Go With Your Gut! – Using Low-Time-Investment Evaluations of Student Work for Identifying High versus Low Quality Responses

Dr. Matthew A. Verleger, Embry-Riddle Aeronautical Univ., Daytona Beach

Matthew Verleger is an Associate Professor of Engineering Fundamentals at Embry-Riddle Aeronautical University in Daytona Beach, Florida. His research interests are focused on using action research methodologies to develop immediate, measurable improvements in classroom instruction and the use of Model-Eliciting Activities (MEAs) in teaching students about engineering problem solving. Dr. Verleger is an active member of ASEE. He also serves as the developer and site manager for the Model-Eliciting Activities Learning System (MEALearning.com), a site designed for implementing, managing, and researching MEAs in large classes.

Go With Your Gut! – Using Low-Time-Investment Evaluations of Student Work for Identifying High versus Low Quality Responses

Abstract

Background
Peer review can be a beneficial pedagogical tool for providing students both feedback and varied perspectives on their work. Despite being a valuable tool, the best mechanism for assigning reviewers to reviewees is still often blind random assignment. While better mechanisms must exist, they necessarily rely on having some prior knowledge about the work being reviewed.

Purpose (Hypothesis)
The purpose of this paper is to present the findings from an effort to classify student team performance on Model-Eliciting Activities (MEAs) using a trained reviewer's gut instinct about the quality of the work.

Design/Method
MEAs are realistic, open-ended, client-driven engineering problems where teams of students produce a written document describing the steps of how to solve the problem. Using an archival data set, nearly 450 MEA solutions were evaluated by two trained student researchers in approximately two minutes per solution. Their evaluations are compared against other, more detailed, analyses of the solutions to identify if their evaluations are sufficiently accurate enough to use as baseline data for making peer review matching decisions with a comparatively miniscule investment in time.

Results
Results indicate that both researchers performed less accurately than computer-based classification but were largely consistent with the more detailed evaluations conducted by teaching assistants.

Conclusion
The conclusion this research made was that gut reaction based classification was not wholly sufficient to address the needs for informed peer review matching. The results may be useful as an additional data source for computer-based classification to reduce the amount of training or to increase accuracy.

Introduction

Peer review is a cornerstone of the modern scientific process. It is meant to act as a gateway, allowing good research through while filtering out junk science; to separate the wheat from the proverbial chaff. Yet many scientists, academics, and even the US Supreme Court agree that peer review, while essential to the scientific process, is far from a perfect system. In the 1993 Supreme Court case "Daubert v. Merrell Dow Pharmaceuticals", Justice Blackmun wrote that it was the opinion of the court that "Publication (which is but one element of peer review) is not a sine qua non of admissibility; it does not necessarily correlate with reliability… But submission to the scrutiny of the scientific community is a component of "good science," in part because it increases the likelihood that substantive flaws in methodology will be detected" (Blackmun, 1993). Effectively, the court recognized that, while peer review is good for science as a whole, it does not necessarily work correctly all the time. The problem with peer review is that it is a theoretically sound process that can easily fail apart on implementation. It is a methodology whose success is heavily dependent on having the most appropriate reviewer for the situation providing the right review.

When used in the classroom, peer review can be a useful tool for providing students with additional feedback and perspectives while not significantly increasing the workload of graders or course administrators. Much like its research counterpart, classroom peer review suffers from issues related to proper reviewer selection. Ballantyne, Hughes, and Mylonas (2002) noted multiple studies describing how students do not necessarily believe that they or their peers are capable reviewers. In their study, 40% of the participants agreed that their peers could not fairly assess their work. Fundamentally, this is a case of "one bad apple spoils the bunch." When a student receives even a single poorly formed peer review, their attitude towards all their received reviews can be spoiled. While this issue can be reduced through significant training and careful rubric design, the need for understanding effective reviewer matching is essential for improving the long-term effectiveness and implementation of peer review in the classroom.

Prior research by the author (Verleger, 2014; Verleger, Diefes-Dux, Ohland, Besterfield-Sacre, & Brophy, 2010; Verleger, Rodgers, & Diefes-Dux, 2016) has highlighted some of the complexities of viewing the reviewer-reviewee relationship as a variable that can be adjusted and explored to different effects. A key outcome of that research was an understanding that, to make effective peer review assignments, the work being reviewed must be at least somewhat accurately classifiable prior to peer review based on the amount of help a reviewee needs. This paper reports on using trained student evaluators to provide a "gut reaction" evaluation for classification purposes. Their results will be compared against some of the other evaluations completed as part of this research.

Background

Peer Review

Editorial peer review has been a cornerstone component of scientific achievement since the mid-1950's (Burnham, 1990). Despite its tremendous post-war boom to become the de facto standard for scientific and technical publications and the largely similar goal of providing feedback to improve quality, peer review is still only moderately used as a pedagogical tool within the higher education classroom. The single greatest hindrance toward utilizing peer review in the classroom is getting students to accept that it is a viable source for feedback and assessment. Ballantyne et al. (2002) undertook a study of 1,654 first- and second-year students spanning three semesters studying four different courses. Despite continual efforts based on feedback from students and faculty to improve the process, some of the attitudes of the participants towards the process remained relatively consistent throughout the entire study. In a follow-up survey given to all 1,654 students, 734 gave a response to a question regarding the worst aspect of the peer review process. 31% of those the 734 responders (14% of the total) mentioned concerns about the competency of either themselves or their peers.

Despite a lack of confidence in the quality of the review, the majority of students report liking peer review. Of the 30 undergraduate computer science students in their study, Moreira and Silva (2003) found that 77% of the students indicated that they liked peer review, and another 13% were neutral towards peer review. Liu et al. (2001) reported that 64% of participants viewed peer review as beneficial and effective for learning. Despite students' concerns about peer review, multiple studies indicate that it improves the quality of the products being submitted subsequent to the review. 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 was helpful for their learning. 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."

In addition to the immediate skills provided by participating in 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 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 long-term benefits recognized by academia, 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 not ever 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, though 35% of the education students in their study recognized the long-term value.

Despite the value of peer review, the best approach that has been used for matching reviewers to reviewees is random assignment. Because it requires no prior knowledge of either the reviewer or reviewee, it is the easiest methodology to implement. To make more informed choices, some amount of quality prediction must be done to identify who needs the most help. Verleger (2010) used teaching assistant scores on an early draft as an indicator. Another approach Verleger attempted was to use natural language processing to computationally predict quality (2014). Both approaches demonstrated flaws that this work attempted to explore.

Model-Eliciting Activities (MEAs)

This research is being explored in the context of Model-Eliciting Activities. Model-Eliciting Activities (MEA) are realistic, client-driven, open-ended problems that are designed to be both model-eliciting and thought-revealing (Lesh, Hoover, Hole, Kelly, & Post, 2000). They require students to mathematize (e.g., quantify, organize, dimensionalize) information in context. An engineering-based MEA requires that students be provided with a realistic problem that a client needs solved. The solution of an MEA requires the development of one or more mathematical, scientific, or engineering concepts that are unspecified by the problem – students must grapple with their existing knowledge to develop a generalizable mathematical model to solve the problem. The point is for students to be involved in the creation of the initial ideas underlying the concept or system, thus establishing the need and motivation to go through cycles of expressing their initial ideas, testing, and refining them. An MEA creates an environment where skills such as communication, verbalization, and an ability to work cooperatively and collaboratively are valued. Carefully constructed MEAs can begin to prepare students to communicate and work effectively in teams; to create, adopt and adapt conceptual tools; to construct, describe, and explain complex systems; and cope with complex systems. The attributes of MEAs support the development of the abilities and skills required of graduates of accredited engineering programs as stated in ABET Criterion 3 a to k (ABET, 2013).

Methodology

Table 1. MEA Rubric (Numeric Items)

Dim.	Item Label	Full Item Wording	Points	
Mathematical Model	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

MEA Description

The MEA used in this research is referred to as the Paper Airplane MEA. Complete details about the problem can be found in Wood, Hjalmarson, & Williams (2008). In broad terms, the MEA has teams of students using data from multiple throws in a paper airplane contest to develop a procedure to help judges award four prizes in the contest; Most Accurate, Best Floater, Best Boomerang, and Best Overall.

MEA Evaluation Rubric

A complete description of the MEA Rubric can be found in Diefes-Dux, Zawojewski, & Hjalmarson (2010). The rubric consists of 7 numeric items and 9 free-response text items. The numeric items are shown in Table 1. Five of the free-response items are targeted questions designed to help the reviewer identify specific attributes of high quality solutions present in the solution being reviewed, while the other four items ask the reviewer to provide feedback on how to improve specific attributes of solution.

Data

The MEAs being classified in this research were initially collected in 2008 at a large, mid-west, public, R1 institution. 147 teams of 3-4 students each completed three solution iterations of the Paper Airplane MEA. An initial draft was submitted and given feedback from a trained teaching assistant. This draft was revised and submitted for peer review by 3-4 randomly selected peers. This feedback was then incorporated into a third iteration that was then evaluated by the same teaching assistant as iteration 1 and was assigned a final grade. Detailed in Table 2, in addition to the TA and peer evaluations, there have been 5 other evaluations of portions of the dataset. The most detailed analysis was conducted by the author as part of his dissertation research (Verleger, 2009) and is, for the purposes of this research, assumed to be representative of the true score.

Table 2. MEA Evaluations

Year	Evaluator	# Teams Evaluated	Iterations Evaluated	Approx. Time/Eval	Rubric Style
2008	Teaching Assistant*	147	1 & 3	30 mins	Numeric & Free Response
2008	Peers		2		
2009	Expert		1, 2, & 3	60 mins	
2014	Algorithmic Assignment	53	1	N/A**	Numeric Only
2015	Expert			2 mins	
2016	Student Researcher 1	147	1, 2, & 3		
2016	Student Researcher 2				

* 15 Teaching Assistants, each with 7-8 or 14-15 teams, depending on the number of sections being evaluated.

**Algorithmic Assignment required multiple hours to generate assignment trees from training data but nearly instant time to apply trees to selected samples. Details of process can be found in Verleger (2014).

For this study, the expert and two trained researcher assistants each evaluated MEAs with the express purpose of evaluating them using a "gut reaction" in 2 minutes or fewer. The expert relied upon his extensive experience evaluating and researching MEAs, while the research assistants were training using the same protocol (described below) as the Teaching Assistants, but with the explicit instructions that they were going to be going the gut reaction evaluation and they should be trying to develop a mental evaluation heuristic.

Training Protocol

The training protocol for teaching assistants and the student researchers consisted of three stages.

Stage 1) General instruction about MEAs and developing a good MEA solution. This consisted of a passive lecture-style information session designed to set a larger context for MEAs.

Stage 2) Development of a personal solution to the MEA. Participants developed their own solution to the MEA. For the research assistants, they were asked to iterate on their solution until the author found their solutions to be of sufficiently high quality. This iteration process was not done for the teaching assistants, as it was deemed too time intensive.

Stage 3) Full evaluation of 5 samples with comparison to an expert evaluator. Participants were presented with 5 sample MEA solutions and asked to complete a full evaluation of the work using both the numeric and free response items. After each evaluation, they are shown their review next to an expert's review of that same sample and asked to reflect on how they might improve their evaluation to more closely align with the expert.

For peer review, the peers went through a similar, but much shorter training process to the teaching assistants, with Stage 3 being reduced to only a single training evaluation and comparison to expert.

Results

For each of the 7 rubric items shown in Table 1, the 6 non-expert evaluations completed between 2008 and 2016 were compared against the expert scores. Peer evaluation scores are a rounded average of the scores assigned by the individual students that provided the peer review. The results are shown in Figures 2-8, with the legend for each graph shown in Figure 1. In Figures 2-8, each bar is the same overall length and represents 100% of the responses that evaluator made for that rubric item. Values less than 2% are shown, but do not have a text label indicating their percentage to reduce clutter. Bars are centered around the "Matched Expert" section, with more generous markings to the right and less generous markings to the left.

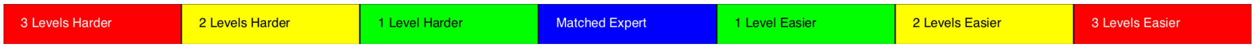

Figure 1. Legend for Figures 2-8.

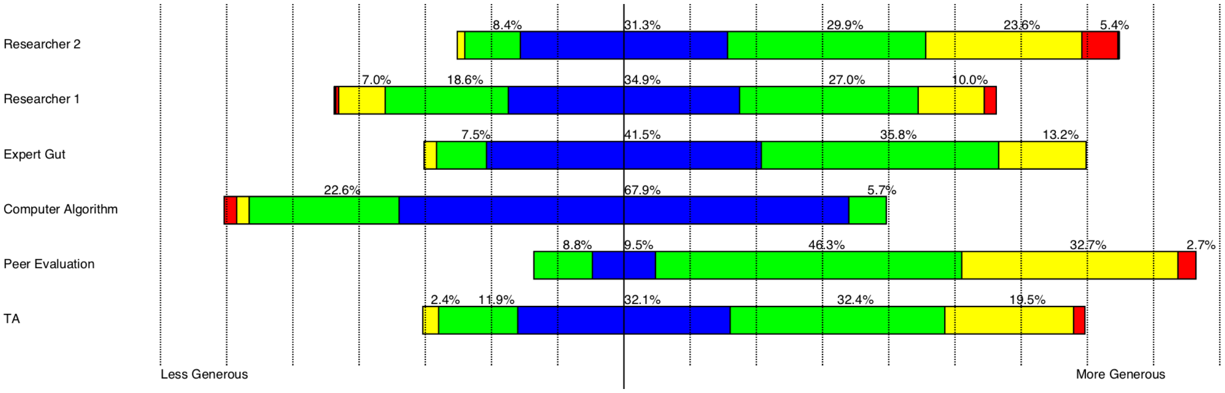

Figure 2. Mathematical Model Complexity

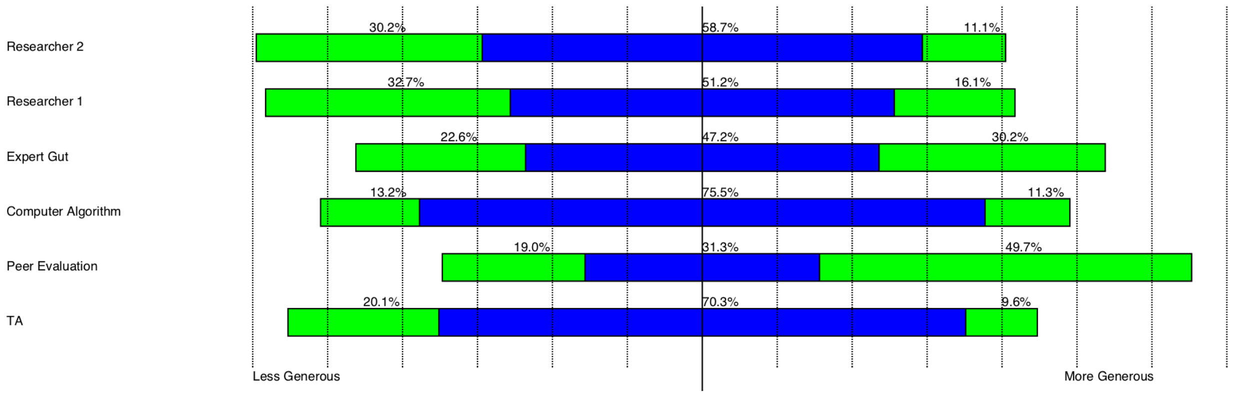

Figure 3. Data Usage

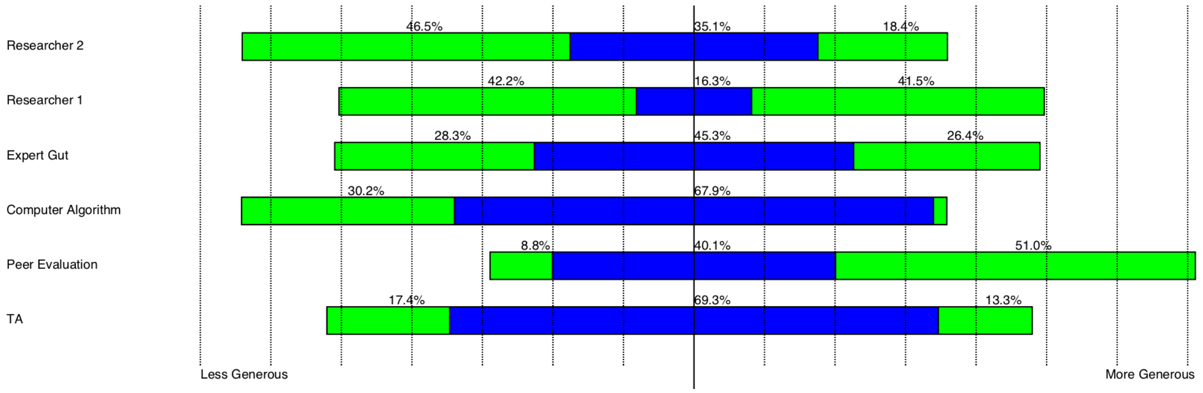

Figure 4. Rationales

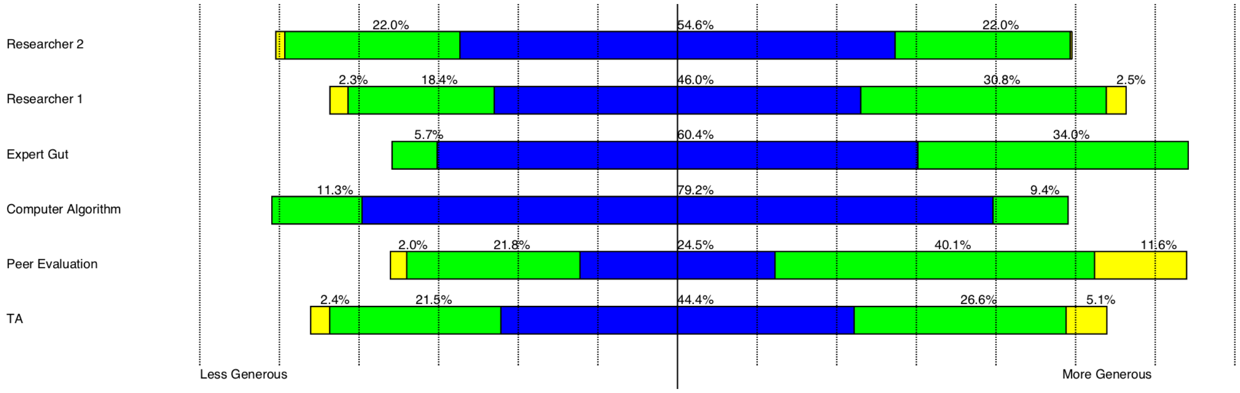

Figure 5. Re-usability/Modifiability

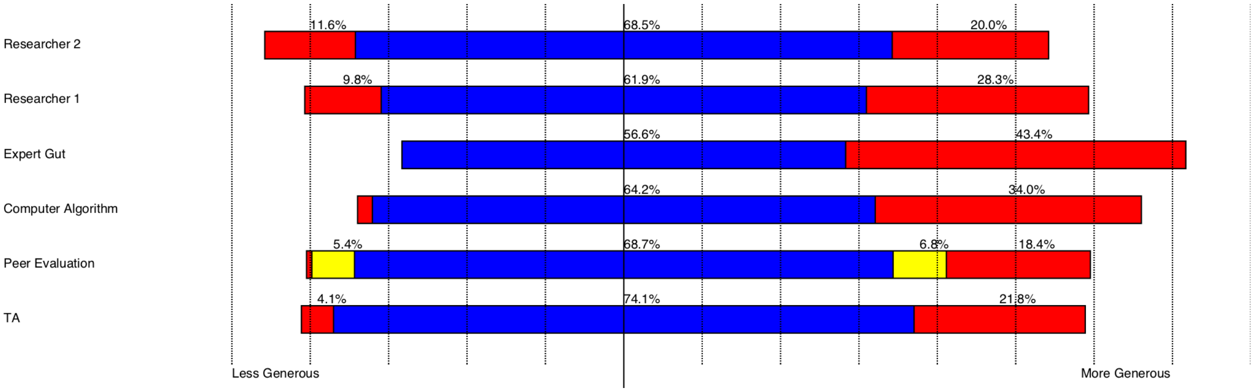

Figure 6. Results

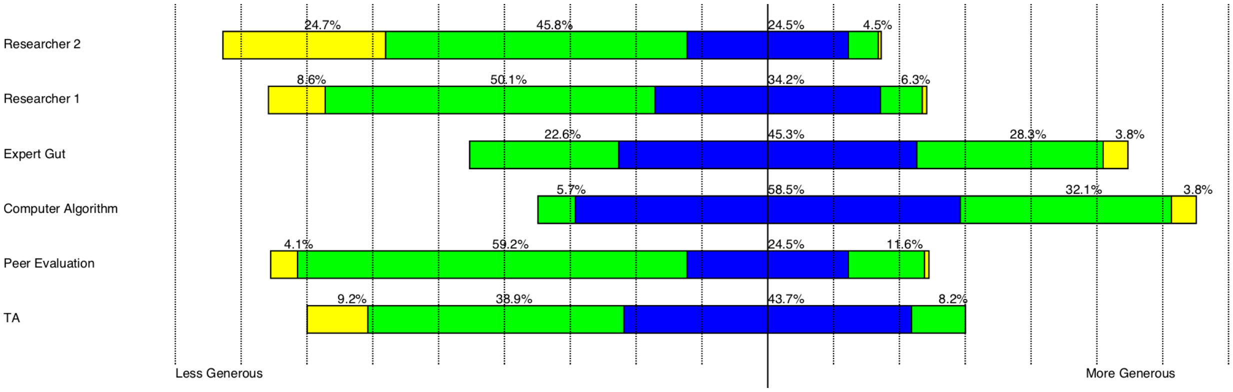

Figure 7. Audience Readability

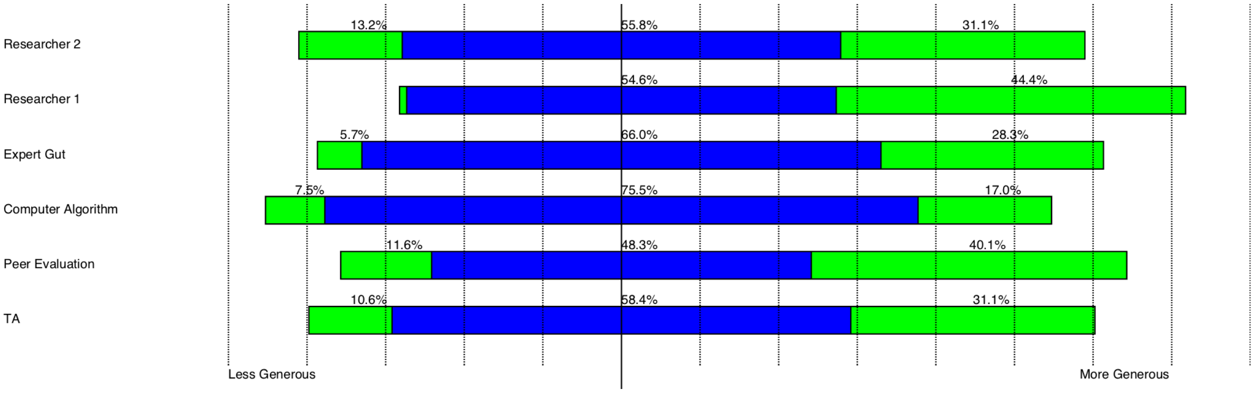

Figure 8. Extraneous Information

Analysis

The data highlights several interesting points. First and foremost, the two reviewers did not produce evaluations that can accurately classify which teams need help. While they only represent two evaluators and broader statistical inference cannot meaningfully be extrapolated to all potential evaluators, they do represent "typical" teaching assistants. Their inaccurate assessment highlights that the current training regimen is not sufficient for them to consistently produce accurate results. In fact, the expert's gut reaction evaluation was only marginally better.

A few specific notes of interest are:

- The expert's gut reaction for Results is the least accurate for that item. In training, the focus of the results item is that a high-quality solution includes both the ranking of teams in the contest as well as the numeric scores that led to that rating. Students often struggle to provide all the numeric values necessary for the ratings. The lower alignment of the expert's gut reaction scores are likely because the expert's heuristic for quick evaluation was largely based on the presence of any numeric results and did not validate that each of the appropriate sets of numbers were present.
- Audience Readability was difficult for the reviewers to assess quickly. A large portion of the readability is based on sentence structure, grammar, and clear writing; all attributes that are difficult to judge quickly. This is further exemplified by the fact that neither of the reviewers are native English speakers.
- For nearly all the evaluations, the computer algorithmic approach was still superior to human evaluation. The underlying problem with computer evaluation was the need for detailed training data.

Conclusions and Next Steps

While the results are not surprising, they do demonstrate why random assignment is so popular in peer review. Attempts to find a meaningful, resource conscious approach to quickly classifying student work for peer review have found the problem to be much larger in scope than originally anticipated. The onset of reviewer fatigue may have caused some issues in reviewer accuracy. While reviews were targeted to 2 minutes each, both reviewers still had decreased accuracy over time, one markedly so. Even in restricting the analysis to earlier evaluations, the accuracy was still not acceptable.

The next step for this research is to explore alternative and hybrid approaches. For example, future work could identify if the computer algorithm would require less training or see improved accuracy if both researchers gut reaction evaluations were included to potentially guide scores. For instance, 23% of the time, both researchers agreed with the expert score. 60% of the time, at least 1 of the researchers agreed with the expert score. 5.6% of the time, both researcher's marks agreed with the expert score where the computer algorithm did not. Using this data, the algorithm may be better able to discern the true score.

While these results are not encouraging for MEA evaluators, the primary recommendation for other researchers seeking to utilize these methods is to focus on improving the training process, as this is a lynchpin need to improving the evaluation process.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1264005. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

ABET. (2013). *Criteria for accrediting programs in engineering*. Baltimore, MA: ABET, Inc.

Ballantyne, R., Hughes, K., & Mylonas, A. (2002). Developing procedures for implementing peer assessment in large classes using an action research process. *Assessment & Evaluation in Higher Education*, *27*(5), 427–441. https://doi.org/10.1080/0260293022000009302

Blackmun, H. Blackmun, H., Daubert v. Merrell Dow Pharmaceuticals, 509 US United States Reports 579 (1993). Supreme Court of the United States,. Retrieved from http://caselaw.lp.findlaw.com/scripts/getcase.pl?court=us&vol=509&invol=579

Boud, D. (2000). Sustainable assessment: Rethinking assessment for the learning society. *Studies in Continuing Education*, *22*(2), 151–167. https://doi.org/10.1080/713695728

Burnham, J. C. (1990). The Evolution of Editorial Peer Review. *JAMA: The Journal of the American Medical Association*, *263*(10), 1323–1329.

Diefes-Dux, H. A., Hjalmarson, M. A., Miller, T. K., & Lesh, R. A. (2008). Model-eliciting activities for engineering education. In J. S. Zawojewski, H. Diefes-Dux, & K. Bowman (Eds.), *Models and modeling in engineering education: Designing experiences for all students* (pp. 17–36). Rotterdam, the Netherlands: Sense Publishers.

Diefes-Dux, H. A., Zawojewski, J. S., & Hjalmarson, M. A. (2010). Using educational research in the design of evaluation tools for open-ended problems. *International Journal of Engineering Education*, *26*(4), 807–819.

Guilford, W. H. (2001). Teaching peer review and the process of scientific writing. *Advances in Physiology Education*, *25*(3), 167–175.

Lesh, R. A., Hoover, M., Hole, B., Kelly, A., & Post, T. (2000). Principles for Developing Thought Revealing Activities for Students and Teachers. In A. Kelly & R. A. Lesh (Eds.), *Handbook of Research Design in Mathematics and Science Education* (pp. 591–645). Mahwah, NJ: Lawrence Erlbaum.

Moreira, D. de A., & da Silva, E. Q. (2003). A method to increase student interaction using student groups and peer review over the internet. *Education and Information Technologies*, *8*(1), 47–54. https://doi.org/10.1023/A:1023926308385

Sitthiworachart, J., & Joy, M. (2003). Web-based Peer Assessment in Learning Computer Programming. In *4th Annual Conference of the LTSN Centre for Information and Computer Sciences*. NUI Galway, Ireland.

Sitthiworachart, J., & Joy, M. (2004). Effective peer assessment for learning computer programming. *ACM SIGCSE Bulletin*, *36*(3), 122. https://doi.org/10.1145/1026487.1008030

Verleger, M. A. (2009). *Analysis of an informed peer review matching algorithm and its impact on student work on model-eliciting activities*. *Engineering Education*. ProQuest Dissertations & Theses Global. (366629637), West Lafayette, IN.

Verleger, M. A. (2014). *Using natural language processing tools to classify student responses to open-ended engineering problems in large classes*. *Paper presented at the ASEE Annual*

Conference. Indianapolis, IN: American Society for Engineering Education. Retrieved from https://peer.asee.org/using-natural-language-processing-tools-to-classify-student-responses-to-open-ended-engineering-problems-in-large-classes

Verleger, M. A., Diefes-Dux, H. A., Ohland, M. W. M. W., Besterfield-Sacre, M., & Brophy, S. (2010). Challenges to informed peer review matching algorithms. *Journal of Engineering Education*, *99*(4), 397–408. https://doi.org/10.1002/j.2168-9830.2010.tb01070.x

Verleger, M. A., Rodgers, K. J., & Diefes-Dux, H. A. (2016). Selecting Effective Examples to Train Students for Peer Review of Open-Ended Problem Solutions. *Journal of Engineering Education*, *105*(4), 585–604. https://doi.org/10.1002/jee.20148

Wood, T., Hjalmarson, M. A., & Williams, G. (2008). Learning to design in small group mathematical modeling. In J. S. Zawojewski, H. Diefes-Dux, & K. Bowman (Eds.), *Models and modeling in engineering education: Designing experiences for all students* (pp. 187–212). Rotterdam, the Netherlands: Sense Publishers.

Zhi-Feng Liu, E., Lin, S. S. J., Chiu, C.-H., & Yuan, S.-M. (2001). Web-based peer review: The learner as both adapter and reviewer. *IEEE Transactions on Education*, *44*(3), 246–251. https://doi.org/10.1109/13.940995