10-2010

Challenges to Informed Peer Review Matching Algorithms

Matthew Verleger
Utah State University, matthew.verleger@erau.edu

Heidi Diefes-Dux
Purdue University

Matthew W. Ohland
Purdue University

Mary Besterfield-Sacre
University of Pittsburgh

Sean Brophy
Purdue University

Scholarly Commons Citation

Verleger, M., Diefes-Dux, H., Ohland, M. W., Besterfield-Sacre, M., & Brophy, S. (2010). Challenges to Informed Peer Review Matching Algorithms. *Journal of Engineering Education, 99*(4). https://doi.org/10.1002/j.2168-9830.2010.tb01070.x

Challenges to Informed Peer Review Matching Algorithms

MATTHEW VERLEGER
Utah State University

HEIDI DIEFES-DUX, MATTHEW W. OHLAND
Purdue University

MARY BESTERFIELD-SACRE
University of Pittsburgh

SEAN BROPHY
Purdue University

ABSTRACT

BACKGROUND

Peer review is a beneficial pedagogical tool. Despite the preponderance of data instructors often have about their students, most peer review matching is done using simple random assignment.

PURPOSE

In fall 2008, a study was conducted to investigate the impact of an informed algorithmic assignment method when conducting peer review of Model-Eliciting Activities (MEAs). An algorithm was developed that utilized peer review calibration scores and Graduate Teaching Assistant scores to assign higher quality reviewers to teams in greater need of feedback. The algorithm showed no statistically significant impact on final response scores. This paper aims to examine the assumptions guiding the algorithm and how the breakdown of those assumptions helped identify possible changes necessary for successfully making informed peer review matches.

DESIGN/METHOD

An expert rater evaluated the solutions of 147 teams' responses to a particular implementation of MEAs in a first-year engineering course at a large mid-west research university. The evaluation was then used to verify the algorithm's operating assumptions, and how those assumptions broke down when compared to a randomly assigned control group.

RESULTS

Noting that an informed algorithm for assigning peer reviewers had no statistically significant impact on the quality of reviewed products, compared to simple random assignment, this work turns a lens on the assumptions of the informed peer review process to identify where that process broke down. As such, guidelines for developing informed peer review matching algorithms emerged from this investigation.

CONCLUSIONS

Conducting informed peer review matching requires significant alignment between evaluators and experts to minimize deviations from the algorithm's designed purpose.

Keywords: Calibration, Model-Eliciting Activity, Peer-Review

I. INTRODUCTION

Peer review has demonstrated that it can be a beneficial tool for helping students learn course content (Topping 1998). In recent years, the availability of computer-based tools to facilitate using peer review in the classroom has considerably increased (Chapman 2003; Gehringer 2001; Liu et al. 2001; Moreira and Silva 2003; Ngu, Shepherd, and Magin 1995; Sitthiworachart and Joy 2003; Trahasch 2004; Tsai et al. 2001). While the availability of tools has increased, the mechanism for matching reviewers to reviewees has not substantially evolved, with most peer review utilities relying on simple random assignment.

During the fall 2008 offering of the core introductory problem solving and computer tools at a Purdue University, an informed algorithm for making peer review assignments was developed and tested during a trained peer review of team solutions to a Model-Eliciting Activity (MEA). MEAs are open-ended, realistic, client-driven engineering problems where the artifact is a generalized written procedure for solving a problem. MEAs are designed to foster a student's mathematical modeling abilities (Zawojewski, Diefes-Dux, and Bowman 2008). For a subset of the students, the scores given by the graduate teaching assistant on the first draft were used to establish the degree of help a team needed on their second draft to improve their solution. Likewise, the level of agreement between the individual reviewers and an MEA rating expert on a training exercise were used to establish the degree of helpfulness a reviewer could provide on a subsequent peer review. One algorithm was explicitly designed to match more accurate reviewers to teams needing more assistance and, conversely, less accurate reviewers to teams needing less assistance. The purpose of the algorithm was to make reviewer-reviewee matches that were more beneficial to reviewees than could be provided through simple random assignment.

An individual with expertise evaluated the responses to an MEA of 74 teams assigned using an experimental algorithm and 73 teams assigned randomly as a control group to investigate the impact the algorithm would have on the quality of MEA Final Response scores. While students did improve the quality of their products as a result of the peer review, the students assigned using the algorithm showed no statistically significant difference in improvement on MEA Final Response scores when compared to the randomly assigned control group. The nature of informed peer review matching algorithms is such that there are often multiple assumptions fundamental to their use. Once the algorithm was found to have no impact, an analysis was made to investigate what happened with each of the following assumptions associated with this algorithm:

1) Students complete assigned work,
2) Teaching Assistants can grade MEAs accurately,
3) Accurate feedback in peer review is perceived by the reviewed team as being more helpful than inaccurate feedback,
4) Teaching Assistant scores on the first draft of an MEA can be used to accurately predict where teams will need assistance on their second draft, and
5) The error a peer review has in evaluating a sample MEA solution is an accurate indicator of the error they will have while subsequently evaluating a real team's MEA solution.

This research addresses the question “To what extent do the assumptions used in making informed peer review matches for the peer review of solutions to Model-Eliciting Activities decay?” Answers to this question will help identify what needs to be done to increase the validity of the assumptions so that informed peer review matching algorithms can be more effectively used and studied.

II. Literature Review

A. Model-Eliciting Activities

Model-Eliciting Activities (MEAs), first described by Lesh, Hoover et al. (2000), are realistic, open-ended, client-driven, team-based engineering problems designed to foster mathematical modeling abilities and reveal the development of modeling skills and construct development. Students work in teams to develop a procedure in the form of a memo to the direct user describing how to solve the problem, the limitations and assumptions of the solution, and the results from applying that solution to a given dataset.

As a pedagogical tool, MEAs have been explored within the engineering context (Diefes-Dux et al. 2006) and found to be effective mechanisms for helping students develop deeper understanding engineering content. They have primarily been studied in the context of very large first-year introductory engineering courses (Diefes-Dux and Imbrie 2008), though analysis on their effectiveness has also been done in higher discipline specific contexts (Bowman and Siegmund 2008). A significant portion of the current research on their use has centered around developing formative and summative feedback mechanisms that are reliable across a wide variety of evaluators (Diefes-Dux, Zawojewski, and Hjalmarson 2009).

One approach for evaluating MEAs is the use of peer review. Due primarily to the complexity of the problems, teams typically require multiple iterations of feedback and revision to converge on a high quality solution (Verleger and Diefes-Dux 2008). Through the use of peer review, teams are able to explore a wider variety of perspectives than could be seen through traditional student-grader interaction. The use of peer review also aligns with the recommendations and goals of

IEEE (2007) and ASCE (2008), as well as functioning as possible evidence for meeting ABET's a-k criterion (Zawojewski et al. 2008; Diefes-Dux et al. 2004).

B. Peer Review

Peer review has demonstrated that it can be a valuable tool for helping students learn, but it is not without its challenges. One of the greatest challenges in utilizing peer review in the classroom is getting students to accept that their peers can be considered valid sources for feedback and assessment. In Moreira and Silva's (2003) survey of 30 computer science undergraduate and 18 computer science graduate students, 10% expressed concerns about having clear judging criteria, specifically with the fact that the individuals evaluating them may not all be applying the criteria in the same way. Liu et al. (2001) showed similar results with 9% of the 143 third-year computer science students surveyed questioning the fairness of peer evaluations and having fellow students be in control of a significant portion of their overall grade.

While student's concerns are not entirely baseless, Billington (1997) found a statistically significant correlation of 0.80 in marks given on posters describing major ecosystem processes by final-year biology students compared to marks given by instructors. While there was high correlation between students and instructors, there was also a significant difference between the mean mark given by students and that given by the instructors. The average instructor grade was 59% while the average peer review mark was 70%. This suggests that despite students and instructors ranking the artifacts in approximately the same order while evaluating, they used different interpretation of the criteria to perform that evaluation. Cheng and Warren (1999) found that of the 16 scales (3 dimensions used in 3 different courses, each with an "overall" calculation for a 4x4 matrix) used in their study of the oral and written assignments produced by

51 first-year electrical engineering students, all of the scales showed positive score correlation between summative peer and instructor ratings, however only 11 of them were significant. Just as Billington noted in his study, while there was positive correlation in all of the scales, 11 of them (although not the same 11 significantly correlated scales) also showed significant differences in the means. This would imply that, while students would tend to rank products in the same order of quality as expert raters, they do not assign the same scores to those products.

Despite students' concerns about peer review, multiple studies indicate that the quality of the products being submitted improved subsequent to the review. Ballantyne et al. (2002) reported that the majority of the 939 respondents "agreed that peer assessment was an awareness-raising exercise because it made them consider their own work more closely, highlighted what they needed to know in the subject, helped them make a realistic assessment of their own abilities, and provided them with skills that would be valuable in the future." Similarly, Sitthiworachart and Joy (2003) indicated that 69% of first-year undergraduate students in computer science reported that they discovered mistakes in their own code while reviewing code written by their peers. Eighty percent of the students felt that seeing other students' work, both high- and low-quality work, was helpful for their learning.

In addition to the immediate skills provided by peer review, many researchers recognize the long-term benefits provided to reviewers. Boud (2000) posited that the focus of assessment as a whole must be rethought to promote lifelong learning skills. Learning to perform peer review and to respond to formative feedback given by both peer and self review are essential skills for succeeding in a continuous working world that doesn't assign an end-of-project grade. Teaching students how to perform peer review and how to utilize constructive criticism for improvement is essential for their future. Yet despite the obvious long-term benefits recognized by academic researchers, students are largely unfamiliar with peer review. Sitthiworachart and Joy (2004) reported that of their 215 first-year students taking a computer programming course, 89% of them had never experienced peer review prior to the start of the course. Guilford (2001) found that only 39% of undergraduate engineering students understood peer review as it related to scientific publishing. Ballantyne et al. (2002) indicated that only 10% of all the students studied recognized the value of peer review towards their future employment.

Despite numerous studies on effective peer review, there has been little research into how to make effective reviewer-reviewee mappings. Crespo, Pardo, and Kloos (2004) proposed the most ambitious attempt at producing higher quality reviewer-reviewee mappings. They developed an adaptive model that assigns students a "proficiency score". They then utilized a genetic algorithm to map reviewers to reviewees in such a way as to produce complementary proficiencies, i.e. high proficiency reviewers mapped to low proficiency reviewees and vice-versa. This strategy produced "promising experimental results", however no discussion of the educational impacts have since been published. What makes their model flawed is that it reduces a participant down to a single numerical value that assumes a reviewer is an equally capable reviewee. Furthermore, their approach was designed for individual-to-individual reviews, generally preventing teams from being used in the process.

III. Methods

Peer review using random assignment represents a baseline method for providing feedback. This research began as an attempt to improve peer review beyond that baseline. In that attempt, multiple necessary assumptions were made; and upon discovering that the resulting attempt was no different than the baseline, an investigation into those assumptions ensued. What follows is a description of the context in which this study took place, followed by a discussion of what happened in each of the five fundamental assumptions necessary for our implementation of informed peer review matching.

A. Course

This study was conducted during the fall 2008 semester of a required introductory problem-solving and engineering computer tools course at Purdue University. Students in the course attended a paired 110-minute lecture and 110-minute lab each week in sections of 120. Each lecture was taught by an engineering faculty member, while each lab was taught by a team of four graduate teaching assistants (TAs), with additional help from a group of two to four undergraduate teaching assistants. The course was limited to students currently enrolled in the first-year engineering program. Each TA in a given lab was responsible for seven or eight teams of three to four students each in that lab. TAs taught a total of two lab sections and were responsible for assessing and providing feedback to a total of 14 to 16 teams across their two labs.

B. Participants

Ten divisions of the class were taught during the fall 2008 semester. Each division was divided into four quadrants, with one TA assigned to each quadrant, for a total of 40 quadrants. Ten quadrants were selected and assigned one of four algorithms. Only two algorithms (Random and UON, each described subsequently) are included as part of this study. The remaining two algorithms utilized different metrics but a similar design philosophy as the UON algorithm but in preliminary analysis of TA scores did not appear to have as significant an impact on final scores as the UON algorithm. The assignment of an algorithm to a particular quadrant was strategically done to minimize issues related to course-specific deadlines and to be able to statistically resolve potential TA interactions. The gender and ethnic demographics of the students in the course and the sample population studied can be seen in Table 1.

	Whole Class		Sample Population	
	Female	Male	Female	Male
Caucasian American	189	702	91	357
Asian American	22	80	10	37
Spanish American	6	26	3	16
Other American	12	20	6	7
African American	10	20	6	11
American Indian	3	6	1	3
Unspecified or International	20	48	10	26
	262	902	127	457

Table 1 Gender and Ethnicity of Students (N = 1164)

Twenty graduate teaching assistants were involved with the course and acted as the primary contact for students during the lab time. As seen in Table 2, the majority of the TAs were male international students. Seven TAs had been a TA for the course in one or more prior semesters. Those seven TAs all had experience with MEAs during their prior experience as TAs for the course.

		Whole Class		Study Population	
		New	Returning	New	Returning
Male	International	7	5	6	4
	Domestic	1	2	1	1
Female	International	5	0	3	0
	Domestic	0	0	0	0
		13	7	10	5

Table 2. Graduate Teaching Assistant Attributes

C. Model Eliciting Activity (MEA)

1) Selected MEA: In this study, student team responses to the Purdue Paper Plane Challenge (PPPC) MEA were analyzed. The PPPC MEA, a variant of which is described in Wood, Hjalmarson et al. (2008), requires that students develop a procedure to assist the judges of a paper airplane contest in selecting the award winning team in four categories, Most Accurate, Best Floater, Best Boomerang, and Best Overall, given measurements of time in air, distance from target, and length of throw for multiple throws on a straight path and a boomerang path. This MEA represented 5% of a student's course grade.

2) Student Introduction to MEAs: Prior to the MEA, students were introduced to open-ended problem solving and MEAs during a lecture led by their lecture instructor. This lecture focused on an image tiling problem (also known as the Sports Equipment problem) (Verleger et al. 2009; Zawojewski, Diefes-Dux, and Bowman 2008) as a means of demonstrating expectations for MEA solutions. The lecture began with a discussion of mathematical models, model development, and MEAs as a vehicle for developing mathematical models to solve open-ended problems. Students then worked in teams to scope the problem and develop a generalized solution for shape fitting. The focus of the lecture is then shifted from simply solving the problem to developing a high quality, generalizable and share-able procedure which meets the client's immediate and future needs, with particular emphasis placed on understanding how the MEA Rubric dimensions relate to developing a high quality solution. The lecture was concluded by highlighting the role of the TA as a facilitator in the process (as opposed to the more traditional teacher role) and providing an overview of the sequence of MEA related events which would take place over the next five weeks of MEA 1 implementation.

3) MEA Feedback and Assessment Rubric: A full discussion on the development, reliability, and validity of the MEA Rubric can be found in Diefes-Dux, Zawojewski et al. (2010). MEAs were evaluated using two variants of a rubric developed specifically for assessing MEAs: (1) a Full MEA Feedback and Assessment Rubric and (2) a Reduced MEA Feedback and Assessment Rubric. The full version was used by the students; the reduced version was used by the TAs. Both variants were divided along three dimensions; Mathematical Model, Re-Usability / Modifiability, and Audience (Share-ability). Each dimension contained numeric and free response feedback items. The numerical items were the same for both the full and reduced versions of the MEA Rubric; however the free-response items were different. On the reduced MEA Rubric, there was one large text box for each dimension for written feedback, each with a general prompt. On the full version of the MEA Rubric, the prompts were more specific, asking for explicit items such as a summary of the mathematics used or recommendations for improving the rationales. These prompts were intended to help students engage in the review of a piece of student work by directing their attention to certain aspects of a team's solution so that better feedback would be generated. The students always used the full MEA Rubric. The TAs used the full MEA Rubric during their assessment training, but only used the reduced MEA Rubric for providing feedback on their first draft and final response. The reduced rubric was designed to decrease the level of summarizing of student work (and thereby time) required from the TAs in the feedback, as grading can between 30-60 minutes per team, depending on the procedure complexity. While the reduced version was not as specific in its requests as the full version, TAs were explicitly trained to consider all the elements of the full version in providing their feedback, even while using the reduced version.

Each quantitative MEA Rubric item was assigned point values which corresponded with levels of achievement. Items are divided into two categories; true/false items and mutually exclusive items. True/False items are assigned one of two possible point values depending on the item. Mutually exclusive items are items where multiple statements are presented, each with its own associated point value, and only a single statement may be selected. All of the items are presented in Table 3. The evaluator selects the statement which best describes the solution being evaluated. The score for each dimension is the minimum of the items in that dimension; the overall score is the minimum score of the three dimensions.

As an example, assume an evaluator selects “True” for the No Progress, Level 2 for the Mathematical Model Complexity, “False” for Data Usage, and “True” for Rationales. As the Mathematical Model dimensional score is calculated as a minimum of the dimension’s items, the Mathematical Model dimension score is a minimum of 0 (No Progress), 2 (Mathematical Model Complexity), 3 (Data Usage), and 4 (Rationales), resulting in a score of 0 for the Mathematical

Model Dimension. The overall score is then calculated as a minimum of the three dimensional scores, meaning that regardless of how this theoretical team performs on the remainder of the items, this team will receive an overall score of 0 (the lowest possible score) because of the 0 given for the Mathematical Model Dimension.

Minimums are taken for the dimensional and overall scores to encourage continuous improvement. This is also a philosophical stance – the student work is only as good as the weakest dimension.

Dim.	Item Label	Full Item Wording	Points	
Mathematical Model	No Progress	No progress has been made in developing a model. Nothing has been produced that even resembles a poor mathematical model. For example, simply rewriting the question or writing a "chatty" letter to the client does not constitute turning in a product.	True	0
			False	4
	Mathematical Model Complexity	The procedure fully addresses the complexity of the problem.	4	
		A procedure moderately addresses the complexity of the problem or contains embedded errors.	3	
		A procedure somewhat addresses the complexity of the problem or contains embedded errors.	2	
		Does not achieve the above level.	1	
	Data Usage	The procedure takes into account all types of data provided to generate results OR justifies not using some of the data types provided.	True	4
			False	3
	Rationales	The procedure is supported with rationales for critical steps in the procedure.	True	4
			False	3
Re-Usability/Modifiability	Re-Usability/ Modifiability	The procedure not only works for the data provided but is clearly re-usable and modifiable. Re-usability and modifiability are made clear by well articulated steps and clearly discussed assumptions about the situation and the types of data to which the procedure can be applied.	4	
		The procedure works for the data provided and might be re-usable and modifiable, but it is unclear whether the procedure is re-usable and modifiable because assumptions about the situation and/or the types of data that the procedure can be applied to are not clear or not provided.	3	
		Does not achieve the above level.	2	
Audience (Share-ability)	Results	Results from applying the procedure to the data provided are presented in the form requested.	True	4
			False	1
	Audience Readability	The procedure is easy for the client to understand and replicate. All steps in the procedure are clearly and completely articulated.	4	
		The procedure is relatively easy for the client to understand and replicate. One or more of the following are needed to improve the procedure: (1) two or more steps must be written more clearly and/or (2) additional description, example calculations using the data provided, or intermediate results from the data provided are needed to clarify the steps.	3	
		Does not achieve the above level.	2	
	Extraneous Information	There is no extraneous information in the response.	True	4
			False	3

Table 3. MEA Rubric – Numerical Items

4) MEA Sequencing: In the fall 2008 semester, MEA administration followed the sequence shown in Figure 1. In lab, students began to work through the sequence with the goal of producing the first draft of a memo to the client that contains a generalizable and shareable procedure. To achieve this goal, they were first given an individual warm-up activity in the form of a mock-newspaper article describing the PPPC MEA and a memo from the client explaining the problem. Upon reading this, students were asked to respond to three free-response questions asking who the client was, what the client needed, and what issues needed to be considered when producing a solution (Diefes-Dux and Salim 2009). After all team members had responded to the individual questions, the members came together as a team to develop a solution to the client's problem. The deliverable at the end of the lab period was the first draft of a memo to the client detailing the solution to the problem.

Following the lab, the graduate teaching assistants provided their student teams with feedback using the Reduced MEA Feedback and Assessment Rubric. After student teams received feedback on their first draft from their TA, the teams made revisions to their memo and submitted a second draft.

As part of the revision process for the first MEA, time was spent in lecture helping students understand how to interpret feedback. This was only done for the first MEA and was focused on helping students to frame their feedback in terms of how it impacted each of the MEA Rubric items. The structure of this lecture had faculty working through a list of common feedback statements for PPPC procedures and discussing how to interpret that feedback. Because the PPPC has been used multiple times in the course, typical responses and the problems associated with those types of responses were able to be anticipated. Likewise, the corresponding feedback was also anticipated. For example, students were shown the feedback "Why does the lowest mean win "Most Accurate"?" A brief discussion was led by the faculty member asking students to explain what they think this feedback means. The faculty member then explains how this is a prompt toward developing rationales within the procedure, and by extension improving their performance on the Rationales MEA Rubric item, by justifying the use of the "lowest mean" as opposed to the "highest mean" or some other statistical test. At the conclusion of this lecture, teams are told to review the feedback from their TA on their first draft and make appropriate revisions before submitting their second draft.

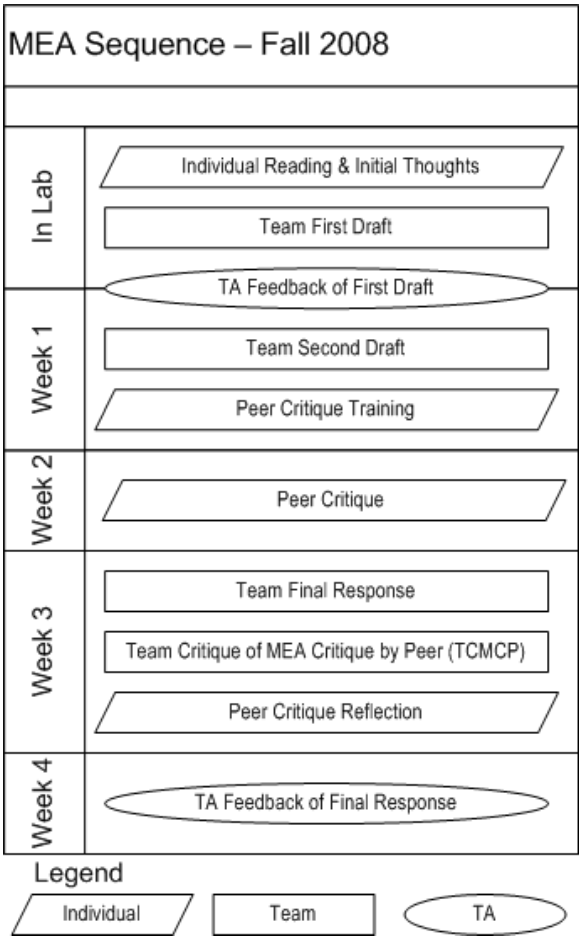

Figure 1. Fall 2008 MEA Sequence

The second draft entered a calibrated double-blind peer review. For this process, individual students used the Full MEA Rubric to evaluate a randomly selected prototypical student solution from a pool of five sample solutions. Prototypical work was selected by the second author from the database of student work from prior semesters. After being selected, the prototypical works were updated to remove any identifying information and to reflect any changes to the problem, including updating results to accommodate changes in the dataset. Care was taken to ensure that the essence of the memos did not change and that the memos appeared to come from a team of peers currently in the class. After students submitted their evaluation of the prototypical work, they were shown their review next to the second author's review of that same memo. They were asked to reflect on how they could improve their ability to review an MEA. During the week following the calibration, students were assigned a solution developed by a team of their peers to review using the same rubric as the calibration.

Because individuals review a team's procedure, each team typically received three or four peer reviews. Teams used these reviews, their own experiences in the review process, and any TA feedback from their first draft (which may or may not have been used in creating their second draft) to make final revisions to their memo. This final response to the client was submitted for evaluation and grading by the TA. Again, the TA used the Reduced MEA Rubric for this evaluation.

In the closing steps, teams evaluated the quality and helpfulness of the peer reviews they received by completing the Team Critique of MEA Critique by Peers (TCMCP). It consisted of targeted items, which can be seen in Table 5, broken down by MEA Rubric dimension evaluated on a 5 point “Strongly Disagree” to “Strongly Agree” Likert scale. This feedback was given back to the peer reviewers in an effort to increase the quality of the reviews they provided during the peer review stages of future MEAs. As students received this feedback after they had submitted their final response, it served no direct purpose toward the PPPC MEA. After providing this feedback to their reviewers, participants individually completed the Peer Review Reflection, reflecting on how comfortable and capable they were in the peer review process.

D. Graduate Teaching Assistant MEA Training

During the week prior to the start of the fall semester, TAs participated in an 8-hour training session split across two days to introduce them to MEAs. They began by working in ad-hoc teams on the PPPC MEA. This served as an opportunity for TAs to develop an understanding of the student experience at solving MEAs. After this, they were given an introduction to the underlying theory surrounding MEAs with a focus on the application of the six design principles of MEA development (Lesh and Doerr 2003) and the goals of using MEAs in the classroom. This presentation transitioned into the TAs discussing the challenges of evaluation and feedback on students’ solutions to open-ended problems. This discussion was then used as a platform for introducing the MEA Feedback and Assessment Rubric and the Instructors’ MEA Assessment/Evaluation Package. The package provides task-specific guidance for applying the MEA Rubric to a particular MEA. Using the MEA Rubric, TAs individually evaluated two pieces of student work. A discussion comparing and contrasting the TAs’ scores to the expert rater’s scores ensued. As a follow-up assignment to the training session, TAs were given electronic access a TA Professional Development Portal. The portal consisted of 5 pieces of student work for which the TAs were required to apply the MEA Rubric. The TAs submitted these reviews to the expert and received feedback on their strengths and weaknesses at applying the MEA Rubric. With feedback, TAs also received access to a side-by-side comparison of their reviews with the expert rater's reviews and were strongly encouraged to use the expert rater's reviews as guides for their own reviews.

E. Assignment Algorithm

All peer review assignments being considered were made using one of two algorithms:

- Random (N = 295 students on 73 teams)
- UON - Unweighted Overall Need (N = 289 students on 74 teams)

Students assigned using the random algorithm served as the control group, as this is the de-facto standard for making peer review assignments. Students assigned using the UON algorithm were considered an experimental group. At no point were the students made aware of the matching process or the quantitative values used to make those matches. Two additional algorithms were used for a portion of the class (N = 580 students on 148 teams), but not studied as part of this research.

1) *Random Assignment:* The reason random assignment has become the de-facto standard for making peer review assignments is that it does not rely on prior knowledge to make assignments. Because it does not utilize information about the reviewer or the reviewee, assignments can be made dynamically at the point in time that the reviewer starts the peer review process. As such, it is not affected if students fail to participate in the review process.

2) *Un-weighted Overall Need (UON):* To inform the UON algorithm, a set of base calculations were performed. TA first draft scores were used to assess the degree of assistance a team

needed. It was assumed that teams who scored lower (i.e., performed poorly) on the first draft would also need more assistance on their second draft, as they had more work to do to improve the quality of their solution than those teams who scored higher on the first draft. Coupled with this was the assumption that TA first draft scores were accurate, and thus could be effectively used as an indicator of a team's need for assistance.

Next, a comparison of the scores on the calibration exercise between the individual reviewers and the expert rater was performed. This was done to assess the accuracy of the reviewer. It was assumed that reviewers who were accurate were also providing more helpful feedback. Coupled with this was the assumption that calibration score accuracy was an accurate indicator of how well a reviewer would perform on the peer review exercise.

The UON algorithm attempts to holistically improve how students perform on the MEA by utilizing a calculation which is based on all three MEA Rubric dimensions. First, each team is assigned one of the most accurate overall reviewers. This step is done to ensure that every team receives at least one accurate review. The remaining reviewers are evenly divided between all of the teams, with the team needing the greatest assistance being reviewed by the most accurate of the remaining reviewers and the teams needing the least amount of assistance being matched to the least accurate reviewers. The UON algorithm attempts to provide the best help to those teams with the greatest need for assistance while not overly penalizing teams that need less assistance. By design, each team should receive an equal number of reviewers with an uneven distribution in review accuracy, assuming that all students who participate in the calibration also complete their peer review.

F. Evaluators

Two evaluators were used for this research. The first author re-evaluated the control and experimental groups. The second author provided the evaluations for the TA training and prototypical work used in the calibration (including the feedback seen by students during the "Comparison to Expert" step).

G. Reliability of Re-evaluation of PPPC Responses

To verify that the algorithm assumptions were valid, the first author re-evaluated all three drafts of both the control and experimental groups. The 147 teams in the Random (N = 73) and UON (N = 74) groups were randomly ordered and assessed without the author being aware of which algorithmic group each team was a member. All of the first drafts were assessed, followed by the second drafts, then the final responses. Re-evaluations took place over the span of 13 weeks.

For each of the two studied groups (Random and UON), 21 teams were randomly selected (for a total of 42 teams). Seven of those 21 (10% of the treatment group's total size) were randomly selected to have their first drafts re-evaluated a second time. Seven of the remaining 14 were randomly selected to have their second drafts re-evaluated a second time, while the final seven teams had their final responses re-evaluated a second time. This re-evaluation occurred 10 weeks after the original re-evaluations were completed. Spearman's Rho correlation coefficients (α = 0.05) were calculated between the author's original re-evaluation (the first evaluation having been done by the TAs and peer reviewers) and the second re-evaluation, as seen in Table 4 (A). All 12 items are sufficiently high enough for the author's evaluation to be considered acceptably reliable.

	(A)		(B)		(C)		(D)		(E)		
	Re-re-evaluation Reliability Correlations		Corr. between First and Second Draft Expert's Scores		Corr. between First Draft TA and Expert's Scores		Corr. of Error in Calibration versus Error in Peer Review		Increase in Avg. Error Between Calibration and Peer Review		
	N = 42		N = 147		N = 147		N = 449		N = 449		
Item/Dimension/Overall Score	Spearman Rho Correlation	Sig. (2-tailed *p*-value)	Spearman Rho Correlation	Sig. (2-tailed *p*-value)	Spearman Rho Correlation	Sig. (2-tailed *p*-value)	Spearman Rho Correlation	Sig. (2-tailed *p*-value)	t-test Sig. (2-tailed *p*-value)	F-Test Sig.	Increase in Average Error
Mathematical Model Dimension	0.66	0.000	0.25	0.003	-0.06	0.464	0.05	0.292	0.000	0.005	-0.053
No progress has been made in developing a model.	1.00	0.000	0.71	0.000	0.40	0.000	0.16	0.001	0.536	0.265	0.045
The procedure fully addresses the complexity of the problem.	0.63	0.000	0.22	0.007	-0.11	0.206	0.02	0.719	0.083	0.649	-0.085
The procedure takes into account all types of data provided to generate results OR justifies not using some of the data types provided.	0.78	0.000	0.38	0.000	0.24	0.004	0.05	0.248	0.049	0.834	0.065
The procedure is supported with rationales for critical steps in the procedure.	0.69	0.000	0.10	0.240	0.20	0.018	0.05	0.319	0.000	0.139	-0.245
Re-Usability / Modifiability Dimension	0.97	0.000	0.53	0.000	0.09	0.263	0.11	0.017	0.000	0.001	-0.261
The procedure not only works for the data provided but is clearly re-usable and modifiable.	0.97	0.000	0.53	0.000	0.09	0.263	0.11	0.017	0.000	0.001	-0.261
Audience (Share-ability) Dimension	0.80	0.000	0.31	0.000	0.41	0.000	0.09	0.052	0.000	0.057	-0.229
Results from applying the procedure to the data provided are presented in the form requested.	0.77	0.000	0.27	0.001	0.51	0.000	0.03	0.583	0.000	0.006	-0.575
The procedure is easy for the client to understand and replicate.	0.75	0.000	0.32	0.000	0.21	0.010	0.01	0.839	0.142	0.299	-0.058
There is no extraneous information in the response.	0.61	0.000	0.40	0.000	0.19	0.025	0.05	0.261	0.000	0.000	0.305
Overall Score	0.77	0.000	0.21	0.013	0.10	0.209	0.10	0.029	0.018	0.039	0.149

Table 4. Item/Dimension/Overall Score Calculations

H. Inter-Rater Reliability on PPPC Calibration Exercises

Two evaluators were used throughout the MEA process. The second author was responsible for providing the feedback on the five prototypical calibration samples, while the first author performed all of the remaining evaluations. To identify the inter-rater reliability, the five prototypical calibration samples were re-evaluated by the first author. Of the 60 total markings (5 samples * [8 items + 3 dimensions + an overall score]), both authors were in perfect agreement on 52 items (87%) and within one level on 7 (12%). One Audience Readability item (2%) had a difference of two levels.

IV. Results

Using the first author's evaluations of the control and experimental groups, the assumptions surrounding the algorithm's implementation are evaluated in the following sections and recommendations based on those findings are made subsequently.

A. Lower First Draft Scores Imply the Need for More Assistance during Peer Review

One of basic premises of the algorithm is that first draft scores provided by the TAs can be used to identify teams who will need greater assistance in revising their second draft. Shown in Table 4 (B), Spearman's Rho correlation coefficients ($\alpha = 0.05$) were calculated between the first author's first draft scores and second draft scores. The higher the correlation value, the more likely it is that the first draft scores can be used as an accurate predictor for the second draft scores

The only correlation value that would traditionally be considered "meaningfully significant" (i.e., $\rho > 0.60$) involves the "No progress has been made" item, though the trend seen across the other items should not be ignored. That the "No Progress" item showed the greatest correlation is not surprising, as, based on the first author's evaluation, only two teams did not obtain a level 4 in the first draft and only one of those teams allowed that problem to persist into the second draft. The remaining 145 teams all received scores of 4 on both the first draft and second draft.

Aside from the rationales item, which showed no statistically significant correlation between drafts, the remaining six MEA Rubric items all showed statistically significant low-level positive correlations, averaging 0.35. While none of these could individually be considered "meaningfully significant", the collectively positive values would seem to indicate that, while first draft scores can be used as a partial indicator of the degree of help a team may need on their second draft, other factors may need to be investigated to adequately predict the degree of assistance a team needs. These results would indicate that the validity of this assumption is weak, but is also not grossly violated.

B. TA First Draft Scores are Accurate

To justifiably use first draft scores as an indicator for the degree of assistance needed, the TA scores must be accurate. Spearman's Rho correlation coefficients ($\alpha = 0.05$) were again used to investigate this assumption. As shown in Table 4 (C), only six of the eight MEA Rubric items and only one of the three dimensions are statistically significantly correlated, and none of them correlate at a meaningfully significant level. The only MEA Rubric item that even approaches meaningfulness is the presence or absence of results, an item which has a clearly "correct" answer and should thus be easier for TAs to properly assess. A detailed analysis using this data of the problems TAs have in applying the MEA Rubric can be found in Diefes-Dux, Verleger et al., (2009).

C. Reviewers Who are More Accurate are Providing More Helpful Feedback

An implied assumption of the algorithm is that more accurate reviewers are providing more helpful feedback. To investigate this assumption, an error metric of the difference between peer review and first author's scores was calculated for each MEA Rubric item (N = 438 peer reviews of the 144 Random and UON assigned teams – three teams did not complete the TCMCP). The dimensional and overall accuracy was calculated as the sum of the squares of those calculated differences. For all items, dimensions, and the overall score, scores closer to zero are an indicator that the peer reviewer is more accurate. The TCMCP scores issued by the reviewed teams to the reviewers were used as an indicator of helpfulness of the peer review that team received, as this is the only mechanism for distinguishing the impact of individual reviews. This does assume the TCMCP is a valid indicator of helpfulness.

Stepwise linear regression analysis ($\alpha = 0.05$) was used to predict TCMCP scores based on the Peer Review Overall Error sum score. The results are presented in Table 5. None of the R^2 values are large, indicating that the overall error can only account for a small portion of the TCMCP variability. Collectively, these results would seem to indicate that, while more accurate feedback is at least partly beneficial for some items, overall accuracy does not represent a large contributing factor towards helpfulness.

TCMCP Item	β	*t* (437)	p	R^2
This peer critique included an honest attempt to use our procedure to produce a solution to the test case.	-0.257	-5.544	0.000	0.066
This peer critique was clearly written.	-0.188	-3.989	0.000	0.035
The Mathematical Model portion of this peer critique was a fair and accurate assessment of our team's procedure.	-0.188	-4.004	0.000	0.035
The Audience (Share-ability) portion of this peer critique was a fair and accurate assessment of our team's procedure.	-0.144	-3.039	0.003	0.021
The Re-usability/Modifiability portion of this peer critique was a fair and accurate assessment of our team's procedure.	-0.084	-1.766	0.078	0.007
The Mathematical Model portion of this peer critique resulted in our making substantive changes to our MEA solution.	-0.051	-1.076	0.283	0.003
The Audience (Share-ability) portion of this peer critique resulted in our making substantive changes to our MEA solution.	-0.031	-0.639	0.523	0.001
The Re-usability/Modifiability portion of this peer critique resulted in our making substantive changes to our MEA solution.	-0.014	-0.299	0.765	0.000

Table 5. Results of TCMCP Prediction from Peer Review Overall Error

D. Calibration is an Accurate Indicator of Peer Review Accuracy

Similar to the assumption that TA first draft scores are an accurate indicator of the degree of help a team needs, there is an equivalent assumption that calibration accuracy is a good indicator of how accurate a peer reviewer will perform on the actual peer review. Saterbak and Volz (2008) found students do align closely with instructors during both calibration and peer review, but only after significant revisions were made to both the prototypical works and the rubric. The need for those revisions highlights how this assumption is highly context dependent. For this study, the authors had reason to believe that this assumption would be valid based on the fact that calibration and peer review were explicitly designed to be nearly identical, with the only indicator that the calibration exercise did not involve rating a peer team being the use of the term “Calibration” instead of the term “Peer Review”.

Spearman’s Rho correlations ($\alpha = 0.05$) were calculated between error on calibration and the error in peer review for the 449 peer reviews of the 147 teams. As seen in Table 4 (D), there is very little correlation between the degree of error a reviewer makes on the calibration exercise and the error they make on the actual peer review, with none of the items being meaningfully significant. This would indicate that calibration accuracy is a poor predictor of peer review accuracy. Further analysis found that calibration had a positive, though non-uniform, impact on peer review error, reducing the gross error by a statistically significant 19%. As seen in Table 4 (E), paired t-tests and F-tests ($\alpha = 0.05$) were calculated for all eight MEA Rubric items between the error measurement for the calibration and the error measurement for the peer review.

E. Students Will Complete their Assigned Tasks

The nature of informed algorithms is such that assignments must typically be made in advance, which means that individuals must be assigned based on the assumption that they will be completing their assigned work. One of the most attractive attributes of random assignment is that reviewer-reviewee assignments can typically be made dynamically on demand, preventing the uneven distributions of reviewers that can occur in informed algorithms when assigned reviewers do not complete their review.

Only 1049 (90%) of the 1164 students enrolled in the course completed the calibration. Because participants who do not complete calibration are automatically prevented from participating in the peer review, they are not assigned a team to review. Only 84.3% (885/1049, 76% of the overall course enrollment) of those individuals who were assigned a team to peer review completed the assignment.

V. Discussion and Suggestions for Applying Informed Peer Review

Despite the breakdown of the assumptions resulting in the informed algorithm not performing as desired, a number of important lessons emerged which can help guide future research on calibration and informed peer review.

A. Calibration

It was expected that students, as a result of calibration, would improve the accuracy of their peer reviews. One of the assumptions for the algorithm was that students who were less accurate than their peers on calibration would typically continue to be less accurate than their peers on the peer review. This assumption, as discussed above, was violated more than expected. However, in spite of this, the average sum of squares error decreased by a statistically significant 19% between calibration and peer review ($p < 0.001$, N = 449). The implication is that calibration was typically helping improve accuracy, though not in a uniform enough manner to make peer review errors reliably predictable.

As this research demonstrated, under certain circumstances, calibration can reduce the error a reviewer makes on subsequent peer reviews. The most likely contributor that allowed calibration to be successful was how closely it resembled the actual peer review. Multiple individuals, when asked in their Peer Review Reflection to comment on the challenges encountered during the peer review process, made reference to “the first team” or “both teams”, demonstrating that many individuals were possibly not aware that the review they were completing of the calibration's prototypical procedure was not for a peer team.

B. Informed Peer Review

The UON algorithm studied in this research utilized the sum of the eight MEA Rubric items to establish reviewee need and reviewer skill levels. The algorithm's functionality was built around five basic assumptions:

1) Students complete assigned work,
2) TA's can grade MEAs accurately,
3) Accurate feedback in peer review is perceived by the reviewed team as being more helpful than inaccurate feedback,
4) First Draft scores can be used to accurately predict Second Draft areas of need, and
5) Calibration error is an accurate indicator of Peer Review error.

For our implementation, all five of the assumptions lacked the validity necessary for the algorithm to operate as desired. The first two assumptions could potentially be resolved through changes to the MEA implementation and the MEA Rubric and training process. The third assumption, based on the discussion found above, is at least partially valid, though additional work is needed to better understand what makes for the most helpful feedback.

The last two assumptions are specific to the UON algorithm, and while neither could be considered reasonably valid for this research, both are predictive modeling problems that are specific to the UON algorithm's implementation. Their breakdown represents a failure in the UON algorithm, but does not close the door for other informed algorithms.

Certainly a limitation of this study is that there are other assumptions that could be impacting the results. Three assumptions for which the authors currently lack either the data or the resources needed to properly validate are: (1) that students are putting forth the same level of effort on all three attempts at developing a solution, (2) that improvements in the second draft and final responses are the result of students having received feedback and not purely a result of students revisiting their work, and (3) that the quality of the numerical responses is highly correlated to the quality of the qualitative feedback a reviewer provides. Any of these assumptions could potentially have an impact on how students progress through the MEA sequence and the quality of the work they are producing. Each also represents a larger set of research questions dealing with (1) the motivation to always produce high quality work, (2) a team's ability to self-evaluate their work and improve it without external influences, and (3) a student’s quantitative versus qualitative evaluation skills and the correlation between those two skills.

The most critical aspect to the future success of informed matching algorithms is in fully identifying and minimizing the number of critical assumptions. In the case of the UON algorithm, assumptions two and five were critical to the basic functionality of the algorithm and the degree of violation represents a breakdown in how the algorithm functioned. Because the algorithm relied on these two independent assumptions, both of which exhibited non-trivial degrees of decay, there was not a large enough subset of participants for which the assumption held true to do an independent analysis of the algorithm’s impact. Improving the conditions so that the impact of these assumptions is minimized could lead to an increased ability to make sure that teams are being assigned the reviewer that will best be able to help them improve the quality of their work.

C. Implications for Engineering Educators

In an age of growing class sizes and shrinking budgets, peer review is an attractive mechanism for instructors to use to provide feedback to even the largest classes without a corresponding increase in time or teaching staff. While the premise of assigning better reviewers to those reviewees that need more help seems like an important and viable method for maximizing the value of peer review, the practice of doing so proves more complex than anticipated. Indeed, the act of making those assignments demonstrated to be no better than simple random assignment.

For engineering educators, this presents a critical lesson that, while peer review can be beneficial, blind random assignment still represents the easiest and fairest mechanism for making reviewer/reviewee assignments. The UON algorithm generally tried to pair the highest quality reviewers to the teams with the greatest need, and likewise it paired the lowest quality reviewers to the teams with the least amount of need. To combat this clearly unfair bias, the algorithm first assigned each team a single high quality reviewer before making the biased informed assignments. The fatal flaw was that, because the mechanism for predicting reviewer quality was not accurate enough, the initial round of assignments did not guarantee that all teams received at least one high quality reviewer. While random assignment does not guarantee that all teams will receive a high quality reviewer, it does not have a clear bias against receiving one. To develop, explore, and ultimately reap the potential benefits of informed peer review matching algorithms, means of accurately predicting reviewer quality are needed.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant Nos. DUE 0535678 and EEC 0835873.

REFERENCES

American Society of Civil Engineers. 2008. 2007-2008 Policies & Priorities. Washington, D.C.: ASCE Washington Office.

Ballantyne, R., K. Hughes, and A. Mylonas. 2002. Developing Procedures for Implementing Peer Assessment in Large Classes Using an Action Research Process. *Assessment and Evaluation in Higher Education* 27 (5):427-441.

Billington, H. L. 1997. Poster Presentations and Peer Assessment: Novel Forms of Evaluation and Assessment. *Journal of Biological Education* 31 (3):218-220.

Boud, D. 2000. Sustainable Assessment: Rethinking Assessment for the Learning Society. *Studies in Continuing Education* 22 (2):151-167.

Bowman, K., and T. Siegmund. 2008. Designing Modeling Activities for Upper-Level Engineering Classes. In *Models and Modeling in Engineering Education: Designing Experiences for All Students*, edited by J. S. Zawojewski, H. Diefes-Dux and K. Bowman. Rotterdam, The Netherlands: Sense Publishers.

Chapman, O.L. 2007. *Calibrated Peer Review - The White Paper: A Description of CPR* [PDF], August 10 2003 [cited November 28 2007]. Available from http://cpr.molsci.ucla.edu/cpr/resources/documents/misc/CPR_White_Paper.pdf.

Cheng, W., and M. Warren. 1999. Peer and Teacher Assessment of the Oral and Written Tasks of a Group Project. *Assessment and Evaluation in Higher Education* 24 (3):301-314.

Crespo, R. M., A. Pardo, and C. D. Kloos. 2004. An Adaptive Strategy for Peer Review. In *34th ASEE/IEEE Frontiers in Education Conference*. Savannah, GA.

Diefes-Dux, H., K. Bowman, J. S. Zawojewski, and M. Hjalmarson. 2006. Quantifying Aluminum Crystal Size Part 1: The Model-Eliciting Activity. *Journal of STEM Education* 7 (1 & 2):51 - 63.

Diefes-Dux, H., and P.K. Imbrie. 2008. Modeling Activities in a First-Year Engineering Course. In *Models and Modeling in Engineering Education: Designing Experiences for All Students*, edited by J. S. Zawojewski, H. Diefes-Dux and K. Bowman. Rotterdam, The Netherlands: Sense Publishers.

Diefes-Dux, H., T. Moore, J. S. Zawojewski, P.K. Imbrie, and D. Follman. 2004. A Framework for Posing Open-Ended Engineering Problems: Model-Eliciting Activities. In *Frontiers in Education (FIE)*. Savannah, GA.

Diefes-Dux, H., and A. Salim. 2009. Problem Identification during Model-Eliciting Activities: Characterization of First-Year Students' Responses. In *2009 Research in Engineering Education Symposium (REES 2009)*. Queensland, Australia.

Diefes-Dux, H., M. Verleger, J. S. Zawojewski, and M. Hjalmarson. 2009. Multi-Dimensional Tool for Assessing Student Team Solutions to Model-Eliciting Activities. In *American Society for Engineering Education National Conference (ASEE 2009)*. Austin, TX.

Diefes-Dux, H., J. S. Zawojewski, and M. Hjalmarson. 2009. Using Educational Research in the Design of Evaluation Tools for Open-Ended Problems. *International Journal of Engineering Education* (In Press).

Diefes-Dux, H., J. S. Zawojewski, and M. Hjalmarson. 2010. Using Educational Research in the Design of Evaluation Tools for Open-Ended Problems. *International Journal of Engineering Education* (In Press).

Gehringer, E. F. 2001. Electronic Peer Review and Peer Grading in Computer-Science Courses. In *Thirty-second SIGCSE technical symposium on Computer Science Education* Charlotte, North Carolina: ACM.

Guilford, W. H. 2001. Teaching Peer Review and the Process of Scientific Writing. *Advances in Physiology Education* 25 (3):167-175.

IEEE. 2009. *IEEE Envisioned Future* 2007 [cited June 24 2009]. Available from http://www.ieee.org/portal/cms_docs_iportals/iportals/aboutus/envisioned_future.pdf.

Lesh, R. A., and H. M. Doerr, eds. 2003. *Beyond Constructivism: Models and Modeling Perspectives on Mathematics Problem Solving, Learning, and Teaching*. Mahwah, NJ: Lawrence Erlbaum.

Lesh, R. A., M. Hoover, B. Hole, A. Kelly, and T. Post. 2000. Principles for Developing Thought Revealing Activities for Students and Teachers. In *Handbook of Research Design in Mathematics and Science Education*, edited by A. Kelly and R. A. Lesh. Mahwah, NJ: Lawrence Erlbaum.

Liu, E. Z., S. S. J. Lin, C. Chiu, and S. Yuan. 2001. Web-Based Peer Review: The Learner as Both Adapter and Reviewer. *IEEE Transactions on Education* 44 (3):246-251.

Moreira, D.A., and E. Q. Silva. 2003. A Method to Increase Student Interaction Using Student Groups and Peer Review over the Internet. *Education and Information Technologies* 8 (1):47-54.

Ngu, A. H. H., J. Shepherd, and D. Magin. 1995. Engineering Peers: Computer-Assisted Approach to the Development of Peer Assessment System. Paper read at Research and Development in Higher Education 18: Blending Tradition and Technologies, HERDSA 95, at Rockhampton, Queensland.

Saterbak, A., and T. Volz. 2008. Implementing Calibrated Peer Review™ to Enhance Technical Critiquing Skills in a Bioengineering Laboratory. In *American Society for Engineering Education (ASEE) National Conference*. Pittsburgh, PA.

Sitthiworachart, J., and M. Joy. 2003. Deepening Computer Programming Skills by Using Web-based Peer Assessment. Paper read at 4th Annual Conference of the LTSN Centre for Information and Computer Sciences, at NUI Galway, Ireland.

Sitthiworachart, J., and M. Joy. 2003. Web-based Peer Assessment in Learning Computer Programming. In *3rd IEEE International Conference on Advanced Learning Technologies*. Athens, Greece.

Sitthiworachart, J., and M. Joy. 2004. Effective Peer Assessment for Learning Computer Programming. In *9th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education* Leeds, United Kingdom.

Topping, K. 1998. Peer Assessment Between Students in Colleges and Universities. *Review of Educational Research* 68 (3):249-276.

Trahasch, S. 2004. Towards a Flexible Peer Assessment System. In *Fifth International Conference on Information Technology Based Higher Education and Training (ITHET) 2004*. Istanbul.

Tsai, C., E. Z. Liu, S. S. J. Lin, and S. Yuan. 2001. A Networked Peer Assessment System Based on a Vee Heuristic. *Innovations in Education and Teaching International* 38 (3):220-230.

Verleger, M., and H. Diefes-Dux. 2008. Impact of Feedback and Revision on Student Team Solutions to Model-Eliciting Activities. Paper read at American Society for Engineering Education (ASEE) National Conference, at Pittsburgh, PA.

Verleger, M., H. Diefes-Dux, C. Liguore, and B. Eick. 2009. *Image Tiling : Problems : Sgmm : Research - School of Engineering Education, Purdue University* 2009 [cited May 26 2009]. Available from https://engineering.purdue.edu/ENE/Research/SGMM/Problems/ImageTiling/SinglePage.

Wood, T., M. Hjalmarson, and G. Williams. 2008. Learning to Design in Small Group Mathematical Modeling. In *Models and Modeling in Engineering Education: Designing Experiences for All Students*, edited by J. S. Zawojewski, H. Diefes-Dux and K. Bowman. Rotterdam, The Netherlands: Sense Publishers.

Zawojewski, J. S., H. Diefes-Dux, and K. Bowman, eds. 2008. *Models and Modeling in Engineering Education: Designing Experiences for All Students* Sense Publishers.

Zawojewski, J. S., M. Hjalmarson, K. Bowman, and R. A. Lesh. 2008. A Modeling Perspective on Learning and Teaching Engineering Education. In *Models and Modeling in Engineering Education: Designing Experiences for All Students*, edited by J. S. Zawojewski, H. Diefes-Dux and K. Bowman. Rotterdam, The Netherlands: Sense Publishers.

Authors' Biographies

Matthew A. Verleger is a Post-Doctoral Researcher in Purdue University's School of Engineering Education. He received his B.S. in Computer Engineering in 2002, M.S. in Agricultural and Biological Engineering in 2005, and Ph.D. in Engineering Education in 2009 all from Purdue. Throughout that time, he has been a teaching assistant in Purdue first-year engineering problem solving and computer tools courses. His research focuses on Model-Eliciting Activities and the first-year experience. Contact information includes: School of Engineering Education, Purdue University, 701 West Stadium Drive, West Lafayette, IN 47907-2045 USA. Email: mverleg1@purdue.edu.

Heidi A. Diefes-Dux is an Associate Professor in the School of Engineering Education at Purdue University. She received her B.S. and M.S. in Food Science from Cornell University and her Ph.D. in Food Process Engineering from the Department of Agricultural and Biological Engineering at Purdue University. Since 1999, she has been a faculty member within the First-Year Engineering Program at Purdue, the gateway for all first-year students entering the College of Engineering. She coordinated (2000-2006) and continues to teach in the required first-year engineering problem solving and computer tools course, which engages students in open-ended problem solving and design. Her research focuses on the development, implementation, and assessment of model-eliciting activities with realistic engineering contexts. She is currently the Director of Teacher Professional Development for the Institute for P-12 Engineering Research and Learning (INSPIRE). Contact information includes: School of Engineering Education, Purdue University, 701 West Stadium Drive, West Lafayette, IN 47907-2045 USA Email: hdiefes@purdue.edu

Matthew W. Ohland is an Associate Professor in Purdue University's School of Engineering Education. He received his Ph.D. in Civil Engineering from the University of Florida in 1996. Dr. Ohland is the Past President of Tau Beta Pi and has delivered over 100 volunteer seminars as a facilitator in the society's award-winning Engineering Futures program. He is Chair of the Educational Research and Methods division of the American Society for Engineering Education and a member of the Administrative Committee of the IEEE Education Society. His research on the longitudinal study of engineering student development, peer evaluation, and high-engagement teaching methods has been supported by over $9.1 million in funding from NSF and Sloan. Contact information includes: School of Engineering Education, Purdue University, 701 West Stadium Drive, West Lafayette, IN 47907-2045 USA. phone (765) 496-1316, fax (765) 494-5819, or e-mail ohland@purdue.edu.

Mary Besterfield-Sacre is an Associate Professor and Fulton C. Noss Faculty Fellow in Industrial Engineering at the University of Pittsburgh. Her principal research interests are in engineering education evaluation methodologies, planning and modeling the K-12 educational system. Her current focus areas lie in measuring aspects of technical entrepreneurship, problem solving, and design. She received her B.S. in Engineering Management from the University of Missouri - Rolla, her M.S. in Industrial Engineering from Purdue University, and a Ph.D. in Industrial Engineering at the University of Pittsburgh. She is a former associate editor for the *Journal of Engineering Education* and current associate editor for the *Advances in Engienering Education*. Email: mbsacre@engr.pitt.edu.

Sean P. Brophy is an assistant professor in the School of Engineering Education at Purdue University and research director for INSPIRE P-12 Engineering Education. Dr. Brophy is currently conducting research on precursors to engineering thinking in young children. This work aligns well with his other research interests relate to using simulations and models to facilitate students' understanding of difficult concepts within engineering. Address: School of Engineering Education, Purdue University, 701 West Stadium Avenue, Neil Armstrong Hall of Engineering, West Lafayette, Indiana 47907-2045; telephone: (765) 496.3316; fax: (765) 494.5819; e-mail: sbrophy@purdue.edu.