution is defined by its own hyperparameters, mean μ_α and variance σ_α^2. Each student's information is then used to estimate these common hyperparameters, and the common distribution they define in turn constrains the estimation of the individual students' intercepts.

$$\alpha_p \sim N(\mu_\alpha, \sigma_\alpha^2)$$

Second, in a similar fashion, the estimation of individual students' inbound peer rating means X_{1p} is constrained by stipulating that pooled together, they share a common mean and variance, i.e., the mean of all individual means $\mu_{[IPR]}$, and the variance of all individual means $\sigma^2_{[IPR]}$.

$$X_{1p} \sim N(\mu_{[IPR]}, \sigma^2_{[IPR]})$$

This technique is known as partial pooling. (Gelman & Hill, 2006) It can be contrasted with no pooling, as in model 5.1a, and with complete pooling, which stipulates that while a parameter may be informative or important, it does not vary for individual students.

All hyperparameters were given uninformative prior distributions.

There is pooling for individual students' inbound peer rating variances $\sigma^2_{p[IPR]}$. Models with such pooling were unstable, especially for the domain-relevant dataset.

4.2.4 Model 5.2a: Rating Dimensions

Model 5.2a represents the distinct dimensions of the inbound peer ratings, in contrast to models 5.1a and 5.1b, which treat all inbound peer ratings as though they correspond to one rating dimension, no matter that they were elicited via different prompting questions.

To incorporate information on dimensions, each observed inbound peer rating is modeled as normally distributed with mean X_{np} that corresponds to the average of the ratings received by author p for rating dimension n and a variance for this dimension $\sigma^2_{n[IPR]}$ that is shared across all pupils. The explanatory variable matrix X is altered to include one column per rating dimension, and the lone regression coefficient β is replaced by regression coefficients β_n for each rating dimension n. (There are $d = 4$ rating dimensions in the domain-relevant condition, and $d = 5$ in the problem-specific condition.) Within each dimension, individual pupils' means of inbound peer ratings X_{np} are pooled by stipulating a shared prior distribution across students. These are uninformative priors, normal with a mean of 0 and a variance of 1000. Peer ratings from each dimension are centered about the respective mean of that dimension.

$$Y_p \sim N(\mu_p, \sigma^2)$$

$$\mu_p = \alpha_p + \sum_{n=1}^{d} \beta_n * X_{np}$$

$$IPR_{np} \sim N(X_{np}, \sigma^2_{n[IPR]})$$

Table 4.3 .3: Model fit (DIC) for domain-relevant and problem specific datasets

Model	Domain-relevant	Problem-specific
5.1a	1416	2423
5.1b	1305	2237
5.2a	1335	1702

Distinguishing the ratings by dimension leads to fewer observed ratings per pupil, per dimension. For example, rather than 20 peer ratings per pupil in the problem-specific condition (5 rating dimensions times 4 reviewers), there are ratings from 4 reviewers per dimension. This precludes estimation of individual per-dimension variances; instead, per-dimension variance parameters were fitted to absorb some of the noise. These variances are denoted $\sigma^2_{n[IPR]}$, with subscript n signifying the rating dimension. Per-dimension variances are treated as independent of each other since rating dimensions ought to be independent.

4.3 Results

The model fitting was examined by verifying that in almost all cases multiple chains of MCMC sampling mixed well, and arrived at similar estimates of the parameters of interest as attested by the Gelman “Rhat” $\hat{R}$ values below 1.2, for all parameters in all models. Visual examination of autocorrelation plots showed that autocorrelation for all models was acceptable, indicating that the MCMC simulation was sufficiently long. Model 5.2a produced the best fit for the problem-specific dataset using a short simulation (6,000 iterations discarding 1,000 as burn-in). Even though the short simulation converged, a longer simulation (100,000 iterations discarding 20,000 as burn-in) was run to ensure that the results were stable. Despite the improvement in overall model fit only to the problem-specific dataset, model 5.2a converged in its estimates of parameters for both datasets.

Model fit was measured via DIC, which trades off fit accuracy against model complexity. Lower DIC indicates better fit. DIC values may be compared with reference to the same dataset but not across the problem-specific and domain-relevant datasets. There is no way of knowing whether a fit is good in an absolute sense, only whether it is better or worse relative to another model. In practice, DIC scores over multiple runs of these models varied within the range of ±15 points over the same dataset, but with different randomly determined initial values.

There are three key findings based on the DIC (Table 4.3 .3). First, partial pooling (model 5.1b) can improve significantly on the baseline (model 5.1a) for both domain-relevant and problem-specific datasets. Second, distinguishing the different rating criteria (model 5.2a) improves the fit for the problem-specific dataset, but actually hurts the fit for the domain-relevant dataset.

As a sanity check that the models reflect what is known of the data, the hyperparameter μ_{IPR} (the mean of all individual means X_{1p}) was estimated as ~0 under model 5.1b for both datasets (Figure C.14). This makes sense, since the ratings were centered about the mean.

Given the high number of parameters in each model (e.g., 28 students * 5 rating dimensions for inbound peer rating means X_{np} in model 5.2a on the problem-specific dataset, not including other parameters in 5.2a on the same dataset), it is not practical to describe each parameter. The discussion below addresses the chief parameters of interest, and suggests they could be interpreted by an instructor or an intelligent tutoring system.

The posterior estimates of the intercepts α_p were estimated to be close to the mean of the instructor-assigned midterm scores. (Figure C.4, Figure C.9, Figure C.21) With ratings centered, the intercept represents the predicted midterm score for a student whose inbound peer ratings averaged to zero, taking into account individual author characteristics. This parameter helps bridge the gap between the response and the the covariates (means of inbound peer ratings). In doing so, it becomes a reference point for understanding the performance of individual students. The models predict that a student's midterm score would be equal to α_p for a student whose means of inbound peer ratings were zero across all dimensions (assuming the ratings were centered). For students whose mean inbound peer ratings are non-zero, their contribution is computed relative to α_p.

Another way of interpreting the intercept is that relatively large absolute values of α_p belong to peer works for which the other terms (i.e., the peer ratings) are not an adequate assessment. As noted, the α_p were estimated to be close to the mean of the instructor-assigned midterm scores. If the α_p were centered about their mean, after the Bayesian estimation procedure, then those centered α_p that were relatively close to zero would provide a relatively small counterweight to the peer ratings term, while those that were relatively far from zero, in either the positive or negative direction, would provide a large counterweight. Thus, they would indicate peer works for which the peer ratings required a counterweight.

Pooling in model 5.1b made intercept estimates an order of magnitude tighter than in model 5.1a, e.g., for the problem-specific dataset, to ±0.8 points with 95% confidence on

Table 4.3 .4: Model 5.2a estimates for β_n.

Dimension (Domain-Relevant)	argument	conclusion	issue	writing
Mean (SD)	1.32*	-2.20	0.43	-0.66

Dimension (Problem-Specific)	idea1	idea2	nda	rop	tsm
Mean (SD)	1.31	0.79*	0.02	0.31*	0.43

the instructor's scoring scale. Pooling for X_{1p} also allowed model 5.1b to share information across students, but the effect was less pronounced given that 5.1a already had tight intervals on these parameters. Partial pooling can be an effective technique for these models.

The regression coefficients β represent the importance of the averaged inbound peer ratings (per-dimension or collapsing dimensions) to estimating the instructor's score. Given the intercept α_p, the model's prediction of an individual student's score depends on the means of the inbound peer ratings weighted by β_n. The coefficients are per unit of the underlying Likert scale (despite centering), so a 1-point change in the mean inbound peer rating for some dimension adds β to the prediction of the midterm score. In the ideal case, the β_n all have the same sign, and their values may be intuitively interpreted as indicating that criteria differ in their impact on approximating instructor scores.

Under model 5.1a, the credible intervals for β included zero (Figure C.2), implying that peer ratings were not significant predictors of instructor scores for that model. With model 5.1b, estimates of β with 95% confidence did show that average peer ratings predicted instructor scores (Figure C.8), emphasizing the value of pooling.

By distinguishing the rating dimensions, model 5.2a improved fit for data from the problem-specific criteria, but not for data from the domain-relevant criteria. This can be seen from the overall fit. (Table 4.3 .3) As the credible intervals show (Table 4.3 .4,Figure C.20), the model is confident in its estimates for two of the four domain-relevant β coefficients (argument development and issue identification, but not conclusion justification or organization quality), and only argument development contributed to estimating the instructor's midterm score. For the problem-specific rubric, four of the five problem-specific β coefficients were estimated confidently (all except the first idea infringement claim);

two of the five dimensions (the second idea infringement claim and right of publicity) contributed to estimating the midterm score, and a third, trade-secret misappropriation, contributed marginally. The problem-specific β estimates are all positive, suggesting that each dimension adds linearly to the intercept. Counterintuitively, the domain-relevant β estimate for justifying a conclusion has a negative sign, as if high performance on justifying a conclusion corresponds to a drop in the midterm score, but note that this β was not estimated confidently. These problems with low confidence and a combination of positive and negative signs for domain-relevant criteria echo the high pairwise correlation between X_{np} for all 6 pairs of rubric dimensions that was reported earlier (Table 3.3 .2). High collinearity may cause instability and interactions among β_n coefficients for the domain-relevant rating criteria (without hurting overall model fit). By contrast, the β_n for the problem-specific ratings may be intuitively interpreted as indicating that criteria differ in their impact on approximating instructor scores.

All β_n, including those that were not significantly different from zero, reflect important patterns in the data and could be informative to an instructor or a tutoring system. Coefficients that were not estimated confidently may indicate a lot of noise in how reviewers applied the corresponding criteria, e.g., the first idea infringement claim. Coefficients that were estimated confidently yet were found to be not significant contributors to estimating the instructor's score (e.g., the claim of breach of non-disclosure) may be empirically redundant with other criteria. Indeed, these two problematic dimensions of the problem-specific rubric are exactly those that correlated significantly with the trade-secret misappropriation claim. (Table 3.3 .2) Knowing that some rubric dimensions are problematic may suggest to the instructor that these dimensions need to be omitted or redefined.

The means of inbound peer ratings for each pupil, X_{np} represent the pupil's proficiency with respect to each assessment criterion n. (Figure C.16) Under model 5.2a, these are posterior estimates with Bayesian credible intervals for the same parameters that are conventionally computed as point values reported to students and instructors in existing peer review systems. Since ratings were centered with respect to each dimension's mean, the value of X_{np} reveals where the competence of pupil p with respect to assessment criterion n falls relative to his or her peers, and it is not affected by skew (i.e., by whether the criterion itself was particularly easy or challenging across all students). The better fit of model 5.1b on the domain-relevant dataset relative to model 5.2a indicates that computing these quantities may be problematic for rubrics whose rating dimensions are not well distinguished from each other.

Per-pupil variances $\sigma^2_{p[IPR]}$ in model m5.1b were estimated somewhat less confidently on problem-specific ratings than on domain-relevant ratings, which makes sense given that for the problem-specific dataset, this model collapsed ratings from demonstrably distinct dimensions.

Per-dimension standard deviations $\sigma_{n[IPR]}$ in model 5.2a ranged approximately between 0.96 for right of publicity to 1.43 for the second idea infringement claim among the problem-specific dimensions. Standard deviations for domain-relevant dimensions were slightly lower, from 0.84 to 1.10. This makes sense given that reviewers using the problem-specific rubric explored used a greater range of the Likert scales.

Insofar as regression models may be used for prediction, the overall standard deviation σ estimates prediction accuracy. Given the inbound peer ratings, the instructor's score can be predicted to within approximately $\pm\sigma$. For problem-specific support, model 5.2a estimated $\sigma = 0.25$, 95% CI [0, 0.66], and for domain-relevant support, $\sigma = 0.47$, 95% CI [0.01, 1.14]. (But note that this is an estimate of error during training, and it would be expected to increase on held-out data.)

4.4 Discussion

This study showed the feasibility of building Bayesian models that describe many different aspects of peer review. The best-fitting model on the problem-specific dataset was 5.2a, which incorporated expert beliefs regarding the distributions of peer ratings, instituted pooling across pupils and distinguished rubric dimensions. The best-fitting model on domain-relevant data was 5.1b, which incorporated expert beliefs regarding the distributions of peer ratings and instituted pooling, but did not distinguish rubric dimensions.

These results may bear on pedagogy in many different ways. The models' parameter estimates may provide a tutoring system or an instructor with actionable assessment information on individual pupils, the whole class, and the assessment rubric itself. Some may even suggest changes in curriculum or assessment. The models report some parameters of interest that have not been described previously, and also enable Bayesian estimates of other parameters.

For instance, an instructor may wish to know which specific concepts were challenging for the authors and which were easy. Challenging concepts could be deserving of instructor attention in grading and instruction. One way to detect challenging concepts is by examining estimates of X_{np}, which pertain to a student's proficiency with regard to each of a rubric's criteria. If criteria can be compared "apples to apples", and if they all use the same well-anchored rating scale (as in the problem-specific rubric studied here), and if the raw ratings have been re-centered about the mean before being entered into the model, then it is meaningful to look at whether distributions of X_{np} skew right for some criteria, which would suggest that these criteria correspond to challenging concepts.

The parameter estimates may further show if the rating dimensions are mutually independent or noisy. Pairwise correlations between dimensions, a metric that is already popular in the literature (e.g., Table 3.3 .2), would benefit from being computed over the Bayesian estimates of X_{np} rather than point values. Inconsistent signs of the β_n may indicate dimensions that produce noisy ratings or have some other problem. On the contrary, consistent signs among the β_n show how the criteria differ in their impact on approximating instructor scores.

Another cue to the instructor that it may be necessary to revise assessment criteria may come from $\sigma^2_{n[IPR]}$, the per-dimension estimates of variation of mean inbound peer ratings. At the ideal level of variance among students, the scale is appropriately grounded and there is sufficient variation among students to explore most of the rating scale. If there is too little variance, then it could be that the Likert scale for this rating dimension does not distinguish between the pertinent levels of author competence, or the scale is not sufficiently grounded, or the students do not actually vary and the criterion is not informative. If there is very high variance, then it could be that the Likert scale is insufficient to describe the actual differences among students. It is also appropriate to compare $\sigma^2_{n[IPR]}$ across dimensions and to ensure that they are not too dissimilar. All these estimates accommodate missing peer ratings because they "borrow strength" from other students' ratings through pooling, and because they are posterior distributions with intervals that speak to the estimates' credibility.

To identify struggling students, an instructor should examine the student's means of inbound peer ratings X_{np} and the individual intercept α_p. The X_{np} constitute the aggregate judgment of the peer reviewers, while the α_p shows the student's individual performance baseline, which is, in effect, information about the quality of the author's work not captured by the peer reviewers.

At a glance, the relative importance of the dimensions are available via β_n: the larger the β_n, the more impact it has on estimating the instructor's score. As a further guide, the instructor should first check the credible intervals for these estimates. A wide interval indicates that the value of β_n is not believable (e.g., it could fluctuate greatly on a different model run). An interval that is narrow but nonetheless includes zero, especially if the point estimate itself is close to zero, indicates a rating dimension that has relatively little impact on the instructor's score. It is important to understand why this would be the case; for example, it could point to a dimension that contains information redundant with other dimensions. Finally, if the credible interval is narrow and does not contain zero, the magnitude of β_n is the weight given to the rating dimension. For example, if α_p is 20, if β_n is 0.25 and if student p's mean inbound peer rating X_{np} is 2, then the model predicts that the instructor will give the student's work a score of $20 + 0.25 * 2 = 20.5$, all other parameters held constant.

To summarize, depending on the rubric dimensions and their rating scales, Bayesian models can inform instructors about the following questions: Are the rating dimensions noisy, appropriately grounded, mutually independent? What is the relative importance of the criteria? Which concepts were challenging? For which peer works are peer ratings not an accurate assessment? Which students are struggling with what concepts?

In some cases, peer assessment is an important perspective on a student's work in its own right; in others, its relevance may depend on how well it approximates assessment by an instructor or other expert. Either way, consumers of peer assessment information, whether instructors or tutoring systems, require precise estimates of the key parameters in peer assessment. They also need to know whether or not the estimates are credible. The Bayesian models described here fill that role.

The old software developers' adage "garbage in, garbage out" applies to peer assessment criteria. Criteria are not all equally useful, clear, or functional. The models and the results they report are only as good as the criteria. The good news is that peer review provides a built-in facility for evaluating the criteria, which can help instructors to refine them and to communicate them to pupils.

5.0 Conclusions

5.1 Summary

It has long been argued that formative assessment is appropriate to analysis of open-ended problems (Sadler, 1983), and that peer review can be a conduit for formative assessment (Sadler, 1989; N. F. Liu & Carless, 2006). This dissertation aimed to examine whether peer review could be used to address the central question of formative assessment for open-ended problems: how to provide formative feedback to students and useful assessment information to the instructor at the same time. This was accomplished through a design for the peer review process that elicited peer feedback in a particular way, and then by mining that feedback. The feedback was evaluated in multiple ways, and the mining of the feedback showed that hierarchical Bayesian models can produce informative summaries of the peer review process.

The dissertation makes several contributions. First, given that the goal of providing formative assessment for open-ended problems imposes constraints on peer review, it argues that these constraints may be satisfied by a particular approach to the design of rubrics. These constraints were that a rubric had to inform students of their current level of performance and of the performance target, it had to provide suggestions for how to reach the target, and it had to generate data that could be summarized for the benefit of the instructor. These constraints led to creating a rubric with the following characteristics:

- An analytic rubric was preferred to a holistic one.
- The rubric had to collect both ratings and comments for each dimension.
- The rating scales had to be anchored to explanations of distinct levels of performance.

This constitutes a theoretically justified approach to rubric design.

Second, the dissertation introduced a novel distinction among three types of rubrics, domain-independent, domain-relevant and problem-specific, each of which could be made to fit the constraints above. This distinction was necessary, because the characteristics of a rubric derived above were insufficient to guide the creation of a rubric. *A priori*, all three of these could be relevant to assessment of open-ended problems.

Third, the dissertation further focused on two of the three types of rubrics: one rubric that emphasized problem-specific details and another rubric that was relevant to many different problems in the domain. To fit these rubrics to the problem and domain, they were oriented, respectively, towards problem concepts and towards rhetorical technique. The evaluation of rubrics took place in a real classroom for empirical perspective and ecological validity, and addressed feedback validity, reliability, reviewer responsiveness to analytic aspects, and helpfulness to student authors. The evaluation showed that rubric-based assessment of student essays converged to instructor assessment, while assessment by multiple-choice test did not.

Finally, the dissertation proposed and evaluated several novel Bayesian models of peer review vis-à-vis feedback elicited via the two types of rubrics. Introducing hierarchical Bayesian models to peer review in education is a contribution in itself. The proposed models are a natural fit for formative peer assessment given that each paper is evaluated by only a few reviewers, which motivates pooling, and given that analytic rubrics should elicit ratings that are distinct across dimensions, which motivates a linearly additive relationship. The models produce information about students as individuals and as a group, and about the assessment rubric itself.

Thus, the research shows how peer review may be guided via different kinds of rubrics and how instructors may receive information about the status of a peer review exercise.

5.2 Limitations, Implications, and Lessons Learned

The data collection and research methods in any study necessarily limit the inferences that one can draw. So it is with this dissertation. This section acknowledges some of the salient limitations.

The aim of the dissertation was to investigate formative peer assessment. As has been argued above, the peer assessment procedure was constructed to yield formative feedback and assessment, e.g., via the rubric design with criteria-connected dimensions, anchored rating scales, and the elicitation of both peer ratings and textual comments. Regretfully, due to time constraints, the findings reported here were based only on peer ratings, rather than on both ratings and comments. For example, peer comments could have been examined to see if they were truly matched to rubric dimensions, if they discussed problems as well as suggested solutions, if there were differences on this between the two rubrics, and if the back-review ratings evaluated the feedback fairly. Further, comments could have revealed to what extent reviewers using the domain-relevant rubric addressed problem-specific aspects, and *vice versa*. Nonetheless, the peer ratings were studied because of their joint formative and summative role: although ratings are numeric, these were anchored to a meaningful scale and not used to rank students relative to each other, i.e., they were criterion-referenced, not norm-referenced.

Was peer assessment formative? The procedure was intended to help students see their strengths and weaknesses, which enabled them to address the weaknesses (although the limitation of only one draft precluded an investigation of the effects of feedback on revision). Additionally, the students reported the feedback to be helpful.

Further, the hierarchical models that map between instructor assessment and peer ratings constitute an empirical argument that peer assessment was, in fact, formative. Learning is not linear. Even if students enrolled together in some course may be said to have similar learning goals for the course (which is questionable), they will differ in background knowledge, aptitude and motivation, and thus they will differ in what skills and concepts they find challenging, in the errors they make while solving problems, and in the mental model repairs that they must undergo to learn. This creates the need for individualized, formative feedback, which in turn requires assessment of student performance.

Traditional instructor scoring of student work, due to real-world constraints, is usually only a summative assessment technique, and has weaknesses where formative assessment is concerned. Even if it references key criteria, it is burdensome for an instructor to include feedback that is structured such that the criteria are explicit, and that includes detailed suggestions for how the student could improve the quality of the work. (Indeed, this is one motivation for peer assessment.) As has been shown in prior work, quantitative peer assessment can correlate to instructor assessment, e.g., by averaging peer ratings. A contribution of the present work is that it shows how peer assessment via (potentially multiple alternative) rubrics, modeled with statistical structures beyond just averaging, can be used to estimate instructor assessment for individual students. Just as a prism can break up light into constituent colors, a rubric can decompose student performance according to key criteria, and these individualized estimates of performance within each dimension provide a basis both for a summative score such as the instructor's and for formative feedback. In a sense, it could be said that the instructor's assessment was validated against peer assessment, rather than the other way around, because the overall instructor score can now be viewed as arising on the basis of rubric dimensions. Furthermore, the Bayesian models' parameter estimates may help the instructor offer students additional feedback and adjust instructional strategies.

As mentioned above, an important contribution of the dissertation is an approach to rubric design that is rooted in theory and validated through practice. This contribution targets a stated need for clarity in rubric creation. (Turner, 2009)

Peer reviewers do pay attention to rubrics: the fact that reviewers distinguished among the dimensions of the problem-specific rubric shows that it is reasonable to use rubric design to influence the feedback that students give to each other. Even though the domain-relevant and problem-specific rubrics shared many similarities by design, empirical evaluation revealed that they have differences as well. One interpretation is that both rubrics focused on argumentation (although this need not be the case for other domain-relevant and problem-specific rubrics), but they approached argument structure in different ways. A domain-relevant rubric is akin to asking a student: "first, discuss all the evidence, then, separately all the warrants, and then, separately once again, all the claims." A problem-specific rubric emphasizes that a particular claim is what ties together the evidence and warrants. In doing so, the problem-specific rubric highlights the conceptual structure of a problem and maps from the conceptual structure to the argument structure. By contrast, when a domain-relevant rubric omits the conceptual structure or relegates it to a single rubric dimension, it risks hiding important aspects of evaluation criteria from reviewers (and authors) and obfuscates the grounding of the argument structure. Thus, a reviewer using the domain-relevant rubric must not only rediscover the conceptual structure, and not only map from the argument under review to the conceptual structure, but also, finally, to break apart the argument structure to align it with the domain-relevant criteria. In hindsight, it seems intuitive that the increase in rubric generality from problem-specific to domain-relevant rubrics (and beyond, to domain-independent rubrics) should correspond to a decrease in a rubric's explanatory power.

Some instructors may prefer domain-relevant rubrics to problem-specific ones because domain-relevant rubrics may not necessitate adjustments to different problems, whereas problem-specific rubrics necessarily do. However, the *concept-oriented* problem-specific rubric provides a path to making these adjustments: one need only enumerate the salient distinct concepts that pertain to the problem, and explain what it means to analyze each concept. The concepts will likely be similar in complexity and in analytical requirements, which means that they will likely share the same rating scale. At the same time, using a domain-relevant rubric does not relieve the instructor of having to be mindful about whether that rubric fits the assignment at hand.

The study only compared one domain-relevant rubric versus one problem-specific rubric, in the context of a single exercise. Thus, findings relating to the differences between the two rubrics cannot be conclusive. In other exercises and in other domains, other domain-relevant and problem-specific rubrics could be devised, and they could affect peer assessment in ways that this study did not detect. Ultimately, the domain-relevant versus problem-specific distinction is just one dimension in the space of rubric design, and it remains to investigate other considerations. For example, rubrics may focus on assessment of procedural knowledge versus assessment of declarative knowledge.

Last but not least, the study holds a lesson for in-classroom assessment. Instructors (as they will readily acknowledge) are fallible. Given that it is possible to automatically evaluate some kinds of assessment (e.g., some aspects of validity, reliability, internal consistency, and inter-dimension relationships, as demonstrated here), instructors should be aided by software that enables such evaluation. This will likely enhance transparency of assessment, and help instructors to refine their professional practice.

5.3 Future Work

This study was largely exploratory. At the outset, it was unclear whether there would be any distinctions between the domain-relevant and problem-specific rubrics. The findings suggest areas of future research in the learning sciences, in open-ended problems, and in educational technology.

Of interest for the learning sciences, it remains to evaluate effects of domain-relevant and problem-specific rubrics on reviewers and on authors who receive this feedback, including outcomes such as quality of feedback comments, change in second draft quality, and retention of problem-specific understanding and domain-relevant knowledge. For instance, problem-specific, concept-oriented rubrics may make it easier for reviewers to recognize a problem's concept structure. If easier recognition reduces the burden of cognitive processing, that may increase the coverage of concepts in feedback.

Another issue is the helpfulness of feedback. In this experiment, problem-specific feedback was rated helpful less often than domain-relevant feedback, which was partly due to the fact that domain-relevant feedback contained praise more often. No student received both types of feedback, however, which would be important for a full comparison of helpfulness.

These questions may be very challenging, and they need to be addressed in multiple domains for generalizability. One of the challenges is that although the distinction between problem-specific content knowledge and argumentation skills can be appreciated intuitively, it can be difficult to define operationally for purposes of measurement. Consider a student author's work that has a weak argument about a claim of trade secret misappropriation. The cause for this weakness could be that the student does not understand some aspects of trade secret law, or it could be that the student has not learned how to make legal arguments. These distinct problems call for appropriate remediation strategies, such as helping the student repair his mental model of trade secret law, or helping the student see weaknesses in the rhetorical structure of the argument. It complicates the distinction further that content knowledge and argumentation skills may be strongly correlated.

With regard to open-ended problems, it would be interesting to see if this research will generalize to domains where students assess peer-produced works that do not involve writing. For instance, while argument is clearly central to the law, its importance has also been recognized in domains such as design. (Buchanan, 2001) A new design or re-design is itself considered an argument, even if this argument is presented not as a written text, but, say, as a tangible object.

In educational technology, issues of interest include robustness and utility of the Bayesian models. This dissertation evaluated the Bayesian models on two datasets, where each author's work was reviewed by four peers. While the models themselves may be used with any number of reviewers per author, it remains to be seen if parameter estimates are robust under fewer reviewers. It is also possible to build models incorporating additional explanatory variables, e.g., reviewer tendency to give ratings that are too high or too low (Cho & Schunn, 2007), and to refine model structure and relax assumptions, e.g., by using logistic and loglinear methods to represent Likert ratings (Muthukumarana, 2010). Finally, the models may need to be adapted to other peer review exercises, such as those where students write two drafts rather than one, or to outcomes of interest other than instructor assessment.

The characterization of peer review via Bayesian models may be informative not only to instructors, but also to an intelligent tutoring system. The models transform peer ratings into information about students that may be used in a student model (at individual and class levels), and in a domain model, such as the relative difficulty of rubric dimensions.

For example, a tutoring system informed by the Bayesian models may be able to support students in reflecting and self-regulating their learning better than suggestive feedback from peers alone. (Boom, Paas, & Merriënboer, 2007)

While much work remains, it is hoped that this dissertation may be useful for formative assessment of open-ended problems, and that it should serve as a starting point for future research.

Appendix A

Intellectual Property Midterm Exam Question

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

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

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

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

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

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

Appendix B

Bayesian Model Source Code in BUGS

2.1 Model 5.1a

```
model {
  for (p in 1:P) {
    midterm[p] ~ dnorm(pupil.mu[p], pupil.tau)
    pupil.mu[p] <- pupil.alpha.p[p] + beta.ipr*ipr.mu.p[p]

    pupil.alpha.p[p] ~ dnorm(0, 0.001)
    ipr.mu.p[p] ~ dnorm(0, 0.001)
    ipr.sigma.p[p] ~ dunif(0, 100)
    ipr.tau.p[p] <- pow(ipr.sigma.p[p], -2)
  }

  for (r in 1:IPR) {
    ipr[r] ~ dnorm(ipr.mu.p[ipr_p[r]], ipr.tau.p[ipr_p[r]])
  }

  pupil.sigma ~ dunif(0, 100)
  pupil.tau <- pow(pupil.sigma, -2)

  beta.ipr ~ dnorm(0, 0.001)
}
```

2.2 Model 5.1b

```
model {
  for (p in 1:P) {
    midterm[p] ~ dnorm(pupil.mu[p], pupil.tau)
    pupil.mu[p] <- pupil.alpha.p[p] + beta.ipr*ipr.mu.p[p]
    pupil.alpha.p[p] ~ dnorm(alpha.mu, alpha.tau)
    ipr.mu.p[p] ~ dnorm(ipr.mu.p.mu, ipr.mu.p.tau)
    ipr.sigma.p[p] ~ dunif(0, 100)
    ipr.tau.p[p] <- pow(ipr.sigma.p[p], -2)
  }

  for (r in 1:IPR) {
    ipr[r] ~ dnorm(ipr.mu.p[ipr_p[r]], ipr.tau.p[ipr_p[r]])
  }

  pupil.sigma ~ dunif(0, 100)
  pupil.tau <- pow(pupil.sigma, -2)

  alpha.mu ~ dnorm(0, 0.001)
  alpha.tau <- pow(alpha.sigma, -2)
  alpha.sigma ~ dunif(0, 100)

  beta.ipr ~ dnorm(0, 0.001)

  ipr.mu.p.mu ~ dnorm(0, 0.001)
  ipr.mu.p.sigma ~ dunif(0, 100)
  ipr.mu.p.tau <- pow(ipr.mu.p.sigma, -2)
}
```

2.3 Model 5.2a

```
model {
  for (p in 1:P) {
    midterm[p] ~ dnorm(pupil.mu[p], pupil.tau)
    pupil.mu[p] <- pupil.alpha.p[p] +
      inprod(beta.ipr.rd[], ipr.mu.p.rd[p,])
    pupil.alpha.p[p] ~ dnorm(alpha.mu, alpha.tau)
    for (rd in 1:RD) {
      ipr.mu.p.rd[p,rd] ~ dnorm(ipr.mu.p.rd.mu.rd[rd],
        ipr.mu.p.rd.tau.rd[rd])
    }
  }

  for (r in 1:IPR) {
    ipr[r] ~ dnorm(ipr.mu.p.rd[ipr_p[r],ipr_rd[r]],
      ipr.tau.rd[ipr_rd[r]])
  }

  for (rd in 1:RD) {
    beta.ipr.rd[rd] ~ dnorm(0, 0.001)

    ipr.mu.p.rd.mu.rd[rd] ~ dnorm(0, 0.001)
    ipr.mu.p.rd.tau.rd[rd] <- pow(ipr.mu.p.rd.sigma.rd[rd], -2)
    ipr.mu.p.rd.sigma.rd[rd] ~ dunif(0, 100)

    ipr.tau.rd[rd] <- pow(ipr.sigma.rd[rd], -2)
    ipr.sigma.rd[rd] ~ dunif(0, 100)
  }

  alpha.mu ~ dnorm(0, 0.001)
  alpha.tau <- pow(alpha.sigma, -2)
  alpha.sigma ~ dunif(0, 100)

  pupil.tau <- pow(pupil.sigma, -2)
  pupil.sigma ~ dunif(0, 100)
}
```

2.4 Model 5.3a

```
model {
  for (p in 1:P) {
```

```
  midterm[p] ~ dnorm(pupil.mu[p], pupil.tau)
  pupil.mu[p] <- pupil.alpha +
    inprod(beta.ipr.rd[], ipr.mu.p.rd[p,]) +
    inprod(beta.ibr.rd[], ibr.mu.p.rd[p,])
  for (rd in 1:RD) {
    ipr.mu.p.rd[p,rd] ~ dnorm(ipr.mu.p.rd.mu.rd[rd],
      ipr.mu.p.rd.tau.rd[rd])
    ibr.mu.p.rd[p,rd] ~ dnorm(ibr.mu.p.rd.mu.rd[rd],
      ibr.mu.p.rd.tau.rd[rd])
  }
}

for (r in 1:IPR) {
  ipr[r] ~ dnorm(ipr.mu.p.rd[ipr_p[r],ipr_rd[r]],
    ipr.tau.rd[ipr_rd[r]])
}

for (r in 1:IBR) {
  ibr[r] ~ dnorm(ibr.mu.p.rd[ibr_p[r],ibr_rd[r]],
    ibr.tau.rd[ibr_rd[r]])
}

for (rd in 1:RD) {
  beta.ipr.rd[rd] ~ dnorm(0, 0.001)
  beta.ibr.rd[rd] ~ dnorm(0, 0.001)

  ipr.mu.p.rd.mu.rd[rd] ~ dnorm(0, 0.001)
  ipr.mu.p.rd.tau.rd[rd] <- pow(ipr.mu.p.rd.sigma.rd[rd], -2)
  ipr.mu.p.rd.sigma.rd[rd] ~ dunif(0, 100)

  ibr.mu.p.rd.mu.rd[rd] ~ dnorm(0, 0.001)
  ibr.mu.p.rd.tau.rd[rd] <- pow(ibr.mu.p.rd.sigma.rd[rd], -2)
  ibr.mu.p.rd.sigma.rd[rd] ~ dunif(0, 100)
```

```
    ipr.tau.rd[rd] <- pow(ipr.sigma.rd[rd], -2)
    ipr.sigma.rd[rd] ~ dunif(0, 100)

    ibr.tau.rd[rd] <- pow(ibr.sigma.rd[rd], -2)
    ibr.sigma.rd[rd] ~ dunif(0, 100)
  }

  pupil.sigma ~ dunif(0, 100)
  pupil.tau <- pow(pupil.sigma, -2)

  pupil.alpha ~ dnorm(0, 0.001)
}
```

Appendix C

Output of Bayesian Data Analysis

To interpret these figures, taking Figure C.2 as an example, there are three display panels.[1] The top is the "trace plot", showing the trace of MCMC sampling across the space of values that β could take. There are three executions of the MCMC sampler ("chains"), corresponding to the colors red, green, and black. One can ascertain visually that the chains' traces overlap a great deal, i.e., the samplers did not get stuck in local maxima (they "mixed"). The middle panel is the autocorrelation plot, which shows how rapidly the chains explored the space of values. From left to right, there is a fair amount of lag between adjacent MCMC samples within a chain, indicating that the model is slow to converge on an estimate for the parameter. The rate of convergence may be an issue for some uses of the model, but it does not affect the model's validity given that the three chains do converge on very similar estimates, as indicated by the Rhat value. Values less than 1.2 are conventionally taken to indicate that the model has converged. Finally, the bottom panel shows the probability density estimate itself, according to each chain, including the mean (thick vertical line) and minimum and maximum bounds at 2.5% and 97.5% (thin vertical lines). This is the 95% posterior credible interval.

An alternative display is used for parameters that can be displayed as a group, e.g., the mean of inbound peer ratings that we calculate for each peer author (Figure C.3). The left panel displays the 95% posterior intervals for each parameter, with the parameter mean shown by a tick mark. The center panel shows autocorrelation lag, if any, which may indicate slow convergence. Finally, the right panel shows the Rhat value, indicating whether or not the chains converge on an estimate.

For all figures showing an estimate of a sigma parameter (e.g., σ and $\sigma_{p[IPR]}$), note that the estimated value is the standard deviation, which must be squared to arrive at a variance.

[1] These visualizations were generated using Howard Seltman's Rube. www.stat.cmu.edu/~hseltman/rube/

3.1 Model 5.1a: Baseline

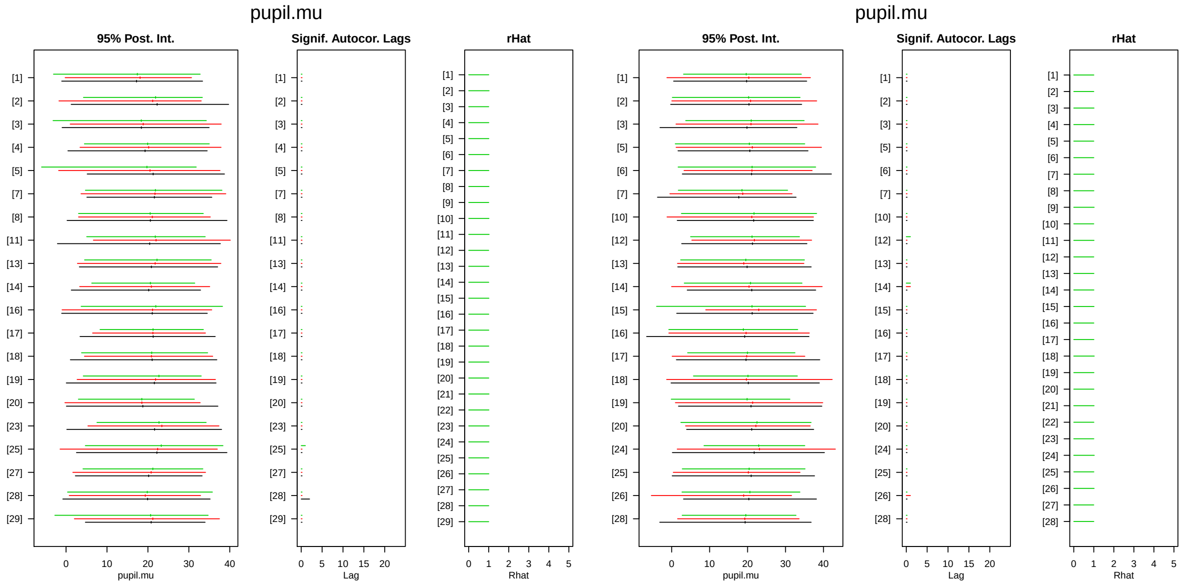

Figure C.1: Model 5.1a estimates for μ_p, domain-relevant (left) and problem-specific (right)

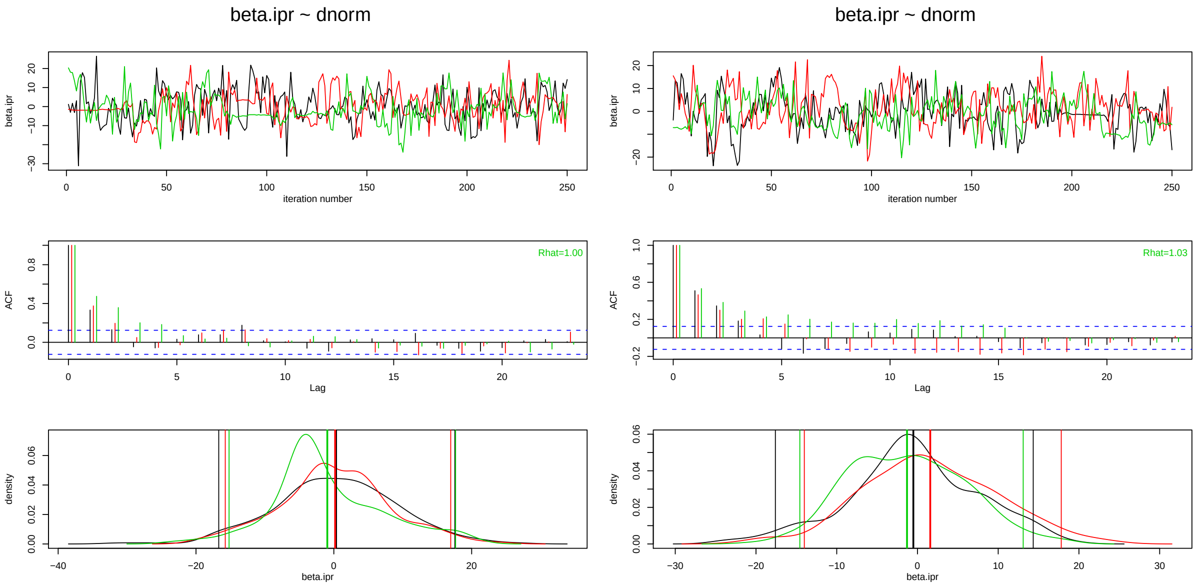

Figure C.2: Model 5.1a estimates for β_1, domain-relevant (left) and problem-specific (right)

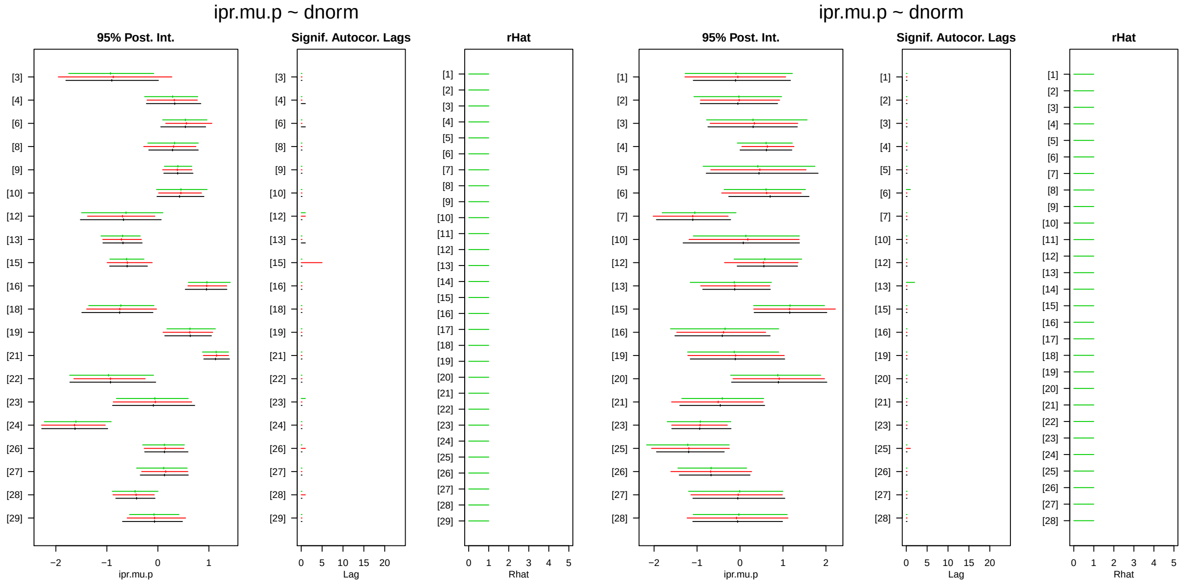

Figure C.3: Model 5.1a estimates for X_{1p}, domain-relevant (left) and problem-specific (right)

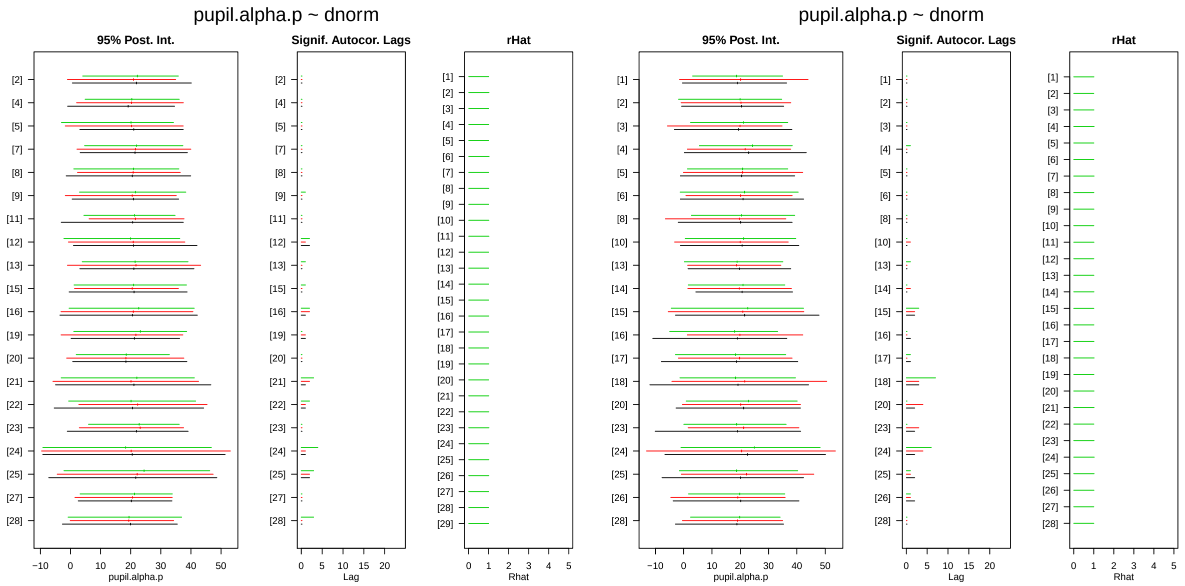

Figure C.4: Model 5.1a estimates for α_p, domain-relevant (left) and problem-specific (right)

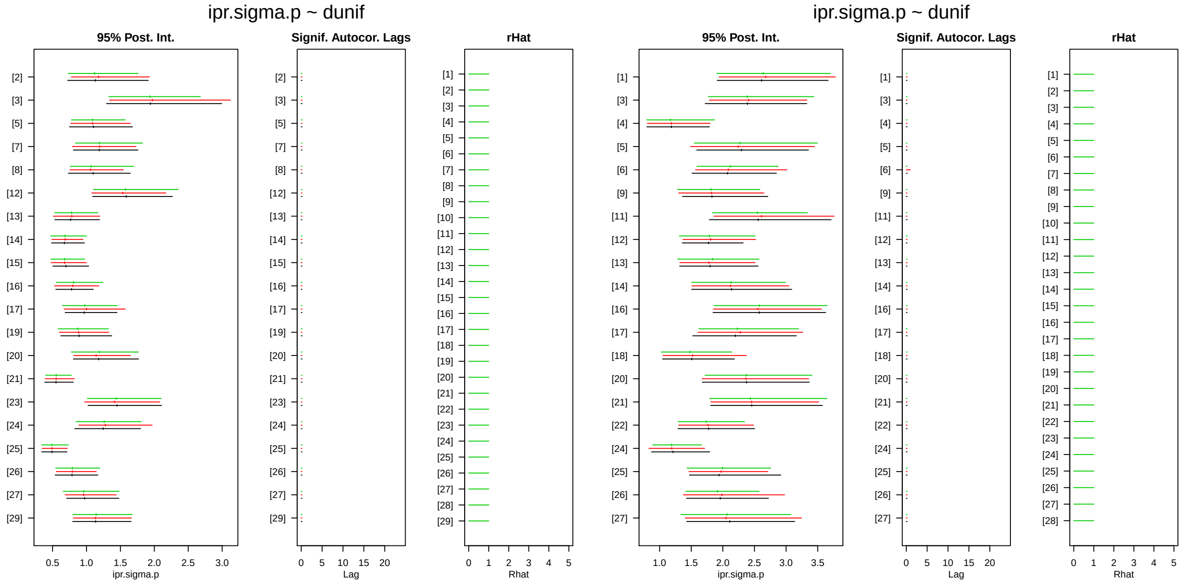

Figure C.5: Model 5.1a estimates for $\sigma_{p[IPR]}$, domain-relevant (left) and problem-specific (right)

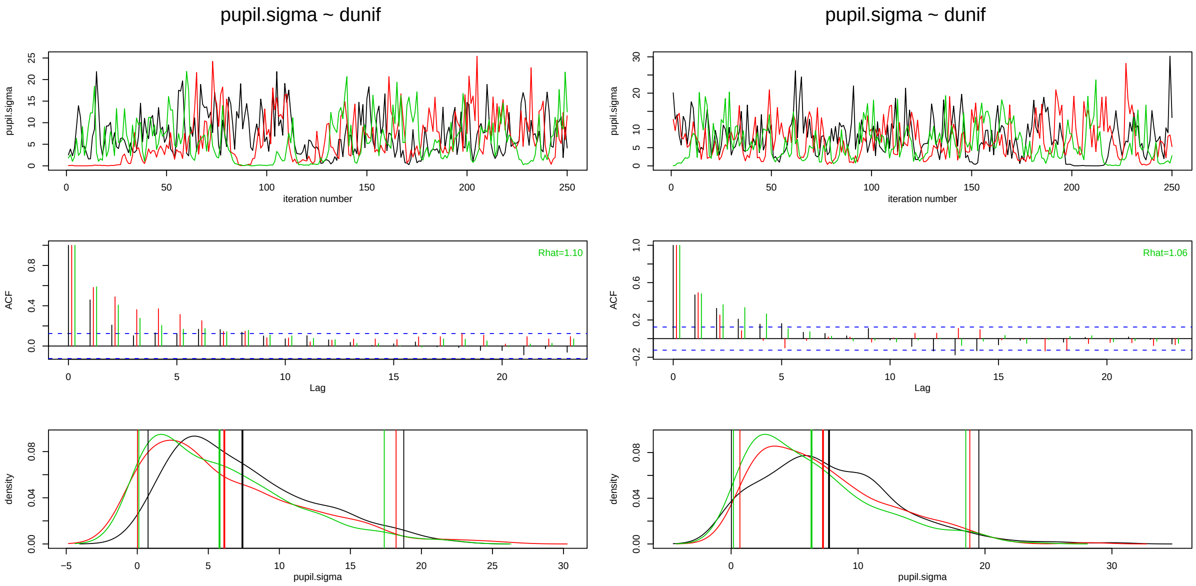

Figure C.6: Model 5.1a estimates for σ, domain-relevant (left) and problem-specific (right)

3.2 Model 5.1b: Contribution of Information Pooling

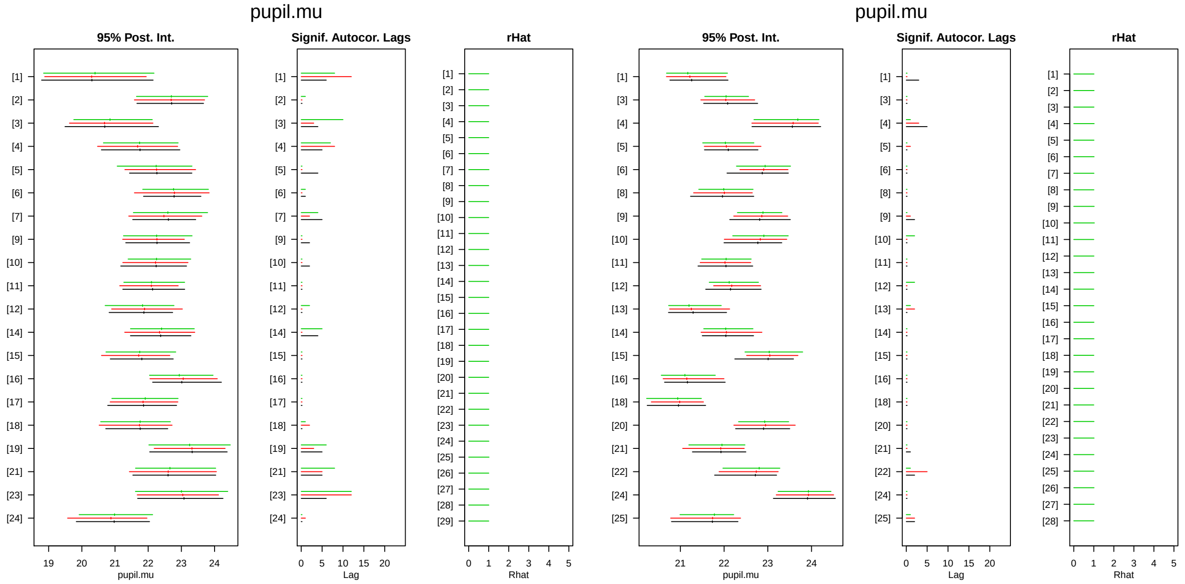

Figure C.7: Model 5.1b estimates for μ_p, domain-relevant (left) and problem-specific (right)

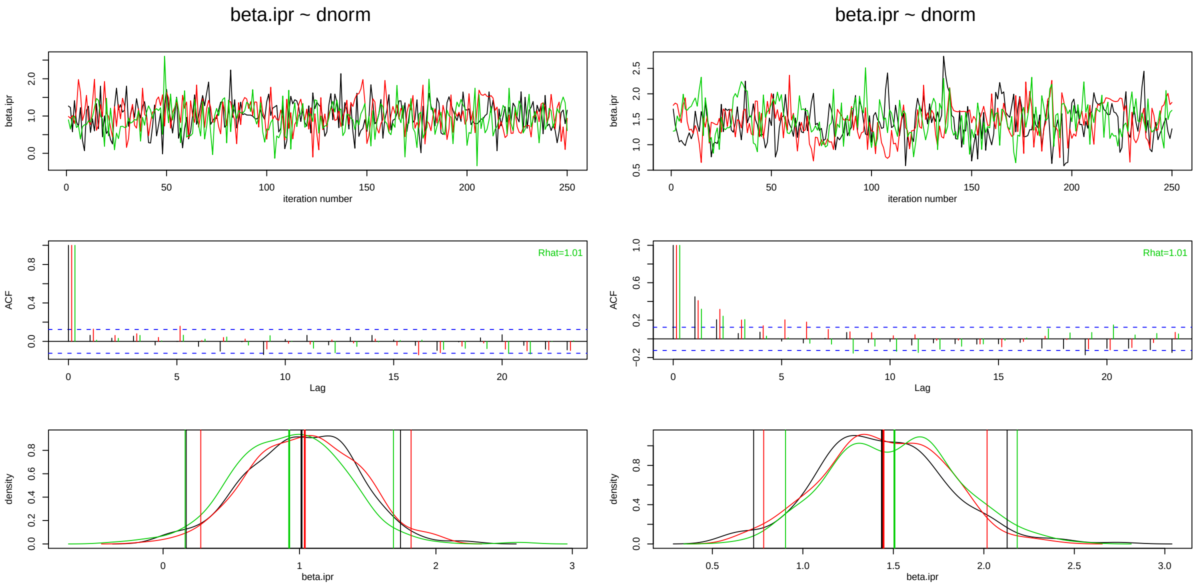

Figure C.8: Model 5.1b estimates for β_1, domain-relevant (left) and problem-specific (right)

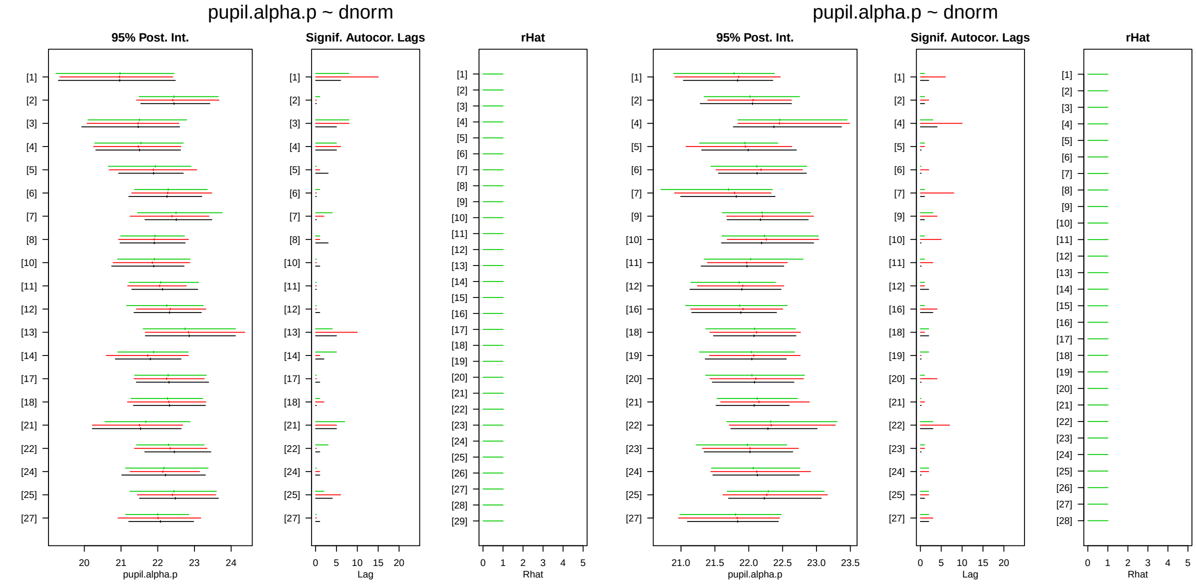

Figure C.9: Model 5.1b estimates for α_p, domain-relevant (left) and problem-specific (right)

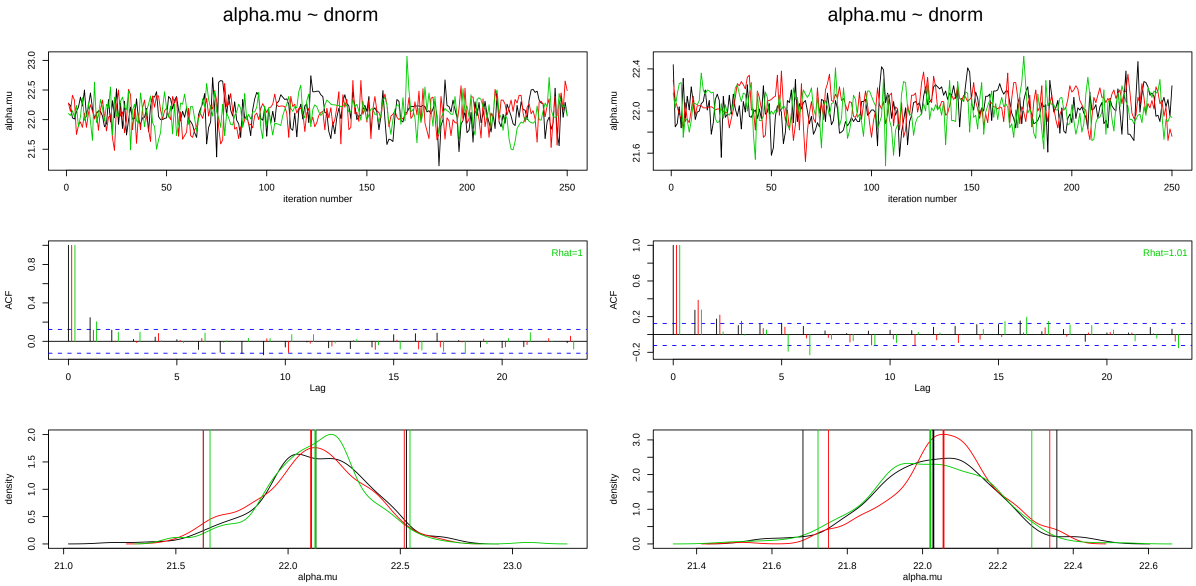

Figure C.10: Model 5.1b estimates for μ_α, domain-relevant (left) and problem-specific (right)

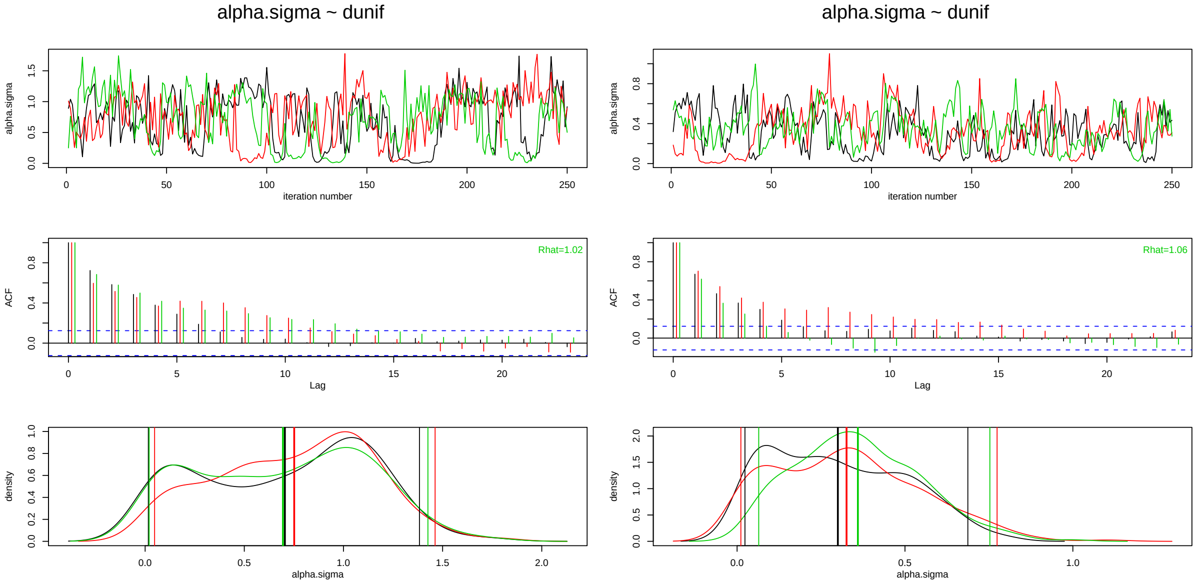

Figure C.11: Model 5.1b estimates for σ_α, domain-relevant (left) and problem-specific (right)

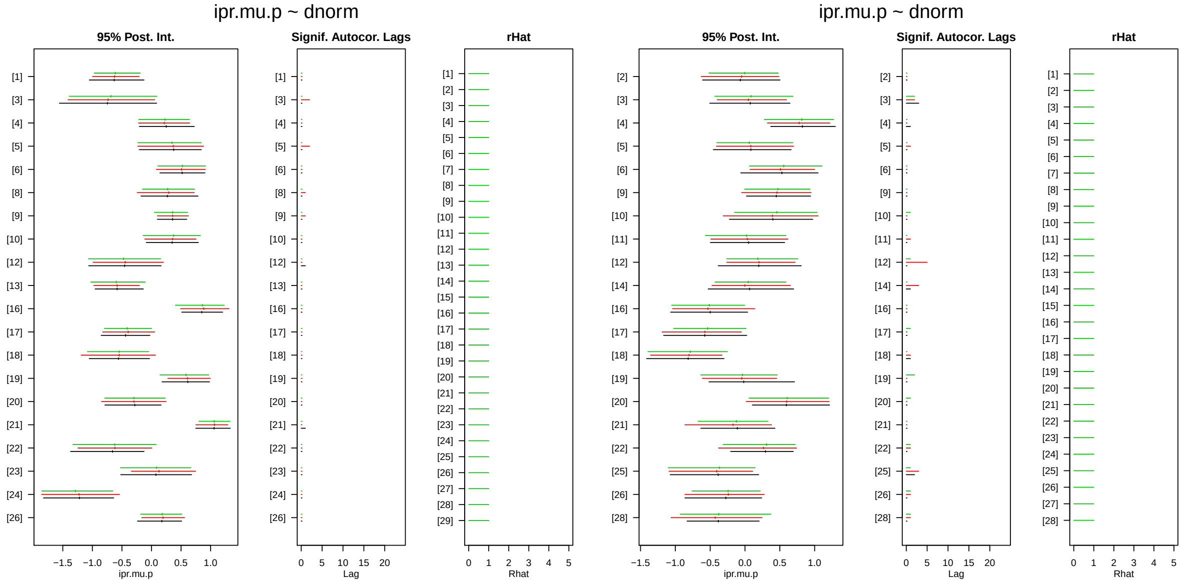

Figure C.12: Model 5.1b estimates for X_{1p}, domain-relevant (left) and problem-specific (right)

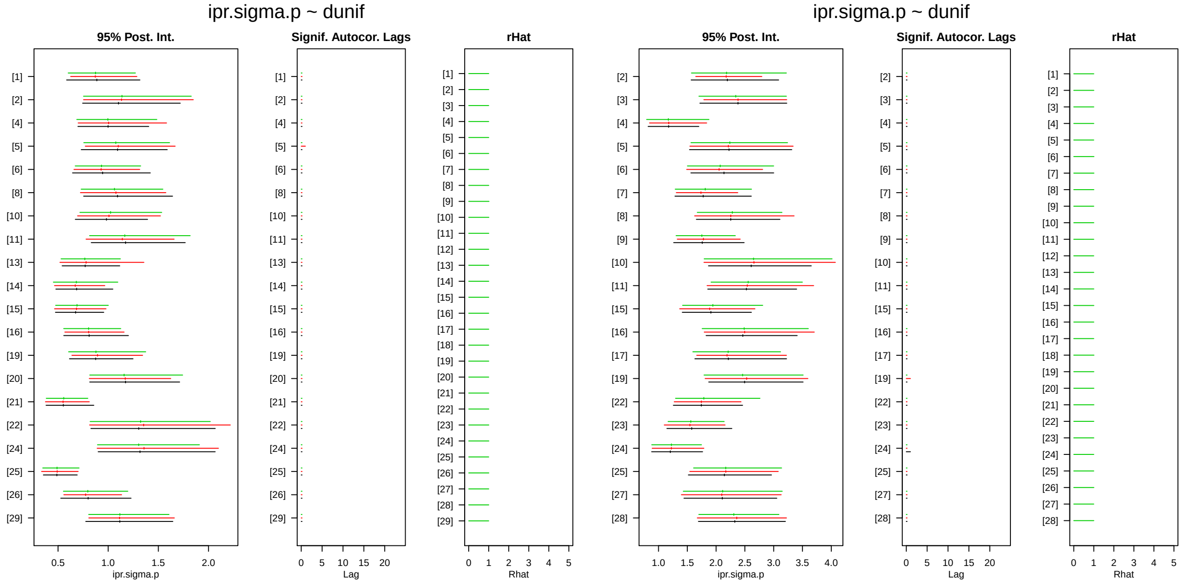

Figure C.13: Model 5.1b estimates for $\sigma_{p[IPR]}$, domain-relevant (left) and problem-specific (right)

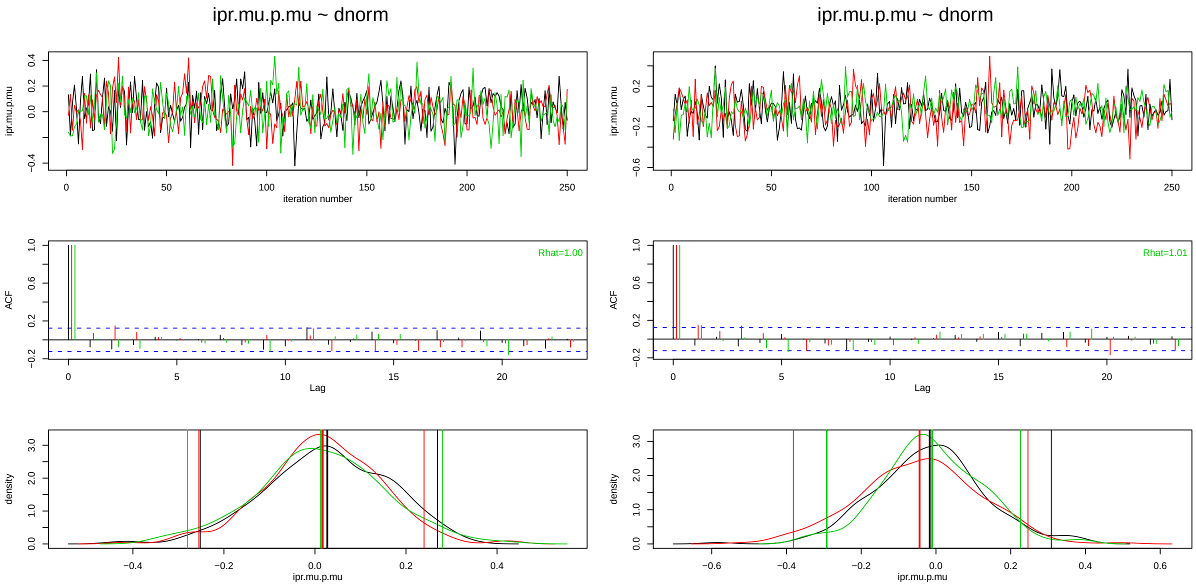

Figure C.14: Model 5.1b estimates for μ_{IPR}, domain-relevant (left) and problem-specific (right)

3.3 Model 5.2a: Contribution of Rating Dimensions

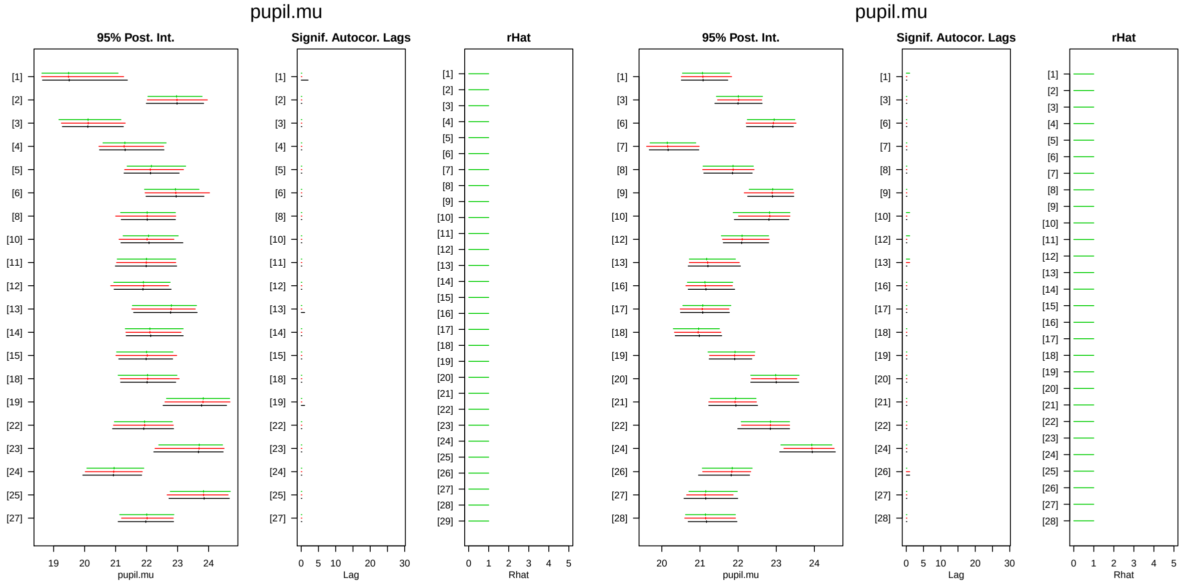

Figure C.15: Model 5.2a estimates for μ_p, domain-relevant (left) and problem-specific (right)

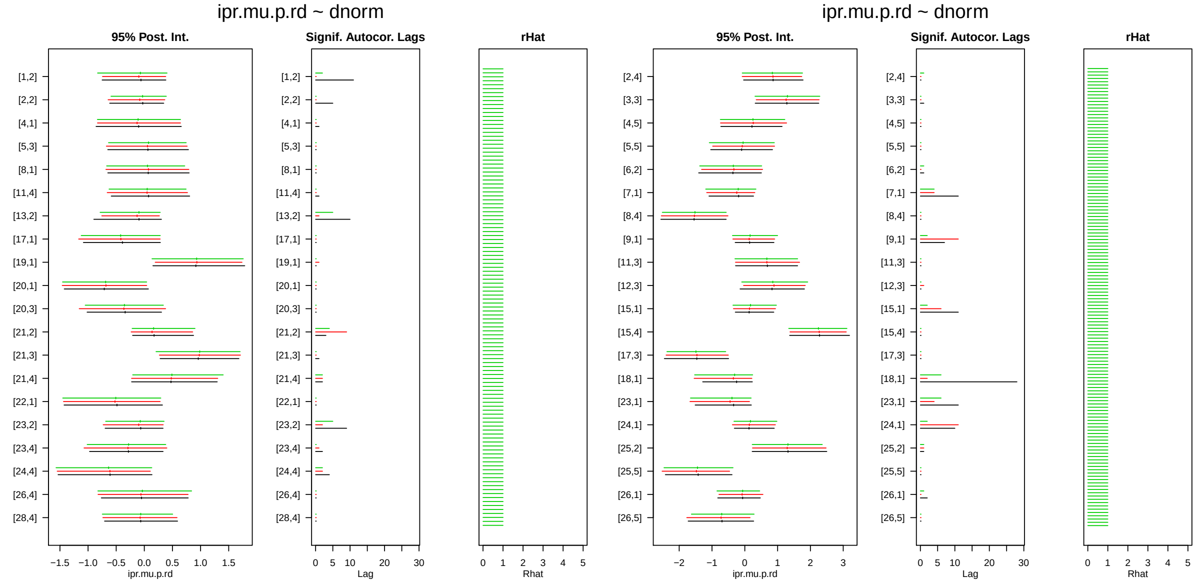

Figure C.16: Model 5.2a estimates for X_{np}, domain-relevant (left) and problem-specific (right)

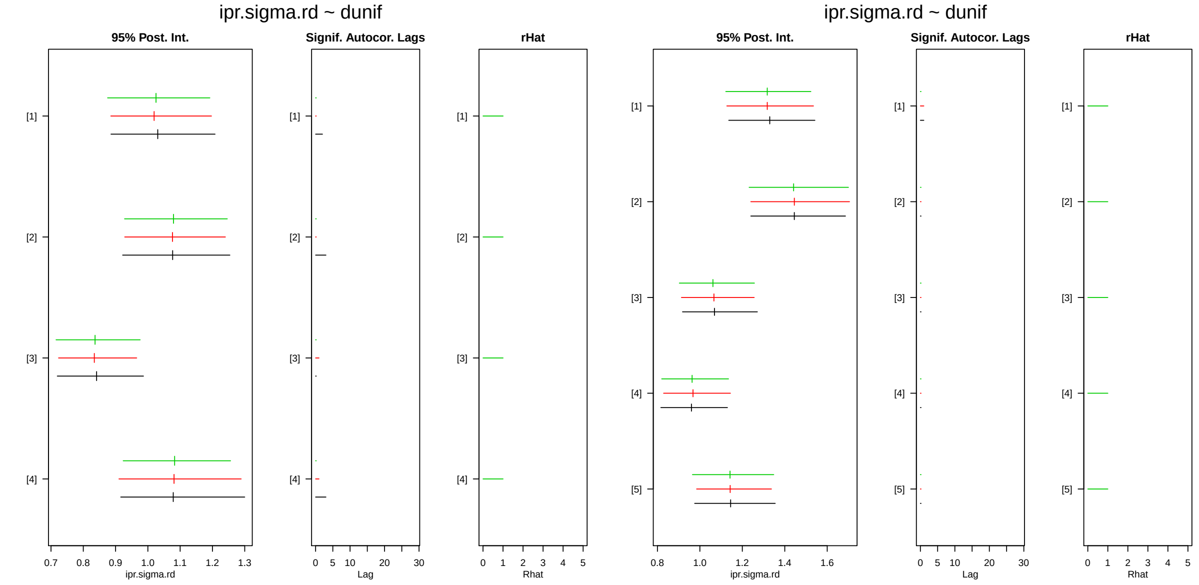

Figure C.17: Model 5.2a estimates for $\sigma^2_{n[IPR]}$, domain-relevant (left) and problem-specific (right)

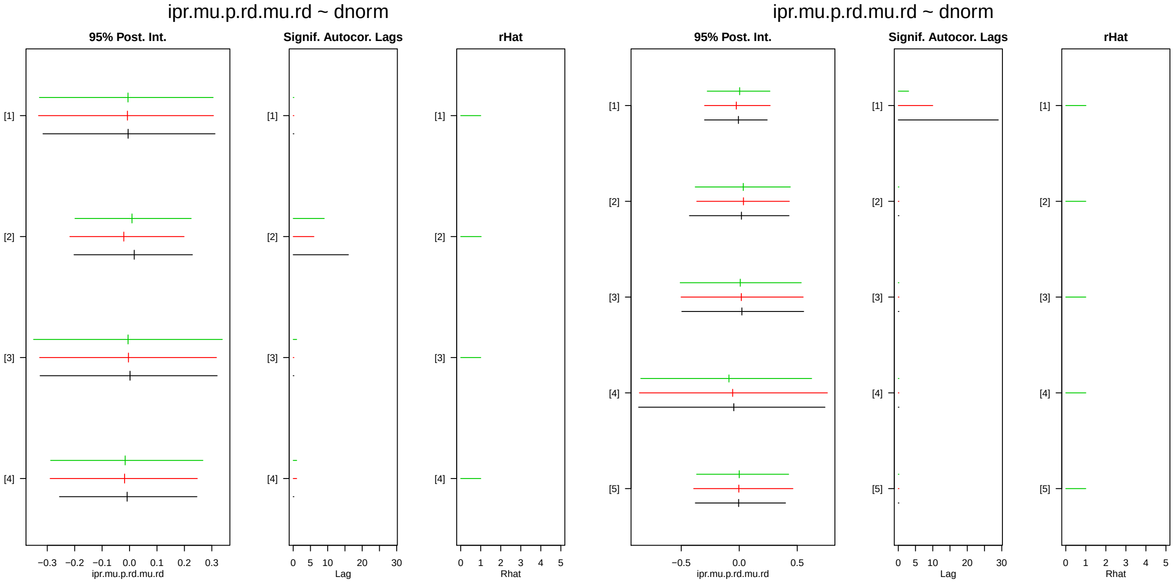

Figure C.18: Model 5.2a estimates for the mean hyperparameter for X_{np}, domain-relevant (left) and problem-specific (right)

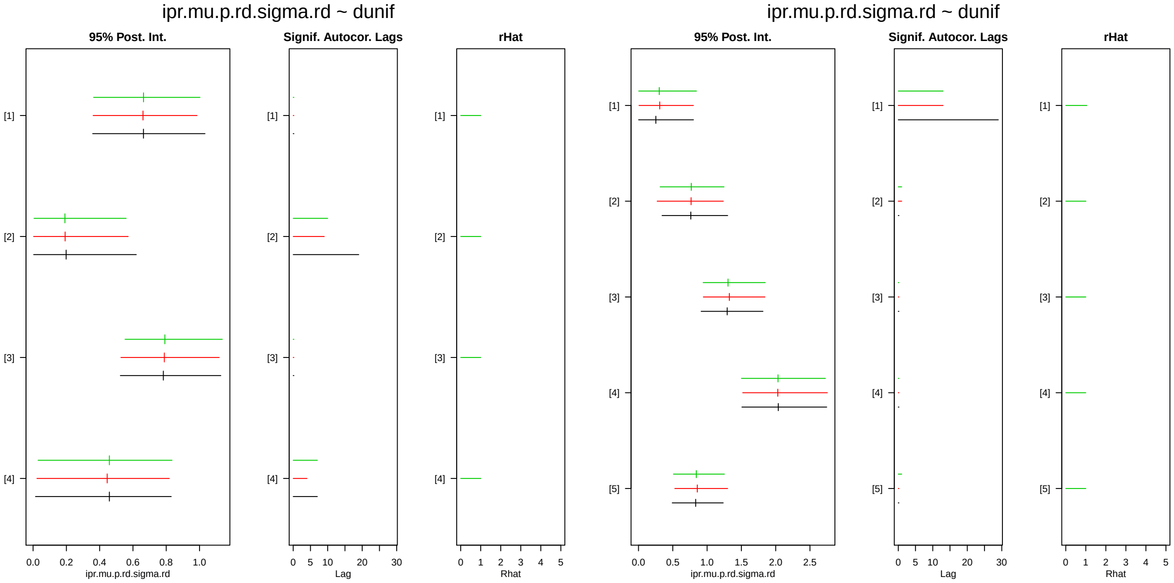

Figure C.19: Model 5.2a estimates for the variance hyperparameter for X_{np}, domain-relevant (left) and problem-specific (right)

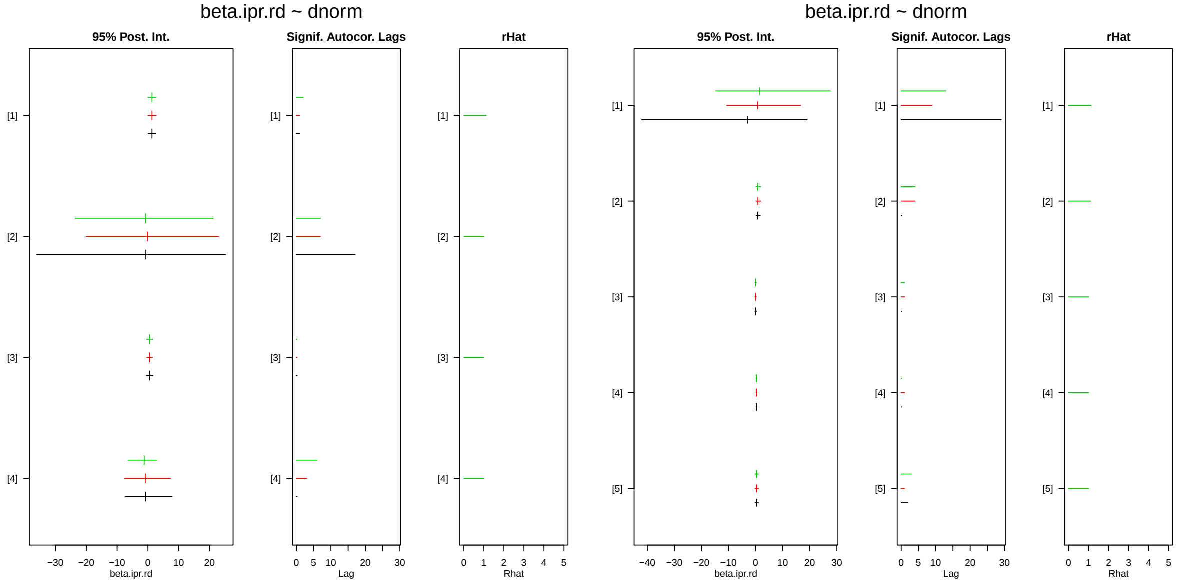

Figure C.20: Model 5.2a estimates for β_n, domain-relevant (left) and problem-specific (right)

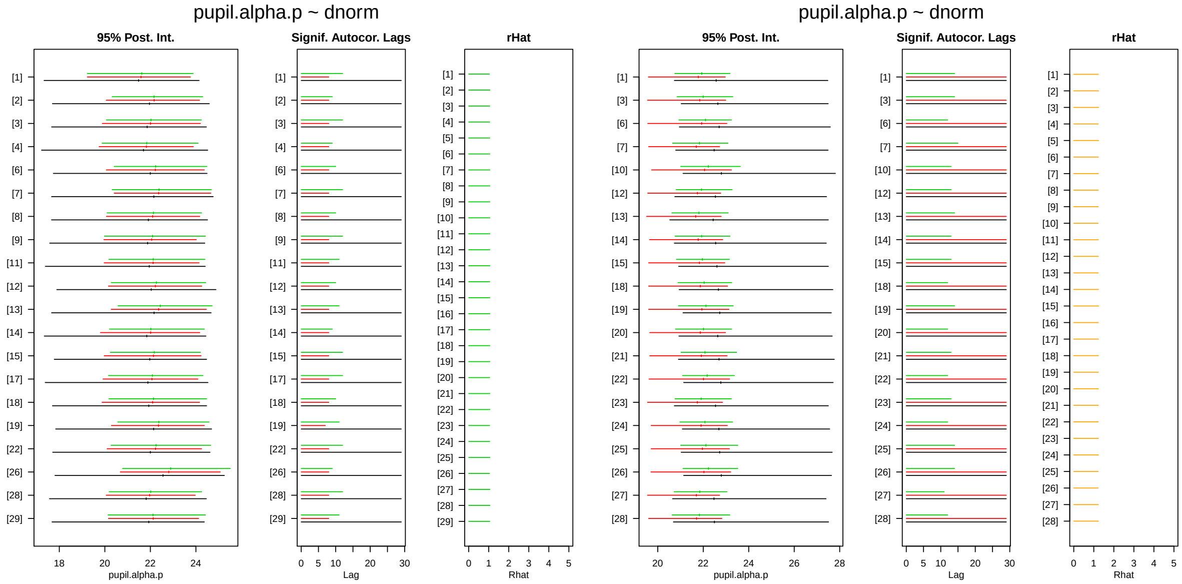

Figure C.21: Model 5.2a estimates for α_p, domain-relevant (left) and problem-specific (right)

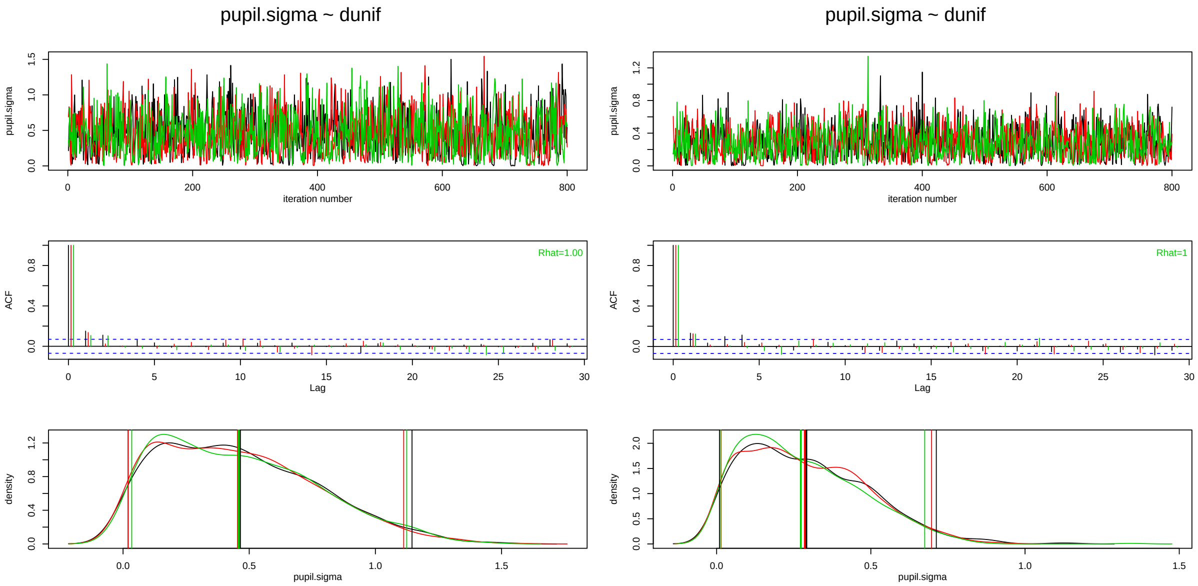

Figure C.22: Model 5.2a estimates for σ, domain-relevant (left) and problem-specific (right)

3.4 Model 5.3a: Contribution of Inbound Back-Reviews

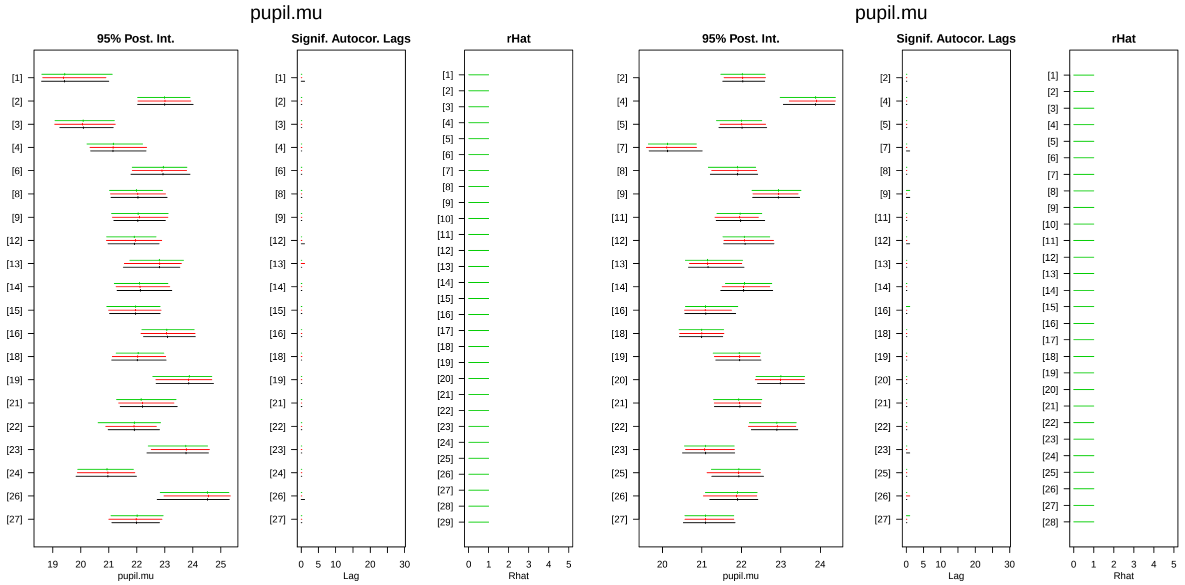

Figure C.23: Model 5.3a estimates for μ_p, domain-relevant (left) and problem-specific (right)

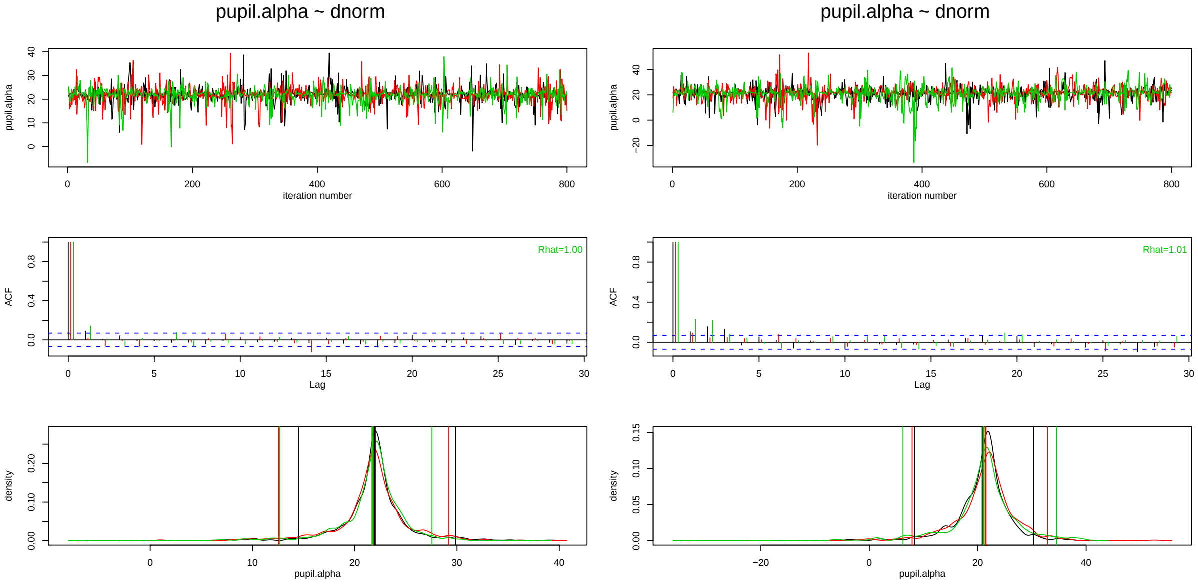

Figure C.24: Model 5.3a estimates for α, domain-relevant (left) and problem-specific (right)

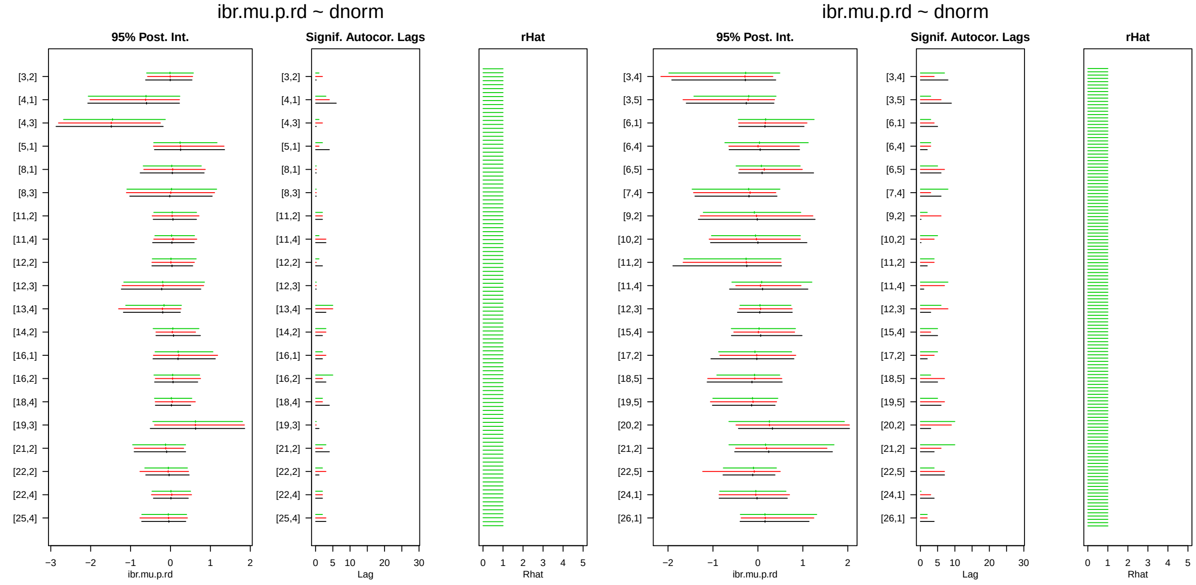

Figure C.25: Model 5.3a estimates for Z_{np}, domain-relevant (left) and problem-specific (right)

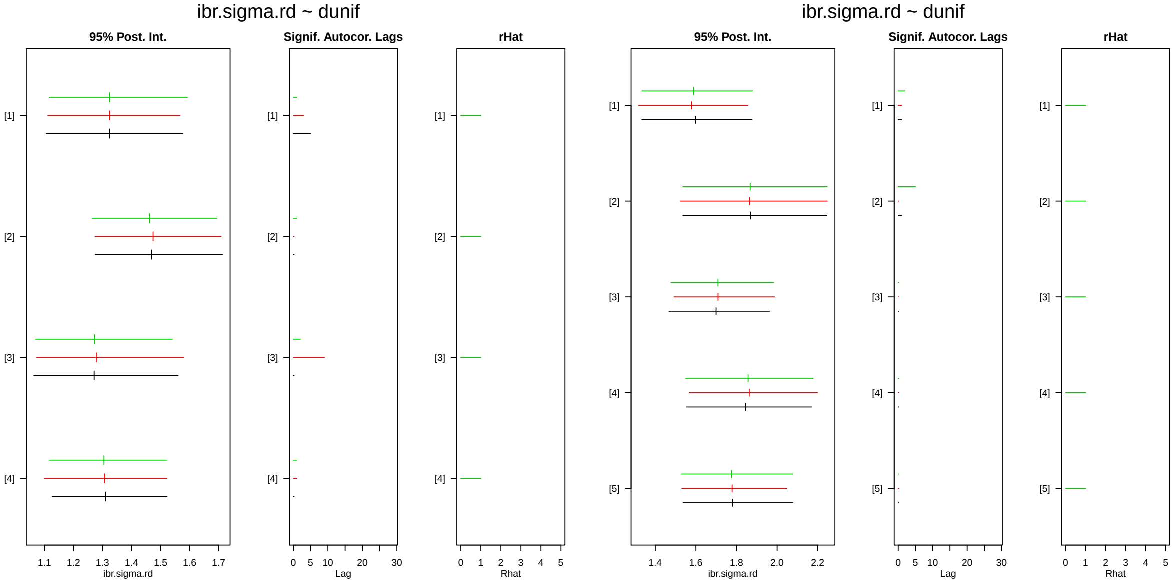

Figure C.26: Model 5.3a estimates for $\sigma_{n[IBR]}$, domain-relevant (left) and problem-specific (right)

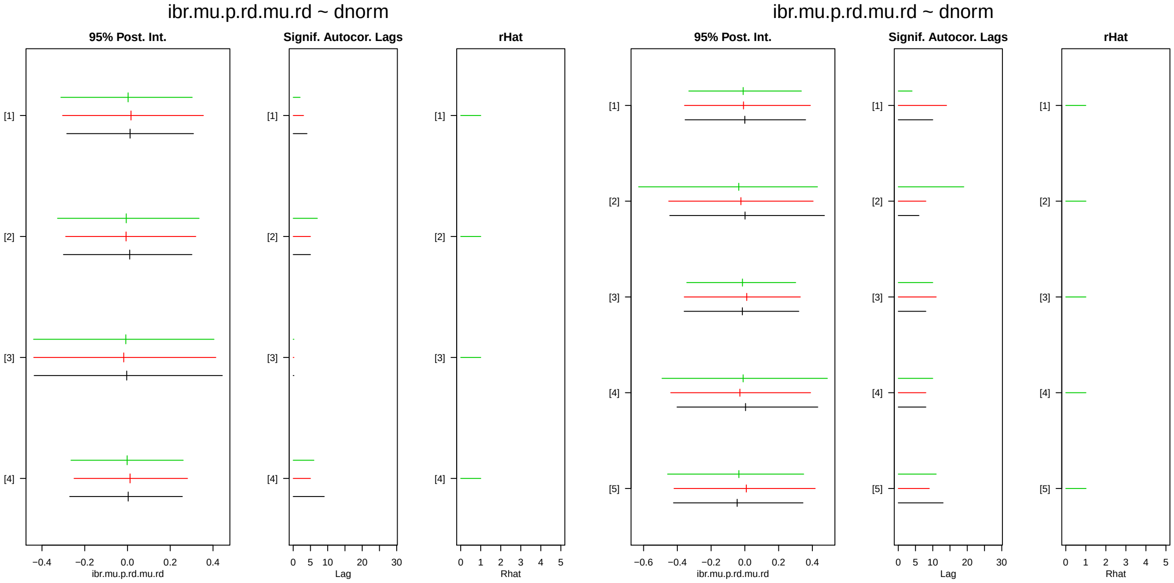

Figure C.27: Model 5.3a estimates for the mean hyperparameter for Z_{np}, domain-relevant (left) and problem-specific (right)

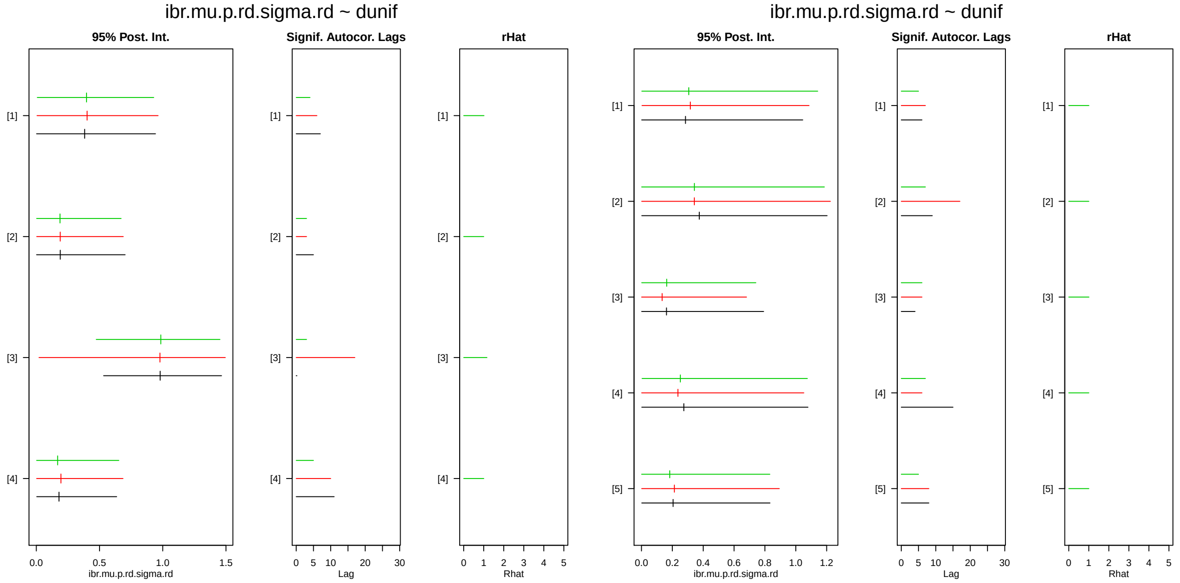

Figure C.28: Model 5.3a estimates for the variance hyperparameter for Z_{np}, domain-relevant (left) and problem-specific (right)

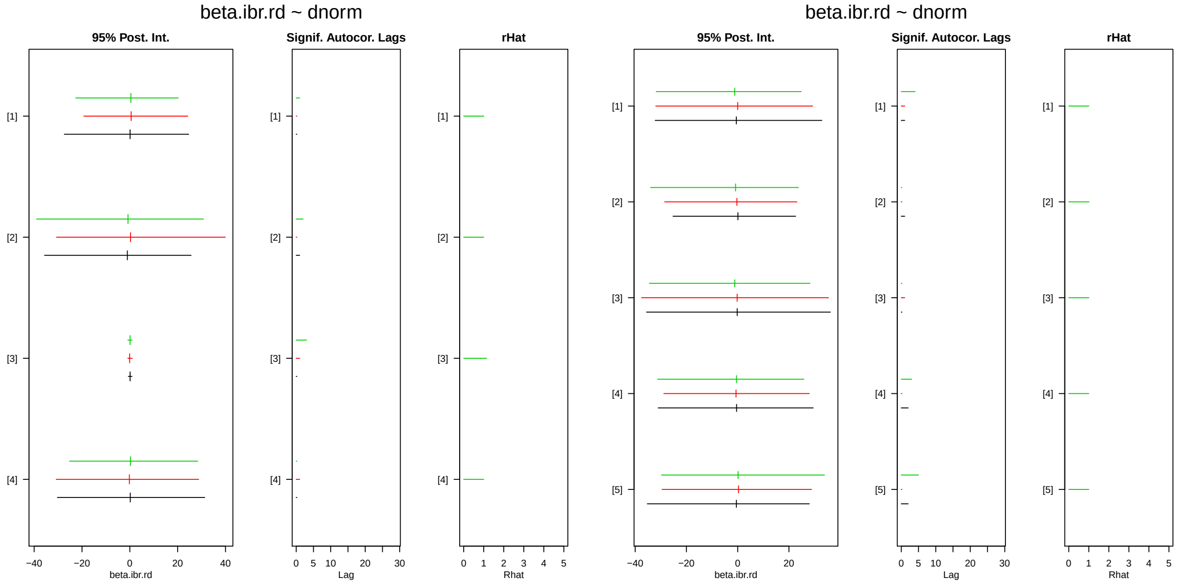

Figure C.29: Model 5.3a estimates for γ_n, domain-relevant (left) and problem-specific (right)

References

Alamargot, D., & Chanquoy, L. (2001). *Through the models of writing*. Springer.

Aleven, V., Lynch, C., Pinkwart, N., & Ashley, K. D. (Eds.). (2009). *Special issue on Ill-Defined domains* (Vol. 19 (3)).

Aleven, V., Ogan, A., Popescu, O., Torrey, C., & Koedinger, K. R. (2004). Evaluating the effectiveness of a tutorial dialogue system for Self-Explanation.

Anderson, N., & Shneiderman, B. (1977). Use of peer ratings in evaluating computer program quality. In *Proceedings of the fifteenth annual SIGCPR conference* (pp. 218–226). ACM.

Andrade, H., & Du, Y. (2007, April). Student responses to criteria-referenced self-assessment. *Assessment & Evaluation in Higher Education*, *32*(2), 159–181.

Andrade, H. G. (2000). Using rubrics to promote thinking and learning. *Educational Leadership*, *57*(5), 13–19.

Andrade, H. L. (2010). Students as the definitive source of formative assessment: Academic Self-Assessment and the Self-Regulation of learning. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 90–105). New York: Routledge.

Armstrong, S. L., & Paulson, E. J. (2008, May). Whither "Peer review"? terminology matters for the writing classroom. *Teaching English in the Two-Year College*, *35*(4), 398–407.

Ashley, K., Pinkwart, N., Lynch, C., & Aleven, V. (2007). Learning by diagramming supreme court oral arguments. In *11th international conference on artificial intelligence and law* (pp. 271–275). Stanford, CA: ACM Press.

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*(2), 153–170. Available from `http://linkinghub.elsevier.com/retrieve/pii/S0191491X06000228`

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

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. Available from `http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED049304`

Boom, G. van den, Paas, F., & Merriënboer, J. J. van. (2007, October). Effects of elicited reflections combined with tutor or peer feedback on self-regulated learning and learning outcomes. *Learning and Instruction*, *17*(5), 532–548. Available from `http://www.sciencedirect.com/science/article/B6VFW-4PXNHCW-1/2/fecdec6f0beac4c55ac70abf307754d2`

Buchanan, R. (1992). Wicked problems in design thinking. *Design issues*, *8*(2), 5–21.

Buchanan, R. (2001, January). Design and the new rhetoric: Productive arts in the philosophy of culture. *Philosophy & Rhetoric*, *34*(3), 183–206. Available from `http://www.jstor.org/stable/40238091` (ArticleType: research-article / Full publication date: 2001 / Copyright © 2001 Penn State University Press)

Burstein, J., Marcu, D., & Knight, K. (2003). Finding the WRITE stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems*.

Carr, C. (2003). Using computer supported argument visualization to teach legal argumentation. In *Visualizing argumentation: Software tools for collaborative and educational Sense-Making* (p. 75–96). London, UK: Springer-Verlag.

Chalk, B., & Adeboye, K. (2005, September). Peer assessment of program code: a comparison of two feedback instruments. In *6th annual conference for the higher edcation academy subject network for information and computer science (HEA-ICS).* University of York, UK.

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive science*, *5*(2), 121–152.

Chi, M. T., Leeuw, N. de, Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*, 439–477. Available from `http://www.pitt.edu/~chi/papers/ChideLeeuwChiuLaVancher.pdf`

Cho, K. (2008). Machine classification of peer comments in physics. In *Educational data mining* (pp. 192–196).

Cho, K., & Cho, Y. H. (2007). Learning from ill-structured cases. In D. S. McNamara &

J. Trafton (Eds.), *29th annual cognitive science society conference* (p. 1722). Austin, TX: Cognitive Science Society.

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. Available from `http://portal.acm.org/citation.cfm?id=1769477`

Cho, K., Schunn, C., & Charney, D. (2006). Commenting on writing. typology and perceived helpfulness of comments from novice peer reviewers and subject matter experts. *Written Communication*, *23*(3), 260–294.

Cho, K., & Schunn, C. D. (2007). Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers and Education*, *48*(3). Available from `http://dx.doi.org/10.1016/j.compedu.2005.02.004`

Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology*, *98*(4), 891–901. Available from `http://doi.apa.org/getdoi.cfm?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, judging* (pp. 3–32). National Council of Teachers of English. Available from `http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED143020`

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*(4), 253–278. Available from `http://dx.doi.org/10.1007/BF01099821`

Crespo García, R. M., Pardo, A., & Delgado Kloos, C. (2006). Adaptive peer review based on student profiles. In M. Ikeda, K. Ashley, & T. Chan (Eds.), *Intelligent tutoring systems* (Vol. 4053, pp. 781–783). Springer Berlin / Heidelberg. Available from `http://dx.doi.org/10.1007/11774303_99` (10.1007/11774303_99)

Cunningham, S. J. (1994). Using a computer conferencing system to support writing and research skill development. *SIGCSE Bull.*, *26*(4), 5–8. Available from `http://portal.acm.org/citation.cfm?id=190652`

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

Does learning to write have to be so difficult. (1983). In A. Freedman, I. Pringle, J. Yalden, C. C. of Teachers of English, C. Bereiter, & M. Scardamalia (Eds.), *Learning to write : First language, second language* (pp. 20–33). Longman.

Draaijer, S., & Boxel, P. van. (2006, July). Summative peer assessment using 'Turnitin'and a large cohort of students: a case study. In M. Danson (Ed.), *10th CAA international computer assisted assessment conference.* Loughborough University: Professional Development Loughborough University.

Falchikov, N., & Goldfinch, J. (2000, January). Student peer assessment in higher education: a meta-analysis comparing peer and teacher marks. *Review of Educational Research*, *70*(3), 287–322. Available from `http://rer.sagepub.com/cgi/doi/10.3102/00346543070003287`

Foltz, P., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in online writing evaluation with LSA. *Interactive Learning Environments*, *8*(2), 111–129.

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*(3), 435–450. Available from `http://www.eric.ed.gov/ERICWebPortal/detail?accno=EJ788267`

Gehringer, E. (2000). Strategies and mechanisms for electronic peer review. In *30th annual frontiers in education conference* (Vol. 1, pp. F1B/2–F1B/7 vol.1).

Gehringer, E. F., Gummadi, A., Kadanjoth, R., & Andrés, Y. M. (2010). Motivating effective peer review with extra credit and leaderboards. Available from `http://soa.asee.org/paper/conference/paper-view.cfm?id=23703`

Gelman, A., & Hill, J. (2006). *Data analysis using regression and Multilevel/Hierarchical models*. Cambridge University Press.

Gielen, S., Dochy, F., & Onghena, P. (2010). An inventory of peer assessment diversity. *Assessment & Evaluation in Higher Education*, *99999*(1), 1. Available from `http://www.informaworld.com/10.1080/02602930903221444`

Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction*, *In Press, Corrected Proof*. Available from `http://www.sciencedirect.com/science/article/B6VFW-4XHC6BX-1/2/93f2479c042252711dea1230b0419e3a`

Godshalk, F. I., Swineford, F., & Coffman, W. E. (1966). *The measurement of writing ability* (Tech. Rep. No. CEEB RM No. 6.) Princeton, NJ: College Entrance Examination Board. Available from `http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED029028`

Goldin, I., Pinkus, R., & Ashley, K. (n.d.). Validity and reliability of an instrument for assessing case analyses in bioengineering ethics education. *Science and Engineering Ethics*.

Goldin, I. M., Brusilovsky, P., Schunn, C., Ashley, K. D., & Hsiao, I. (Eds.). (2010). *Workshop on Computer-Supported peer review in education, 10th international conference on intelligent tutoring systems*. Pittsburgh, PA. Available from `http://cspred.org`

Gouli, E., Gogoulou, A., & Grigoriadou, M. (2008). Supporting self-, peer-, and collaborative- assessment in e-learning: the case of the PEer and collaborative ASSessment environment (PECASSE). *Journal of Interactive Learning Research*, *19*(4), 615.

Graesser, A., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Person, N., & Tutoring Research Group the. (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments*, *8*, 149–169. Available from `http://internal.autotutor.org/papers/using.pdf`

Graesser, A. C., & Wiemer-Hastings, P. (2000, August). Select-a-Kibitzer: a computer tool that gives meaningful feedback on student compositions. *Interactive Learning Environments*, *8*(2), 149–169. Available from `http://www.informaworld.com/openurl?genre=article&doi=10.1076/1049-4820(200008)8:2;1-B;FT149&magic=crossref||D404A21C5BB053405B1A640AFFD44AE3`

Hacker, D. J., Keener, M. C., & Kircher, J. C. (2009, June). Writing is applied metacognition. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (1st ed., pp. 154–172). Routledge.

Hamer, J., Ma, K. T. K., & Kwong, H. H. F. (2005). A method of automatic grade calibration in peer assessment. In *Proceedings of the 7th australasian conference on computing education - volume 42* (pp. 67–72). Newcastle, New South Wales, Australia: Australian Computer Society, Inc.

Harris, C., Pritchard, M., & Rabins, M. (2000). *Engineering ethics: Concepts and cases* (Vol. 2nd). Belmont, CA: Wadsworth.

Hattie, J., & Timperley, H. (2007, March). The power of feedback. *Review of Educational Research*, *77*(1), 81 –112. Available from `http://rer.sagepub.com/content/77/1/81.abstract`

Higgins, D., Burstein, J., Marcu, D., & Gentile, C. (2004). Evaluating multiple aspects of coherence in student essays. In *Proceedings of the annual meeting of HLT/NAACL.*

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

Hsiao, I., & Brusilovsky, P. (2008, October). Modeling peer review in example annotation. In *16th international conference on computers in education* (pp. 357–362). Taipei, Taiwan.

Hübner, S., Nückles, M., & Renkl, A. (2006). Prompting cognitive and metacognitive processing in writing-to-learn enhances learning outcomes. In *28th annual conference of the cognitive science society.* Available from `http://www.cogsci.rpi.edu/csjarchive/proceedings/2006/docs/p357.pdf`

Huot, B. (1990, May). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication*, *41*(2), 201–213. Available from `http://www.jstor.org/stable/358160` (ArticleType: research-article / Full publication date: May, 1990 / Copyright © 1990 National Council of Teachers of English)

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*(1), 27–44. Available from `http://www.eric.ed.gov/ERICWebPortal/detail?accno=EJ683510`

Jordan, P., Makatchev, M., Pappuswamy, U., VanLehn, K., & Albacete, P. (2006). A natural language tutorial dialogue system for physics. Available from `http://andes3.lrdc.pitt.edu/why/atlas-papers/flairs06-1.pdf`

King, A. (1997). ASK to THINK-TEL WHY: a model of transactive peer tutoring for scaffolding higher level complex learning. *Educational Psychologist*, *32*(4), 221–235.

Kwok, R. C., & Ma, J. (1999). Use of a group support system for collaborative assessment. *Computers & Education*, *32*(2), 109–125. Available from `http://linkinghub.elsevier.com/retrieve/pii/S0360131598000591`

Landauer, T., Laham, D., & Foltz, P. (2003). Automatic essay assessment. *Assessment in Education*, *10*(3), 295–308.

Landauer, T., Laham, D., Foltz, P., Shermis, M., & Burstein, J. (2003). Automated scoring and annotation of essays with the intelligent essay assessor. In *Automated essay scoring* (pp. 87–112).

Lauw, H. W., Lim, E.-p., & Wang, K. (2007). Summarizing review scores of "unequal" reviewers. In *IN PROCEEDINGS OF THE 2007 SIAM INTERNATIONAL CON-*

FERENCE ON DATA MINING. Available from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.150.6039

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). Princeton, NJ: Educational Testing Service.

Lemaire, B., & Dessus, P. (2001). A system to asess the semantic content of student essays. *Journal of Educational Computing Research*, *24*(3), 305–320.

Li, L., & Kay, J. (2005). Assess: Promoting learner reflection in student Self-Assessment. In *Workshop on learner modelling for reflection, to support learner control, metacognition and improved communication between teachers and learners at 12th international conference on artificial intelligence in education* (pp. 32–41). Amsterdam.

Lin, S., Liu, E., & Yuan, S. (2001). Web-based peer assessment: feedback for students with various thinking-styles. *Journal of Computer Assisted Learning*, *17*(4), 420–432. Available from http://dx.doi.org/10.1046/j.0266-4909.2001.00198.x

Lindblom-ylanne, S., Pihlajamaki, H., & Kotkas, T. (2006, March). Self-, peer- and teacher-assessment of student essays. *Active Learning in Higher Education*, *7*(1), 51–62. Available from http://alh.sagepub.com/cgi/content/abstract/7/1/51

Litman, D., & Purandare, A. (2008, May). Content-Learning correlations in spoken tutoring dialogs at word, turn and discourse levels. Coconut Grove, Florida. Available from http://www.cs.pitt.edu/%7Eamruta/pubs/2008/FLAIRS1PurandareA2.pdf

Liu, E. Z., Lin, S. S., & Yuan, S. (2002, December). 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.), *International conference on computers in education.* Auckland, New Zealand: IEEE Computer Society.

Liu, N. F., & Carless, D. (2006). Peer feedback: the learning element of peer assessment. *Teaching in Higher Education*, *11*(3), 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. Available from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED143020

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores.* Addison-Wesley Reading, MA:.

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*(2), 100–115.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS – a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, *10*(4), 325–337.

Lynch, C., Ashley, K., Aleven, V., & Pinkwart, N. (2006). Defining ill-defined domains; a literature survey. In *Proceedings of the workshop on intelligent tutoring systems for Ill-Defined domains at the 8th international conference on intelligent tutoring systems* (pp. 1–10). Citeseer.

Maner, W. (2002, April). Heuristic methods for computer ethics. *Metaphilosophy*, *33*(3), 339–365. Available from `http://www.blackwell-synergy.com/links/doi/10.1111%2F1467-9973.00231`

Masters, J., Madhyastha, T., & Shakouri, A. (2008, January). ExplaNet: a collaborative learning tool and hybrid recommender system for student-authored explanations. *Journal of Interactive Learning Research*, *19*(1), 51–74. Available from `http://go.editlib.org/p/21960`

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. Available from `http://doi.apa.org/getdoi.cfm?doi=10.1037/1082-989X.1.1.30`

McManus, M. M., & Aiken, R. M. (1995). Monitoring computer-based collaborative problem solving. *J. Artif. Intell. Educ.*, *6*(4), 307–336. Available from `http://portal.acm.org/citation.cfm?id=225653.225656`

McNamara, D., Kintsch, E., Songer, N., & Kintsch, W. (1996). Are good texts always better? interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, *14*(1), 1 – 43.

McNamara, T. (1990, June). Item response theory and the validation of an ESP test for health professionals. *Language Testing*, *7*(1), 52 –76. Available from `http://ltj.sagepub.com/content/7/1/52.abstract`

Miller, P. J. (2003). The effect of scoring criteria specificity on peer and Self-Assessment. *Assessment & Evaluation in Higher Education*, *28*(4), 383–94.

Mitrovic, A., Martin, B., & Mayo, M. (2002). Using evaluation to shape ITS design: Results and experiences with SQL-Tutor. *User Modeling and User-Adapted Interaction*, *12*(2), 243–279.

Muthukumarana, S. (2010). *Bayesian methods and applications using WinBUGS*. PhD dissertation, Simon Fraser University.

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

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

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

Paré, D., & Joordens, S. (2008, October). Peering into large lectures: examining peer and expert mark agreement using peerScholar, an online peer assessment tool. *Journal of Computer Assisted Learning*, *24*(6), 526–540. Available from `http://doi.wiley.com/10.1111/j.1365-2729.2008.00290.x`

Patterson, E. (1996, February). The analysis and application of peer assessment in nurse education, like beauty, is in the eye of the beholder. *Nurse Education Today*, *16*(1), 49–55. Available from `http://linkinghub.elsevier.com/retrieve/pii/S0260691796800931`

Pinkus, R. (1997). *Engineering ethics: balancing cost, schedule, and risk– lessons learned from the space shuttle*. New York: Cambridge University Press. Available from `http://www.loc.gov/catdir/description/cam027/96012332.htmlhttp://www.loc.gov/catdir/toc/cam024/96012332.html`

Pinkwart, N., Lynch, C., Ashley, K., & Aleven, V. (2008). Reevaluating LARGO in the classroom: Are diagrams better than text for teaching argumentation skills? In *9th international conference on intelligent tutoring systems.* Montreal.

Ploegh, K., Tillema, H. H., & Segers, M. S. (2009, June). In search of quality criteria in peer assessment practices. *Studies In Educational Evaluation*, *35*(2-3), 102–109. Available from `http://www.sciencedirect.com/science/article/B6V9B-4WKJ58M-1/2/b4841f775f7492f8642e7f7a7f5af555`

Popescu, O., Aleven, V., & Koedinger, K. (2005). Logic-Based natural language understanding for cognitive tutors. *Natural Language Engineering*, *1*(1), 1–15.

Ramachandran, L., & Gehringer, E. F. (2010). Automated metareviewing. Pittsburgh, PA.

Revelle, W., & Zinbarg, R. E. (2008, December). Coefficients alpha, beta, omega, and the glb: Comments on sijtsma. *Psychometrika*, *74*(1), 145–154. Available from `http://www.springerlink.com/index/10.1007/s11336-008-9102-z`

Rittel, H. W. J., & Webber, M. M. (1973, June). Dilemmas in a general theory of planning. *Policy Sciences*, *4*(2), 155–169. Available from `http://www.springerlink.com/`

content/m5050140x48140m3/

Robertson, J., Good, J., & Pain, H. (1998). BetterBlether: the design and evaluation of a discussion tool for education. *International Journal of Artificial Intelligence in Education*, *9*, 219–236. Available from http://aied.inf.ed.ac.uk/members98/archive/vol_9/robertson/paper.pdf

Russell, A., Cunningham, S., & George, Y. (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*(1), 60–79.

Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education*, *13*(2), 191–209.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional science*, *18*(2), 119–144.

Sanders, K., & Thomas, L. (2007, June). Checklists for grading object-oriented CS1 programs: concepts and misconceptions. *ACM SIGCSE Bulletin*, *39*, 166–170. (ACM ID: 1268834)

Schunn, C. D., Ashley, K. D., & Goldin, I. M. (Eds.). (n.d.). *Redesigning peer review interactions using computer tools* (Vol. Special Issue).

Scriven, M. (1966, March). *Social science education consortium. publication 110, the methodology of evaluation.* Available from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED014001

Sequeira, N. (2010, April). Peer graders get belated payday. *theVARSITY.ca*. Available from http://thevarsity.ca/articles/30540

Shepard, L. A. (2006). *Classroom assessment* (4th ed.; R. L. Brennan, Ed.). Praeger. Available from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED493398

Shute, V. J. (2008, March). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. Available from http://rer.sagepub.com/cgi/content/abstract/78/1/153

Sluijsmans, D. (1998). *The use of self-, peer- and co-assessment in higher education: a review of literature*. Heerlen: Educational Technology Expertise Centre Open University of the Netherlands.

Sluijsmans, D., & Prins, F. (2006). A conceptual framework for integrating peer assessment in teacher education. *Studies In Educational Evaluation*, *32*(1), 6–22. Available from http://linkinghub.elsevier.com/retrieve/pii/S0191491X0600006X

Soller, A. (2004). Computational modeling and analysis of knowledge sharing in collaborative distance learning. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research*, *14*(4), 351–381.

Spandel, V., & Stiggins, R. J. (1996). *Creating writers: Linking writing assessment and instruction* (2Sub ed.). Addison Wesley Publishing Company.

Steinhart, D. (2001). *Summary street: An intelligent tutoring system for improving student writing through the use of latent semantic analysis*. Unpublished doctoral dissertation, University of Colorado, Department of Psychology.

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

Strijbos, J., & Sluijsmans, D. (Eds.). (2010). *Unravelling peer assessment* (Vol. 20 (4)).

Suraweera, P., Mitrovic, A., & Martin, B. (2005). A knowledge acquisition system for constraint-based intelligent tutoring systems. In *12th international conference on artificial intelligence in education.* Amsterdam.

Suthers, D., Hundhausen, C., Dillenbourg, P., Eurelings, A., & Hakkarainen, K. (2001). Learning by constructing collaborative representations: An empirical comparison of three alternatives. In *First european conference on Computer-Supported collaborative learning* (p. 577–584). Maastricht, the Netherlands.

Thornton, A. E., Stilwell, L. A., & Reese, L. M. (2006, October). *The validity of law school admission test scores for repeaters: 2001 through 2004 entering law school classes* (LSAT Technical Report No. TR 06-02). Newtown, PA: Law School Admission Council. Available from http://lsacnet.org/LsacResources/Research/TR/TR-06-02.asp

Topping, K. J. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research*, *68*(3), 249–76.

Topping, K. J. (2010). Peers as a source of formative assessment. In H. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 61–74). New York: Routledge.

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

Toulmin, S. E. (2003). *The uses of argument* (Updated ed.). Cambridge University Press.

Turner, S. A. (2009). *Peer review in CS2: the effects on attitudes, engagement, and conceptual learning*. Unpublished doctoral dissertation, Virginia Tech, Blacksburg, VA. Available from `http://scholar.lib.vt.edu/theses/available/etd-08272009-003738/`

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, *16*.

VanLehn, K., Lynch, C., Schulze, K., Shapiro, J., Shelby, R., Taylor, L., et al. (2005). The andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence and Education*, *15*(3).

Voss, J., Greene, T., Post, T., & Penner, B. (1983). Problem solving skill in the social sciences. In G. Bower (Ed.), *Psychology of learning and motivation: Advances in research theory* (Vol. 17, pp. 165–213). Academic Press.

Voss, J., Hitchcock, D., & Verheij, B. (2006). Toulmin's model and the solving of Ill-Structured problems. In *Arguing on the toulmin model: New essays in argument analysis and evaluation.* Springer.

Voss, J., Post, T., Chi, M., Glaser, R., & Farr, M. (1988). On the solving of ill-structured problems. In *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*(4), 66–73.

Webb, N. M., Nemer, K. M., & Zuniga, S. (2002). Short circuits or superconductors? effects of group composition on high-achieving students' science assessment performance. *American Educational Research Journal*, *39*(4), 943.

Weber, G., & Brusilovsky, P. (2001). ELM-ART: an adaptive versatile system for web-based instruction. *International Journal of Artificial Intelligence in Education*, *12*(4), 351–384.

Weisberg, J. (2007, September). *The complete bushisms. weisberg, jacob (Ed.).* http://www.slate.com/id/76886/pagenum/all/. Available from `http://www.slate.com/id/76886/pagenum/all/`

Wightman, L. F., & Ramsey, H., Jr. (1998). *LSAC national longitudinal bar passage study* (Tech. Rep.). Newtown, PA: Law School Admission Council. Available from `http://lsacnet.org/LsacResources/Research/RR/Wightman-LSAC-98.asp`

Wiley, J., & Voss, J. (1999). Constructing arguments from multiple sources: Tasks that promote understanding and not just memory for text. *Journal of Educational Psychology*, *91*(2), 301–311.

Williamson, M. (1994). The worship of efficiency: Untangling theoretical and practical considerations in writing assessment. *Assessing Writing*, *1*(2), 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. Available from `http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED423541`

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

Xiong, W., Litman, D., & Schunn, C. D. (2010). Assessing reviewers' performance based on mining problem localization in Peer-Review data. In R. Baker, A. Merceron, & P. J. Pavlik (Eds.), *3rd international conference on educational data mining.* Pittsburgh, PA.

Yancey, K. B. (1999, February). Looking back as we look forward: Historicizing writing assessment. *College Composition and Communication*, *50*(3), 483–503. Available from `http://www.jstor.org/stable/358862` (ArticleType: research-article / Issue Title: A Usable Past: CCC at 50: Part 1 / Full publication date: Feb., 1999 / Copyright © 1999 National Council of Teachers of English)

Zeller, A. (2000). Making students read and review code. *SIGCSE Bull.*, *32*(3), 89–92. Available from `http://portal.acm.org/citation.cfm?id=343090`

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, revelle's β, and mcdonald's ω_h: their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123–133.