Grading Reliability of Teaching Assistants New to Assessment of Realistic Open-Ended Problems

Farshid Marbouti and Heidi A. Diefes-Dux
Purdue University, fmarbout@purdue.edu, hdiefes@purdue.edu

***Abstract* - Many first year engineering courses enroll a large number of students. Open-ended problems are common in engineering courses. When implementing realistic open-ended problems in large educational settings with multiple instructors (or teaching assistants), it is a challenge to design valid and reliable assessment tools that can be consistently used to grade students' responses. The purpose of this study is to evaluate the reliability with which teaching assistants (TAs) who are new to assessing student work on realistic open-ended problems use a valid generic four-dimension rubric that is supported by problem-specific guides and designed to assess student work on mathematical modeling problems. The new TAs reliably used the rubric's seven items to score student work across all dimensions. From the analysis of the TA written feedback on the student responses that were scored differently by the TA and expert, three themes emerged: 1) TAs did not identify errors present in student responses, 2) TAs misunderstood the rubric items, 3) TAs correctly identified errors in student responses but scored the items incorrectly. These three issues can be addressed through modifications to the TA training and the problem-specific guides.**

Index Terms – Assessment tools, open-ended problems, professional development, grading reliability.

INTRODUCTION

Open-ended problems help engineering students achieve critical thinking and problem solving skills. Development of these is essential for fulfilling accreditation requirements in engineering programs and preparing for future success [1-2]. When implementing open-ended problems in large educational settings, such as first-year engineering courses, with multiple instructors (or teaching assistants), it is a challenge to design valid and reliable assessment tools that can be consistently used to grade students' responses [3-4]. This is due in part to the variety of acceptable solutions that can result from such problems. Reliability of grading is both a practical concern and a learning concern. All students should receive an appropriate and fairly determined grade. Students should also receive consistent messages about what they should be learning and whether they are learning.

There are two aspects to ensuring the reliability of grading. The first aspect deals with the design of valid and reliable assessment tools. The second aspect deals with the application of the tools by instructors or teaching assistants (TAs) and the training and support necessary for reliable application. A great deal of research has been conducted on the design of valid and reliable assessment tools (e.g. [3, 5, 6]). However, less research has been done on consistency of applying the assessment tools by instructors (or TAs). In addition to adding specific instructions to the assessment tools, training of instructors (or TAs) increases grading reliability [5]. Some researchers (e.g. [7]) suggest moving from TA training to TA professional development. TA professional development goes beyond practical issues and tries to address pedagogy and development strategies including feedback and assessment strategies [8].

Realistic Open-ended Problems

One way to engage students in developing their problem solving skills is via Model-Eliciting Activities (MEAs). MEAs are realistic open-ended problem solving activities that ask student teams to create and improve a generalizable mathematical model for a direct user through an iterative sequence of documented model development [9-11]. MEAs involve students in communication, teamwork, critical thinking and problem solving, which are all necessary skills in science, technology, engineering and mathematics (STEM) education [1, 2, 12, 13].

MEAs have been designed and used in the First-Year Engineering (FYE) Program at Purdue University for over 10 years [4, 14]. There, MEAs are part of the two semester required FYE engineering course sequence. In these courses, MEAs are implemented in an iterative process. The whole sequence for each MEA includes three iterations, in which student teams generate two drafts and a final response. At the end of the first iteration, students receive feedback from peers on draft1. The second draft and final response are evaluated by the TAs.

The MEA implementation sequence and assessment tools have been revised and improved for the FYE program multiple times. In addition, one of the MEAs, *Just-in-Time* (*JIT*) *Manufacturing* [15], has been used and revised several times. This MEA provides student teams with potential shipping companies that can be used by a client company (DDT company) to deliver product between two subsidiaries in a JIT fashion. Data consists of a record of the shipment delays for each shipping company. Students are asked to develop a procedure to rank the shipping companies using this data. The data sets provided to students are not normally distributed, encouraging close examination of the distribution of data in addition to central tendency and variance to solve the problem [16].

Development of valid and reliable assessment tools

Prior to Fall 2007, TAs used a holistic Quality Assurance Guide (QAG) adapted from the models and modeling perspectives [17, 18]. The same QAG was used for all MEAs. Due to inconsistency in TAs' grading and no attention to the conceptual framework intended by the QAG, [19], researchers noted the need to design specific evaluation packages for each MEA. These authors used the *JIT Manufacturing* MEA to guide the design of two evaluation tools: an MEA-specific *Instructors' MEA Assessment/Evaluation Package* (*I-MAP*) and a generic *MEA Feedback and Assessment Generic Rubric* (*MEA Rubric*). An *I-MAP* was designed specifically for the *JIT Manufacturing* MEA. The *MEA Rubric* was designed to be used with any MEA.

TAs' practice grading with these tools was evaluated in Fall 2007 and Fall 2008 [19]. Based on the findings of this study, as well as experts' and TAs' feedback, a four dimension *MEA Rubric* emerged. The four dimensions are mathematical model, re-usability, modifiability and share-ability. These dimensions maintain fidelity to what engineers value in high quality work [19]. In addition to the development and evaluation of these tools, TA training was modified to explain the *I-MAP* and the *MEA Rubric* to the TAs and provide practice with applying the *MEA Rubric* to prototypical student work.

In a later study, in an attempt to examine quality of student responses in each dimension, the *MEA Rubric* and the *I-MAP* for *JIT Manufacturing* MEA were evaluated for Spring 2009 [20]. Based on this study and as a result of a better understanding of student responses, the *I-MAP* was again revised. One of the suggestions made in this study was to add explicit examples of acceptable responses for some dimensions in the *I-MAP*.

TAs use an online system to grade students' responses and provide feedback [21]. The current *MEA Rubric* has four dimensions (Table 1). Two of the dimensions (mathematical model and share-ability) have more than one scored item. In total, there are seven scored items across the four dimensions. Table 2 summarizes these seven scored items. Teams are issued a grade for the final response based on the minimum of these seven scored items.

TA Training

The TAs who joined the FYE instructional team in Fall 2011 participated in two MEA training series, one in Fall 2011 and one in Spring 2012. These occurred prior to their assessing student responses to the *JIT Manufacturing* MEA. In Fall 2011, the training was focused on a MEA that was being used in that semester (*Travel Mode Choice* MEA [22]). The training series for Spring 2012 was based on the *JIT Manufacturing* MEA. In both of these training series more developed MEA-specific *I-MAPs* (compared to the prior semesters) in addition to the *MEA Rubric* were used. Fall 2011 training included more face-to-face time spent on specific examples of students' responses.

For the Spring 2012 training, prior to the face-to-face session, TAs were asked to respond to the *JIT Manufacturing* MEA context setting and individual questions that preface the MEA (e.g. [23]) and solve the MEA via the online tool, then apply the *MEA Rubric*, supported by the *JIT Manufacturing I-MAP*, to their MEA response. These pre-face-to-face training activities are intended to familiarize the TAs with the MEA and the assessment tools, so that the face-to-face sessions can focus on issues of assessing student work. During the face-to-face session, the *I-MAP* for the *JIT Manufacturing* MEA and the *MEA Rubric* were explained to the TAs and two sample responses were reviewed. To familiarize the TAs with grading rubric, the *MEA Rubric* was applied to actual student responses [8]. After the face-to-face session, TAs evaluated three more student responses and compared their evaluation to that of an expert. The comparison of TA and faculty grading of sample work helps the TAs self-calibrate their grading techniques [8].

As a result of the improvements to the MEA assessment and evaluation tools and TA training, it is expected that TA grading consistency has improved and more importantly become reliable for the *JIT Manufacturing* MEA. The goal of the current study is to evaluate the reliability of TAs' grading to understand whether or not the new TAs are reliable in applying the *MEA Rubric* and the *I-MAP* given current TA training strategies.

RESEARCH QUESTION

Can TAs reliably apply the four-dimension *MEA Rubric* to assess student work? If not, why not?

METHODS

Participants & Setting

In Spring 2012, the *JIT Manufacturing* MEA was implemented with 1650 students divided into 416 teams spread across 15 sections. In each of the 15 sections, one Graduate TA (GTA) and 5 Undergraduate TAs (UTA) assessed the student teams' responses to the MEA. The 79 TAs that were involved in this course had different levels of experience with this course. Forty-two of them were TAs for the course prior to Fall 2011 (for anywhere from one

TABLE 1
MEA RUBRIC DIMENSIONS [20].

Dimension	Description
Mathematical Model	A mathematical model may be in form of a procedure or explanation that accomplishes the task, makes a decision, or fills a need for a direct user. A high quality model fully addresses the complexity of the problem and contains no mathematical error.
Re-usability	The procedure can be used by the direct user in new but similar situations.
Modifiability	The procedure can be modified easily by the direct user for use in different situations.
Share-ability	The direct user can apply and replicate results. If the mathematical model is not developed in enough detail to clearly demonstrate that it works on the data provided, it cannot be considered sharable.

TABLE 2
SUMMARY OF *MEA RUBRIC* ITEMS AS IT PERTAINS TO THE *JIT MANUFACTURING* MEA [20].

Dimension	Item	Item Description	Score
Mathematical Model	Mathematical model complexity (specific to JIT MEA)	The procedure looks past measures of central tendency and variation to look at the actual distribution of the data. The distribution of data can be accounted for in a number of ways, including but not limited to, (1) determining the frequency of values (particularly minimum and maximum values), (2) determining the frequency of values within intervals, (3) considering the difference between median and mean, and (4) quantifying the shape of the distribution.	0-4
	Accounting for data types	Mathematical model takes into account all types of data provided to generate results. Even if justifications are provided, no data should be discarded for this problem. Justifications might be provided for things like (1) removal of any part of the time data for any company (e.g. removal of outliers: outliers cannot be removed because there is no information tagged to why a shipping company is late), (2) dropping shipping companies: the requested procedure is supposed to rank the shipping companies, all of the shipping companies.	3-4
Re-usability	Re-usability	Procedure provides the following information: (1) identification of direct user: DDT's logistic manager, (2) deliverable: procedure, (3) function and criteria for success: rank shipping companies in order of best to least able to meet DDT's timing need, (4) constraints: given historical data for multiple shipping companies the time late for shipping runs between two specified locations, (5) overarching description: should provide an overview of how the ranking is determined, and (6) assumption and limitations concerning the use of procedure: what, if anything, limits the use of procedure depends on the details of the procedure.	2-4
Modifiability	Modifiability	Procedure contains acceptable rationales for critical steps and clearly states assumptions associated with procedural steps. Critical steps that need justification/rational: (1) when teams use any statistical measures, these measures must be justified – explain what these measures tells the user, (2) why some measures are being used over others should be explained (3) when developing intermediate ranking or weighting methods, these must be justified; it should be explained why all weightings are the same or different, (4) hard-coded values imbedded in procedural steps require explicit explanation of where the values come from.	2-4
Share-ability	Results	The mathematical model is applied to the data provided to generate results. Quantitative results are also provided. In a level 4 response, both ranking and quantitative results for each shipping company are provided. Units are given and are correct.	1,2,4
	Apply and replicate results	Procedure is easy-to-read and use. If this has not been delivered, the solution is not level 3 work. At a minimum, the results from applying the procedure to the data provided must be presented in the form requested.	2-4
	Extraneous information	The mathematical model should be free of distracting and unnecessary text. This might include (1) outline formatting, (2) indications of software tools (e.g. MATLAB™, Excel™ or, more generally, spreadsheets) necessary to carry out computations, (3) explicit instructions to carry out common computations, (4) discussions of issues outside the scope of the problem, (5) general rambling.	3-4

semester to several years), 20 TAs joined in Fall 2011, and 17 TAs joined the course in Spring 2012. Only those TAs that joined in Fall 2011 were considered for this study as the foundational TA training with MEAs was provided in the Fall semester. Further, focusing on these TAs reduces the effects of different experience levels on the reliability of grading. Four Graduate TAs and 16 Undergraduate TAs joined in Fall 2011. These TAs participated in the TA training in Fall 2011 and Spring 2012.

Data Collection/Analysis

Student teams used an online system to submit their responses. TAs also used this online system to grade student team responses and provide feedback [21]. The student team draft2 responses (which was the first draft graded by TAs) and TAs' scores and feedback for each of the 20 TAs that joined in Fall 2011 constitute the data for this study. From this, one team graded by each TA was randomly selected for inclusion in the study.

To become an expert in grading students' responses, the researcher first applied the *MEA Rubric* and *JIT Manufacturing* MEA *I-MAP* to three sample team responses and compared the scores with that of an expert and read the expert's feedback on the sample responses. Then teams' draft2 were graded by the researcher (expert).

Following the model used by [19], for each scored item in the student team response, the TA score was reported as a difference to the expert score. For example, if a TA graded a team response 3 and the expert graded it 2, the TA score was reported as +1. Following criteria adapted from [24] and used by [19], 90% was set as the criteria for reliability. That is, if 90% of TA scores for a given scored item were within one point of the expert score, the grading was considered reliable. For the items with only two possible scores (accounting for data types and extraneous information) the percentage of TA scores that was the same as the expert was considered for the criteria for reliability.

RESULTS AND DISCUSSION

This section reports and discusses the results based on the four dimensions of the *MEA Rubric*. Table 3 illustrates the difference between the TA and expert scores on seven scored items.

Mathematical Model

The mathematical model dimension has two scored items: mathematical model complexity and accounting for data types. The possible score range for mathematical model was 0 to 4. All TAs' scores were within one point of expert scores and 70% (14) of the scores were the same as the expert score. Five TAs scored lower and only one higher

compared to the expert. Despite the complexity and more detailed nature of the mathematical model complexity item, the reliability for this item, which was specific to *JIT Manufacturing* MEA, was high. This might be because there is a detailed explanation of each level of response for this item in the *I-MAP*. For the five student team responses that TAs scored lower than the expert, students took into account distribution of data (at different levels). Further analysis of the written feedback revealed that in most of these five cases, TAs provided relevant feedback indicating they had a good understanding of the student team response, but the scored level was not consistent with the feedback. Adding more examples of sample student team responses to the TA training or *I-MAP* might help TAs score this item more reliably. The only TA who scored higher than the expert gave the highest score possible while there were some mathematical errors in the student team's model. This TA did not provide detailed feedback as s/he stated "this part is good". The TA may not understand the complexity of the *JIT Manufacturing* problem or may not be able to identify minor mathematical errors. This problem can be addressed during TA training by providing examples of student responses that are well developed but have minor mathematical errors.

TABLE 3
THE DIFFERENCE BETWEEN THE TA AND EXPERT SCORED (n=20).

Dimension	Scored Item	Difference with expert score				
		-2	**-1**	**0**	**+1**	**+2**
Mathematical Model	Complexity	0	5	14	1	0
	Accounting for data types	N	2	18	0	N
Re-usability	Re-usability	0	4	9	7	0
Modifiability	Modifiability	0	4	14	2	0
Share-ability	Results	0	N	18	N	2
	Apply and replicate results	1	1	14	3	1
	Extraneous information	N	0	18	2	N
Final Score		0	4	16	0	0

N: difference is not valid for this item.

The possible scores for accounting for data types were 3 and 4. For this scored item, 90% (18) of TAs scores were the same as the expert scores. Two TAs scored lower than the expert. The second item in mathematical model, accounting for data types, was also reliable. Based on an analysis of the written feedback, the two TAs who scored student work differently than the expert had a different understanding of this item. In one case, the team used a point system to rank the companies. In assigning points to the companies, a weighting system was used to "reward" on-time deliveries and "punish" late ones. The weighting system students used included a zero weight, leading the TA to decide that they did not take into account all data. The other TA decided the team did not take into account all types of data because the team's model only included "one type of data" (i.e. the number of zeros and the number of maximum values). Adding more explicit instructions to the TA training or *I-MAP* might be beneficial to mitigate these types of misunderstandings.

Re-usability

The possible scores for the re-usability dimension item were 2 to 4. Only 45% (9) of TAs scores were the same as the expert score. However, all scores were within one point of the expert scores. Thus, while the TAs are reliable by the criteria (more than 90% of the scores were within one point score of expert score), the results are concerning (55% of the scores were different). This may be due to the more detailed nature of the scoring required for this dimension. To calculate the reusability score, TAs first assign six sub-scores based on the defined criteria in the *I-MAP*, and then they calculate the reusability score based on the sum of these six sub-scores. TAs only submit the calculated re-usability score in the online system. One suggestion is to add these sub-scores to the online system, so the TAs do not have to assign the sub-scores and do the calculations off-line and then enter the reusability score in the system.

Based on an analysis of the written feedback, the 11 students responses that TAs scored different than the expert may have different reasons. In one case, the TA commented about the mathematical model and outliers, which may be an indication that the TA did not understand the reusability dimension. In six cases, the TAs commented appropriately about what was correct and what was missing in the students' responses, but assigned a different score level compared to the expert. While further analysis is required to understand the cause of the expert-TA disparity, one possible explanation is that these TAs did not assign the sub-scores provided in the *I-MAP* but rather assigned an overall score for this dimension. In four cases, TAs did not provide any feedback indicating they thought all parts of the response were acceptable. This might be due to misunderstanding of this item (or at least some of the sub-items). While reusability dimension is one of the more detailed parts of the *I-MAP* and the designers of the *I-MAP* expected that this dimension has a high reliability due to the straightforward identification of the presence of elements in student responses, these findings revealed that TAs have difficulty with this item.

Modifiability

The possible scores for the modifiability dimension are 2 to 4. All TAs scores were within one point of the expert scores and 70% (14) of TAs scores were the same as the expert score. Four TAs scored lower and two higher than the expert. Further analysis of the written feedback revealed that some of the TAs may not understand this dimension. For example, one TA commented on the complexity of the problem in this dimension.

Share-ability

Share-ability has three scored items: results, apply and replicate results, and extraneous information. The possible scores for the results item are 1, 2 and 4. All TAs scores for

this item were within two points of the expert scores, which is one scoring level difference for this item and 90% (18) of TAs scores were the same as the expert score. Only two TAs scored higher than the expert. In both cases, the units were missing, and in one of them, the quantitative results (i.e. calculated scores for each company) were also missing. But TAs did not identify the problems and assigned the highest score possible to the responses. Both quantitative results and units are clearly stated in the *I-MAP* as being required in student responses; this was discussed during TA training.

The possible scores for the apply and replicate results item were 2 to 4. For this item, 90% (18) of TAs scores were within one point of the expert scores and 70% (14) of TAs scores were the same as the expert score. In one case, the TA scored two points higher than the expert. Further analysis of the written feedback revealed that the TA provided relevant feedback but scored differently. The TA that scored two points lower than the expert did not provide detail feedback about this item. Other cases that TAs scored one point different than the expert, had different reasons.

The possible scores for the extraneous information item were 3 and 4. For this scored item, 90% (18) of TAs scores were the same as the expert scores. Two TAs scored higher than the expert. According to the *I-MAP*, students, in their response to the MEA, should avoid unnecessary information such as how to calculate a statistical measure or using software for calculations. For the two student team cases for which the TA scores were higher than the expert, in one case, how to calculate mean was explained in the response. While the TA commented about this in the feedback, s/he assigned the highest score possible. In the other case, MATLAB™ commands were included in the response. In this case, the TA did not identify the problem. Both of these cases were explained in the *I-MAP* and TA training.

Final Score

A final score for each team is calculated as the minimum of the seven assigned scores. The possible scores for the final score were 0 to 4. For the final score, all TAs scores were within one point of the expert scores and 80% (16) of the TAs' final scores were the same as expert final score.

Themes Emerged from Analysis of Feedback

Three main themes emerged from analysis of the written TA feedback for those cases in which the TAs scored student work differently than the expert:

TAs did not identify errors in the student response. When TAs do not detect errors, they do not provide written feedback. This makes it difficult for researchers to understand the exact nature of the scoring issue. Not identifying errors was one reason for differences between the TA and expert scores for share-ability (especially results and extraneous information). Highlighting the *I-MAP* instructions for these items and providing training using examples as seen here might be beneficial.

TAs misunderstood the scored item. This was the main reason for differences in scores between the TAs and expert on two items: modifiability and accounting for all data types (mathematical model dimension). Spending more time in TA training explaining these items might help TAs have a better understanding of these items.

TAs correctly identified errors in student response but scored differently. Some TAs that scored student work differently than the expert commented in their written feedback on what should be included in the student team response but assigned a score that did not match their comments. This was noted for two items: mathematical model complexity and reusability. More practice in applying the rubric to the students' responses and making comparisons to the expert might be helpful.

LIMITATIONS

One limitation of this study was the small sample size. Only 20 student team responses out of 416 total student team responses and 95 student team responses that the 20 new TAs graded were selected for this study. While the selection of the team responses for each TA was random, it might not be a representative of these TAs' grading for several reasons. First, a particular team response might have used unusual mathematical methods and made it difficult to grade. Second, the quality of TA grading may have changed from the first to the last team response s/he graded.

In addition, only draft2 was included in this study. The reliability of TAs' grading may have changed from draft2 to the final response. The current study only focused on evaluating reliability of new TAs regardless of whether they are undergraduate or graduate TAs. Reliability of these two groups of TAs may have been different. Furthermore, as explained earlier, this study only focused on TAs new to grading open-ended problem; more experienced TAs may be more or less reliable on the various *MEA Rubric* items. Another limitation may be the classification of "new" in this study. Some of the TAs were not new to MEAs and the *MEA Rubric* since they worked on MEAs (and likely the *JIT Manufacturing* MEA) as first-year engineering students and had experience via peer feedback with applying the *MEA Rubric*, albeit without the *I-MAP*.

CONCLUSIONS

The purpose of this study was to evaluate the reliability of TAs' grading who are new to grading open-ended problems to understand whether or not the new TAs are reliable in applying a valid rubric designed to assess student work on mathematical modeling problems. All seven scored items within four dimensions (mathematical model, re-usability, modifiability, and share-ability) were reliably applied by the TAs. However, re-usability scores were concerning because less than half of the TAs' re-usability scores were the same as the expert.

In summary, after analysis of the written feedback for the student responses that scored differently by the TA and

expert, three themes emerged that can be addresses via TA training or clarification in assessment tools: 1) TA did not identify errors in student response, 2) TA misunderstood the scored item, 3) TA correctly identified errors in student response but scored differently.

Future research can investigate the reasons that the scores for the reusability dimension were so different. In addition, evaluating the reliability of TA grading on a larger sample size would be beneficial. Evaluating more than one student team response per TA can further illustrate the reliability of TAs' grading. It would also be useful to compare TA's grading of draft2 and final response. Additionally, comparison of reliability of GTAs and UTAs would reveal if these two groups of TAs are similar in terms of reliability or not.

ACKNOWLEDGMENT

This work was made possible by a grant from the National Science Foundation (DUE 0717508). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Accreditation Board of Engineering and Technology Accreditation Department (ABET) (2011). *Criteria for accrediting engineering programs, 2011 - 2012*. ABET Inc.: Baltimore, MD, Retrieved from http://www.abet.org/eac-current-criteria/

[2] National Academy of Engineering (NAE) (2004). *The engineering of 2020: Visions of engineering in the new century*. Washington, DC: The National Academic Press.

[3] Davis, D. C., Gentili, K. L., Trevisan, M. S. & Calkins, D. E. (2002). Engineering design assessment process and scoring scales for program improvement and accountability. *Journal for Engineering Education*, 91(2), 211–221.

[4] Diefes-Dux, H. A. & Imbrie, P. K. (2008). Modeling Activities in a First-Year Engineering Course. In J. S. Zawojewski, H. A. Diefes-Dux and K. J. Bowman (eds) *Models and Modeling in Engineering Education: Designing Experiences for All Students* (pp. 55-92). Rotterdam, the Netherlands: Sense Publishers.

[5] Stellmack, M. A., Konheim-Kalkstein, Y. L., Manor, J. E., Massey A. R. & Schmitz, J. A. (2009). An Assessment of Reliability and Validity of a Rubric for Grading APA-Style Introductions, *Teaching of Psychology, 36*(2),102-107.

[6] Newell, J. A., Dahm, K. D., & Newell, H. L. (2002). Rubric development and interrater reliability issues in assessing learning outcomes. *Chemical Engineering Education*, 36, 212–215.

[7] Seymour, E. (2005). *Partners in innovation: Teaching assistants in college science courses*. Lanham MD: Rowan & Littlefield Publishers.

[8] Diefes-Dux, H. A., Osburn, K., Capobianco, B. M., & Wood, T. (2008). On the front line: Learning from the teaching assistants. In J. S. Zawojewski, H. A. Diefes-Dux and K. J. Bowman (eds) *Models and Modeling in Engineering Education: Designing Experiences for All Students* (pp. 225-256). Rotterdam, the Netherlands: Sense Publishers.

[9] Lesh, R., Hoover, M., Hole, B., Kelly, A., & Post, T. (2000). Principles for developing thought-revealing activities for students and teachers. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of Research Design in Mathematics and Science Education* (pp. 591-645). Mahwah, NJ: Lawrence Erlbaum.

[10] Lesh, R., Hoover, M., & Kelly E. (1993). Equity, assessment, and thinking mathematically: Principles for the design of model eliciting activities. In I. Wirszup & R. Streit (Eds.), *Developments in school mathematics education around the world* (Vol. 3). Reston, VA: National Council of Teachers of Mathematics.

[11] Diefes-Dux, H. A., Hjalmarson, M. A., Miller, T. K., & Lesh, R. (2008). Model-Eliciting Activities for engineering education. In J.S. Zawojewski, H. A. Diefes-Dux, & K. Bowman (Eds.), Models and modeling in engineering education: designing experiences for all students (pp. 17-35). Rotterdam, The Netherlands: Sense Publishers.

[12] Common Core State Standards Initiative (CCSSI). (2010). Common core state standards for mathematics. Retrieved from http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf

[13] National Council for Accreditation of Teacher Education (NCATE) (2008). Professional standards for the accreditation of teacher preparation institutions. Retrieved from http://www.ncate.org/LinkClick.aspx?fileticket=nX43fwKc4Ak%3D&tabid=669

[14] Salim, A. & Diefes-Dux, H. (unpublished). First-year students' problem formulation in Model-Eliciting Activities.

[15] Hjalmarson, M. (2007). Engineering students designing a statistical procedure. *Journal of Mathematical Behavior*, 26(2), 178–188.

[16] Diefes-Dux, H. A. & Cardella, M. A. (2012). A first take on an individual data generation assignment for open-ended mathematical modeling problems, *in Proceedings of the 119th ASEE Conference*, San Antonio, TX.

[17] Carmona-Dominguez, G. (2004). *Designing and assessment tool to assess students' mathematical knowledge*, Dissertation presented at Purdue University.

[18] Clark, K. & Lesh, R. (2003). Whodunit? Exploring proportional reasoning through the footprint problem. *School Science and Mathematics*, 103(2), 92–99.

[19] 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.

[20] Carnes, M. T., Diefes-Dux, H. A., & Cardella, M. E. (2011). Evaluating student responses in open-ended problems involving iterative solution development in Model Eliciting Activities (MEAs), *in Proceedings of the 118th ASEE Conference*, Vancouver, B.C., Canada.

[21] Verleger, M., Diefes-Dux, H., (2010). Facilitating teaching and research on open-ended problem solving through the development of a dynamic computer tool, *in Proceedings of 117th ASEE Conference*, Louisville, KY.

[22] Zawojewski, J., Diefes-Dux, H. A., & Bowman, K. (2008). *Models and modeling in engineering education: designing experiences for all students.* Netherlands: Sense.

[23] Salim, A. & Diefes-Dux, H., (2010). Graduate teaching assistants' assessments of students' problem formulation within Model-Eliciting Activities, *in Proceedings of 117th 2010 ASEE Conference*, Louisville, KY.

[24] Herman, J. L., Aschbacker, P. R., & Winter, L. (1992). *A Practical Guide to Alternative Assessment*. Arlington, VA: Association for Supervision and Curriculum Development (ASCD).

AUTHOR INFORMATION

Farshid Marbouti, PhD Student and Graduate Research Assistant, Purdue University, fmarbout@purdue.edu,
Heidi A. Diefes-Dux, Associate Professor, Purdue University, hdiefes@purdue.edu