AC 2009-1184: MULTIDIMENSIONAL TOOL FOR ASSESSING STUDENT-TEAM SOLUTIONS TO MODEL-ELICITING ACTIVITIES

Heidi Diefes-Dux, Purdue University

Heidi 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. She coordinated (2000-2006) and continues to teach in the required first-year engineering problem solving and computer tools course. Her research focuses on the development, implementation, and assessment of model-eliciting activities with realistic engineering contexts.

Matthew Verleger, Purdue University

Matthew Verleger is a doctoral candidate in the School of Engineering Education at Purdue University. He received his B.S. in Computer Engineering and his M.S. in Agricultural and Biological Engineering, both from Purdue University. His research interests are on how students develop mathematical modeling skills through the use of model-eliciting activities and peer review as a pedagogical tool.

Judith Zawojewski, Illinois Institute of Technology

Judith Zawojewski is an Associate Professor of Mathematics and Science Education at Illinois Institute of Technology. She received a B.S.Ed. in Mathematics and Education at Northwestern University, a M.S.Ed in Mathematics Education from National College of Education, and a Ph.D. in Education at Northwestern University. Judith teaches mathematics education courses to practicing teachers and doctoral students. Her research interest is in the use models and modeling for the development of problem solving experiences as sites for research and assessment in the context of program improvement.

Margret Hjalmarson, George Mason University

Margret Hjalmarson is an Assistant Professor of Mathematics Education at George Mason University. She received a B.A. in Mathematics from Mount Holyoke College, an M.S. in Mathematics and a Ph.D. in Mathematics Education from Purdue University. She teaches mathematics education courses for teachers and mathematics specialists in the Mathematics Education Leadership master's and doctoral programs. Her research interests are in students' learning of mathematics in engineering, design-based research, curriculum, and assessment.

Multi-Dimensional Tool for Assessing Student Team Solutions to Model-Eliciting Activities

Abstract

The effective use of open-ended problems requires reliable and high quality instructor feedback and assessment to substantially boost the quality of student learning and work products. Model-Eliciting Activities (MEAs) are open-ended, realistic, client-driven problems set in engineering contexts requiring teams of students to create a generalizable (shareable, reusable, modifiable) mathematical model for solving the client's problem. Two significant challenges are associated with the assessment of student team solutions to MEAs: (1) evaluation reliability among multiple instructors and (2) fidelity to what is valued in engineering practice. In this paper, we describe the dimensions of a new assessment tool used by graduate teaching assistants to assess student team work on MEAs in a required first-year engineering course, and we demonstrate its application to a specific MEA implemented in Fall 2008. Further, we assess the reliability of the tool by comparing its application by new and returning graduate teaching assistants to that of an Expert. Finally, we discuss how the results of this study are informing subsequent revisions to the tool and graduate teaching assistant professional development with MEAs.

I. Introduction

The need for engineering curricula that develops students' teaming and communication skills, proficiency in engineering science and design, and abilities to address open-ended problems replete with ambiguity and uncertainty is well recognized[1,2]. Such curricula should engage students in authentic learning experiences that reflect engineering practice. High quality and reliable feedback and assessment strategies must accompany these learning experiences to ensure that student learning is achieved (e.g. misconceptions are addressed) and the quality of student work increasingly reflects what is valued in engineering practice.

Model-Eliciting Activities (MEAs) are one instructional approach to developing these and other competencies[3,4]. These client-driven, open-ended, team-oriented problems have been implemented in a large (N = 1200-1600) required first-year engineering problem solving and computer tools course since Fall 2002[5,6]. Over 20 different MEAs have been implemented and a number of feedback and assessment strategies have been employed with varying degrees of success[6]. What these strategies lacked was a clear articulation of core elements of performance valued in engineering practice that could be translated into a rubric (and supporting materials) that could be reliably applied by the 18-20 graduate teaching assistants responsible for assessing student work.

As part of a larger study, a panel of engineering experts identified three core elements of performance for student team work on any MEA:

- Appropriateness of the mathematical model. The complexity of the problem must be addressed in the mathematical model.
- Attention to audience. The product should clearly and effectively communicate the model to the client. Share-ability is another term used to describe this idea.

- Generalizability of the product. The product should be a tool that goes beyond being useful to its creators (the student team members) to being useful to others (e.g. the client). Re-usability and modifiability are among the terms used to describe the idea of generalizability[7].

These core elements of performance were translated into the *MEA Feedback and Assessment Rubric* (*MEA Rubric*). This rubric was designed so that it can be applied to student team work on any MEA. What these core elements of performance mean within the context of a given MEA are articulated in an MEA specific *Instructors' MEA Assessment/Evaluation Package* (*I-MAP*).

In this paper, we describe the dimensions of the *MEA Rubric* used by graduate teaching assistants (GTAs) to assess student work on MEAs implemented in a first-year engineering course. Its application to a specific MEA implemented in Fall 2008 is demonstrated and reliability results are presented. We discuss how these results are informing subsequent revisions to the rubric and GTA professional development with MEAs.

Three research questions guide this work:

1. To what extent do GTAs' assessments of student work match an Expert's assessment?
2. Are there differences between new and returning GTAs' assessments of student work as compared to an Expert?
3. What is the nature of the differences between assessments completed by the GTAs and an Expert?

II. *Paper Plane Challenge* MEA

In this study, we will looked at how the GTAs applied the *MEA Rubric* to the *Paper Plane Challenge* MEA. This MEA was the first of three MEAs implemented in Fall 2008. The *Paper Plane Challenge* MEA used in this implementation was a further modification of the problem originally developed by Richard Lesh[8]. This MEA requires teams of students to use their knowledge of competitions and statistics to develop a procedure that competition judges for a paper airplane throwing contest can use to select winning teams. The student teams are provided with the following description of the competition:

> "*During this competition, each competing team will construct one plane that can undergo minor adjustments prior to each throw. The two required throw paths are shown below: the Straight path and the Boomerang path. For each path, there is a fixed target point where the plane should land. Competing teams throw their planes a predetermined number of times along each path. At the conclusion of the competition, one award is made in each of the following categories: Most Accurate, Best Floater, Best Boomerang, and Best Overall. This event is adjudicated by judges for the American Institute of Aeronautics and Astronautics (AIAA) Paper Airplane Competitions.*"

The student teams receive a memo from Mandi Conner, AIAA Education Chair, is which the student team task is described.

> "*In past competitions, the judges for the American Institute of Aeronautics and Astronautics (AIAA) Paper Airplane Competitions have had problems deciding how to select a winner for each of the four*

awards (Most Accurate, Best Floater, Best Boomerang, and Best Overall). The competition judges do not know which measurements to consider from the throw path results to determine who wins each award.

I am asking your team to consider how to use the measurements. Some sample data from last year is provided below.

Write a memo to my attention. In your memo include procedures to determine Most Accurate, Best Floater, Best Boomerang, and Best Overall. Your team does not have to use all the measurements, but you do have to be able to justify your methods. Within the procedure, clearly state the reason for each step, heuristic (i.e. rule), or consideration in your procedure. Use your procedure and the sample data provided to determine the winners in each category. State the winners with quantitative results in your memo."

The students are shown definitions of the throw paths and measurements (Table 1) and are provided with sample team data (Table 2).

Table 1. Measurements Taken During AIAA Paper Airplane Competitions

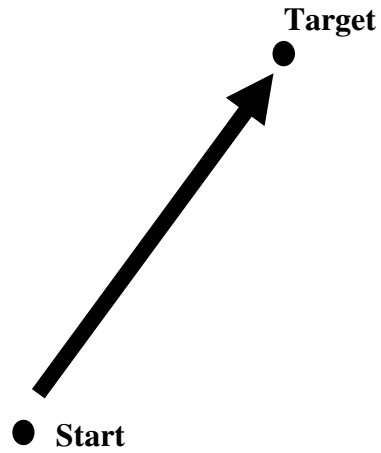
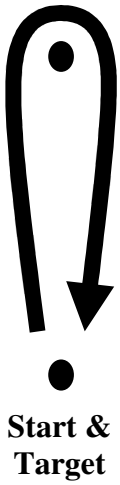
	Straight	Boomerang
		
Path description	Straight line from start to finish	The plane goes out and comes back to the thrower after going around a chair. The chair is 4 meters from the thrower
Amount of time in air	Number of seconds from time of throw to landing	
Length of throw	Straight-line distance from the start point to the landing point	**Made turn**: Distance from starting point to turning point plus distance from turning point to landing point. **Missed turn**: Distance from starting point to landing point
Distance from target	Distance from the landing point to target.	**Made turn**: Distance from landing point to target. **Missed turn**: Distance from landing point to turning point plus distance from turning point to target.

Table 2. Sample Data for *Paper Plane Challenge* MEA

	Straight Path			Boomerang Path		
TEAM	**Amount of time in air (seconds)**	**Length of throw (meters)**	**Distance from target (meters)**	**Amount of time in air (seconds)**	**Length of throw (meters)**	**Distance from target (meters)**
Team 1	3.1	11	1.8	0.7	1.8	6.8
	0.1	1.5	8.7	1.2	3.7	6.7
	2.7	7.6	4.5	2.7	8.4	4.4
Team 2	3.8	10.9	1.7	2.3	8.1	6.1
	4.2	13.1	5.4	0.2	1.6	6.9
	1.7	3.4	8.1	2.1	6.9	5.2
Team 3	4.2	12.6	4.5	---	8.5	5.5
	5.1	14.9	6.7	2.4	7.7	8.7
	3.7	11.3	3.9	0.2	1.9	6.7
Team 4	2.3	7.3	3.3	1.4	4.9	4.9
	2.7	9.1	4.9	2.7	7.2	8.1
	0.2	1.6	9.1	---	---	---
Team 5	4.9	7.9	2.8	2.5	7.7	5.7
	2.5	10.8	1.7	2.1	9.8	9.8
	5.1	12.8	5.7	3.2	10.4	5.8
Team 6	0.2	1.8	8.8	0.1	1.2	8.2
	2.4	10.1	4.6	1.3	4.9	4.9
	4.7	10.3	5.4	1.8	5.5	2.7

By way of instruction to the student teams about the form of the memo, they were told in the activity instructions to use the outline below (Figure 1) to help them organize their response and ensure that the team did not forgotten necessary items.

> "*Use the following outline to help organize your team's response and ensure that your team has not forgotten necessary items. CAUTION: The memo that your team submits should be in narrative form, not in outline form. Items I A-C are typically all covered in the first paragraph and item II is typically in easy-to-follow numbered steps. Item III could be in a combination of paragraph and tabular form, depending on the nature and quantity of the results generated by your team's solution. (NOTE: Your team cannot receive a grade higher than a D if you do not present results (Item III). Why? A client would want to see results. Without results, your team has only attempted part of the task (provided the client with a solution); your team would not have provided evidence that it actually works.).Item IV includes any other requested information.*"

TO: Name, Title
FROM: Team #
RE: Subject

I. Introduction
   A. In your own words, restate the task that was assigned to your team (~1-2 sentences). This is your team's consensus on who the client is and what solution the client needs.
   B. Describe what the procedure below is designed to do or find – be specific (~1-2 sentences)
   C. State your assumptions about the conditions under which it is appropriate to use the procedure. Another way to think about this is to describe the limitations of your procedure.
II. List the steps of the procedure. Provide clear rationales for the critical steps and clarifying explanations (e.g. sample computations) for steps that may be more difficult for the client to understand or replicate.
III. Provide results of applying the procedure to specified data
IV. Other requested information

Figure 1. MEA Solution Outline.

III. Method

A. Setting

In Fall 2008, three MEAs were implemented in a required first-year engineering problem solving and computer tools course with an enrollment of approximately 1200 students. In preparation for these activities, the GTAs were provided with extensive professional development. Prior to the start of the semester, the GTAs employed in the teaching of the course were provided with eight hours of MEA training that included introductory topics such as open-ended problems, understanding first-year students and their reactions to open-ended problems, and classroom implementation of MEAs. Each of these topics is described in more detail by Diefes-Dux, Osburn, Capobianco, and Wood[9]. Particular emphasis was placed on the use of the *MEA Rubric* to assess student work along three dimensions that map back to the three core elements of performance:

- Mathematical Model: Does the mathematical model adequately address the complexity of the problem?
- Audience (Share-ability): Can the client easily reproduce the results using the test case data provided in the MEA?
- Re-usability & Modifiability: Can the client use the model on similar types of data and can the client modify the model for use in similar but different situations?

The GTAs were guided through the assessment of select pieces of prototypical student work on the first MEA to be implemented, the *Paper Plane Challenge* MEA. Following this workshop, each GTA independently practiced assessing five additional pieces of prototypical student work

using the *MEA Rubric* and I-MAP specifically for the *Paper Plane Challenge* MEA. A course instructor reviewed GTAs' assessments, summarized common problems experienced by all GTAs, and provided individual guidance to each GTA. Additional workshop training, assessment practice, and course instructor feedback was provided to the GTAs in preparation for the second and third MEA implemented in Fall 2008.

The GTAs (in teams of 4 and supported by undergraduate assistants) co-facilitated the laboratory implementation of the MEA in which ~30 student teams of 3-4 students developed the first draft of their solution. Following class, each GTA individually assessed the work of 14-15 student teams. Student teams used their GTA's feedback to revise their solutions. Student teams revised their solutions a second time based on feedback received through a blind peer review. This final team solution was graded by the student team's GTA. Feedback provided at this stage was intended to help students perform better on the next MEA.

B. Participants

In Fall 2008, twenty (20) GTAs were employed in the teaching of the first-year problem solving and computer tools course - 13 were new to teaching in this course and 7 had at least one semester of teaching experience in this course. The MEA assessment work of 15 of these GTAs is included in this study - 10 new and 5 returning GTAs. This subset of GTAs was a convenience sampling based on the data analysis available from another study. Of the 5 returning GTAs, 3 were familiar with the *Paper Plan Challenge* MEA from its Fall 2006 implementation in which a different rubric was applied[5]. One of these three returning GTAs also had experience with this MEA from the Spring 2008 implementation in which a rubric similar to the *MEA Rubric* was used.

An Expert independently assessed Fall 2008 student team MEA solutions. The Expert was a doctoral student in Engineering Education with 7 years of teaching experience in the first-year engineering program and 4 years of experience with research on MEAs, including the development of the *MEA Rubric*.

C. Data Collection & Analysis

MEAs are conducted via a web-based interface connected to a database system. This overall system manages the organization and sequencing of the various stages of a MEA implementation; it also facilitates the interactions between individual students, their team, their GTA, and their peers. All student, GTA, and peer responses and interactions during an MEA are stored in the database for later review by the MEA research team. For this study, the data consists of students' first draft solutions to the *Paper Plane Challenge* MEA and select GTAs assessment of that student team work using the *MEA Rubric*.

The Expert independently assessed 7 team MEA solutions per participating GTA, for a total of 105 graded pieces of student work. Only 7 of the 14-15 pieces of student team work were assessed per GTA as for some GTAs this was the maximum number that had undergone data analysis for another study. The selection of 7pieces of work provided an equal number of pieces of analyzed team work from each GTA. When more than 7 pieces of student work had been

assessed by the Expert for a given GTA, 7 pieces were randomly selected for inclusion in this study. The Expert applied the *MEA Rubric* to each piece of student team work. The Expert often constructing sub-rubrics for items that were underspecified in the *MEA Rubric* or I-MAP; this was done by the Expert to ensure internal consistency across student team work. These sub-rubrics provided an opportunity to explore differences in GTA and Expert assessments of student team work. It is recognized that the GTAs may or may not have developed similar sub-rubrics to aid in their own assessment of student team work.

For each item of the *MEA Rubric*, a simple quantitative comparison is made between the GTAs' and Expert's assessment of the student work. Simple comparisons are also made between new and returning GTAs' assessments and the Expert's assessments. Differences in means were statistically analyzed using t-tests; while differences in variance were analyzed using F-tests. The Expert's sub-rubrics were in some instances used to investigate differences between the GTAs' and Experts' use of the *MEA Rubric*.

IV. Results & Discussion

In this section, we present each *MEA Rubric* item with a description of what constitutes high quality work for that item. Results of the simple comparisons and statistical analyses are shown and discussed. Results of investigations into differences between the GTAs' and Expert's assessments of the student team work are also given and discussed. A sample piece of prototypical student team work on the *Paper Plane Challenge* MEA and the Expert's assessment of this work are provided in Appendix A and B, respectively.

When applying the *MEA Rubric* to student team work on a MEA, the GTA selects a maximum level of achievement for each rubric item. For each item, the maximum level of achievement is 4, though the minimum level varies from 0 to 3. In practice (in the classroom), the minimum level of achievement across all items in a dimension becomes the level of achievement for that dimension. The minimum level of achievement across all three dimensions is the level of achievement for the given piece of student team work; this value can be translated into a grade on the team's MEA solution.

A. Mathematical Model

The Mathematical Model dimension of the *MEA Rubric* is broken down into three items: complexity, rationales, and accounting for data types. Each is discussed below.

<u>Complexity</u>

Mathematical Model complexity, for the *Paper Plane Challenge* MEA, is focused around three issues for each of the four competition award categories:

- is a mathematical model idea present,
- is there a clear definition of what constitutes the "best" or "most", and
- is the model free of errors and does the actual model map to the stated definition of "best" or "most".

For example, the student team must clearly (and separately from their procedure) define what they mean by "Best Floater" and then develop an error-free mathematical model that operationalizes that definition. In the sample student team work (Appendix A), the "Best Floater" is defined to be the plane that floats for the longest period of time. The team fails to include in the definition that both the straight and boomerang path throws are considered, though this is what actually occurs in the mathematical model.

Complexity is assessed by the GTAs through the *MEA Rubric* item shown in Table 3. In training, GTAs were instructed to use a sub-rubric in which they assign one point per issue per competition award category, for a total of 12 possible points. Points are then to be translated to a level of achievement (as shown in Appendix B).

Table 3. Mathematical Model Complexity Rubric Item

Level	Description
4	The procedure fully addresses the complexity of the problem.
3	The procedure moderately addresses the complexity of the problem and/or contains embedded errors.
2	The procedure only somewhat addresses the complexity of the problem and/or contains embedded errors.
1	The procedure does not address the complexity of the problem and/or contains significant errors.
0	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 direct user does not constitute turning in a product.

As shown in Tables 4, GTAs assigned the same level as the Expert to only 35 out of 105 (33.3%) pieces of student work (sum of values on the diagonal). The difference between the GTAs' and Expert's assessment of the student work is significant ($p < 0.001$). GTAs tended to assign higher levels than the Expert, with 60 of the 105 pieces of student team work being rated at a higher level by the GTAs than the Expert. Collectively, the quality of the first draft solutions was low in terms of addressing the complexity of the problem, with an average level of achievement of only 1.21 being assigned by the Expert. This means that the majority of responses only minimally began to address the complexity of the problem. GTAs assigned an average level of 1.92.

Table 4. Mathematical Model Complexity Level Assigned by Expert versus GTAs

		GTA Level				
		1	**2**	**3**	**4**	**Total**
Expert Level	**1**	27	36	20	2	85
	2	8	8	1	1	18
	3	0	2	0	0	2
	4	0	0	0	0	0
	Total	35	46	21	3	105

In the most common solution type (52/105), teams had mathematical model ideas for each of the four competition award categories, but no definitions of "best" or "most" (which implies no agreement between the model and the definition). Of these 52 pieces of student team work, 37 (71%) were assigned a higher level by the GTAs than the Expert. An additional 20 pieces of student team work had 3 models included in the solutions but with no definitions. Of these 20, the GTAs assigned a higher level score than the Expert on 13 pieces of student team work; these differences were spread over 7 GTAs. This would indicate that the GTAs are identifying definitions within the student team work when the Expert is not or allowing weak definitions when the Expert is not. This was an issue identified by the course instructor during GTA training, where it was seen that GTAs often allowed a procedural step to serve also as the definition for the award category.

Table 5 shows the percent of student work assessed by the GTAs as easier, harder or the same as the Expert according to the GTAs' experience. An E+3 means that the GTA assigned the student work a level three higher than the Expert (E); a 0 means the GTA and Expert assigned the student work the same level; and an E-3 means that the GTA assigned the student work a level three lower than the Expert. As shown in Table 5, the returning and new GTAs' assessments of student work were quite similar to each other, with the levels they assigned to student work being higher than the Expert's. The mean difference between the returning GTAs assigned level of achievement for complexity and the Expert's was 0.74; the new GTAs' average difference was 0.70; this difference is not statistically significant ($p = 0.54$). The standard deviation of the difference between the GTAs' and Expert's assigned level of achievement was 0.82 and 1.03 for returning and new GTAs, respectively. This difference is slightly significant ($p = 0.02$). The GTAs' experience has a slight impact on the variance in the complexity level assigned to student work; with the new GTAs showing greater variance in their assessments of complexity.

Table 5. Percentage of Mathematical Model Complexity Levels Assigned Relative to the Expert

GTA Status	Number of Pieces of Student Work Assessed	GTA Easier than Expert				GTA Harder than Expert		
		E+3	E+2	E+1	0	E-1	E-2	E-3
Returning	35/105	0%	20%	37%	40%	3%	0%	0%
New	70/105	3%	20%	34%	30%	13%	0%	0%

Two problems were identified. First was the tendency for GTAs to over-estimate the level of achievement of student work. Of the 15 TAs, only 1 (new) GTA consistently underestimated the complexity of the students' mathematical models. Five of the 15 GTAs over-estimated the level of complexity of the mathematical model of all of their student work. Four additional GTAs over-estimated the level achieved on 6 of their 7 pieces of student work. The second problem identified was a lack of internal consistency, or randomness in assigning a level. One (new) GTA, for example, had 6 teams, each of which had a procedure that only contained the four mathematical model ideas associated with the four competition award categories but no definitions of "best" or "most". These teams were all assigned a level 1 by the Expert. This GTA assigned one a level 2, three a level 3, and two a level 4, demonstrating a lack of internal consistency in how this GTA applied the evaluation rubric.

Rationales

Student teams are required to rationalize the critical steps of their procedure. This not only encourages the students to more carefully think through their approaches to solving the problem but it allows instructors to see the students thinking about the mathematics in the problem. The presence or absence of rationales is assessed through the true-false statement, "The procedure is supported with rationales for critical steps in the procedure", where true corresponds to level 4 achievement and false corresponded to a level 3. GTAs are provided the sample list below of items that require rationales, though it is not an all-inclusive list.

- Dropping out individual team throws (e.g. drop lowest of three times in air for each team)
- Dropping out a type of throw
- Dropping out a type of data
- Methods for breaking ties
- Methods for handling missing data
- Steps that involve combining data types

Because many teams may or may not perform some of the steps above, it is difficult to formalize this rubric item into an objective format. The Expert assessed this question using a 3-point Likert scale, corresponding to "well rationalized", "partially rationalized", and "no or minimal rationales provided". Teams who were well or partially rationalized were marked at a level 4, while teams with no or minimal rationales were marked at a level 3.

An example of a rational for a critical step in a procedure can be seen in the sample student work in Appendix A. The procedure for the "Best Boomerang" contains the step "All teams who throw below four meters are eliminated from contention". The rationale for this step states that "Throws must at least reach the chair, so a minimum of four meters is necessary."

As shown in Table 6, the GTAs' assessments of rationales in students' work aligned well with the Expert, with alignment on 77 out of 105 (73.3%) pieces of student work. The average level of achievement with regards to rationales as assigned by the Expert was 3.20; for the GTAs, it was 3.24. There is no significant difference ($p = 0.45$) between the GTAs and the Expert. Two GTAs account for 7 of the 12 pieces of student team work for which the Expert found appropriate rationales and the GTAs did not. Two different GTAs account for 6 of the 16 pieces of student team work in which the Expert did not find adequate rationalization and the GTAs did.

Table 6. Rationales Level Assigned by Expert versus GTAs

		GTA Level		
		3	**4**	**Total**
Expert Level	**3**	68	16	84
	4	12	9	21
	Total	80	25	105

As shown in Table 7, the mean difference between returning GTAs' assessments of rationales in student work and the Expert's was 0.20, while the difference between new GTAs' assessments

and the Expert's was -0.04. There is a slightly significant difference (p = 0.03) between the returning and new GTAs. The standard deviation of the difference between the returning and new GTAs' assessments versus the Expert's was 0.53 and 0.49, respectively. The standard deviations are not significantly different for the returning and new GTAs (p = 0.60). Here, the new GTAs were slightly better able to use the *MEA Rubric* to assess the presence of rationales.

Table 7. Percentage of Rationale Levels Assigned Relative to the Expert

GTA Status	Number of Pieces of Student Work Assessed	GTA Easier than Expert		GTA Harder than Expert
		E+1	0	E-1
Returning	35/105	26%	69%	6%
New	70/105	10%	76%	14%

While there is generally agreement between the GTAs and the Expert, it should be noted that very few teams provided adequate rationales. Students, while working on their first draft solution to an MEA, tend to focus on simply producing *a* procedure at the expense of providing rationales for their procedural steps. This results in many first drafts being produced with no rationales. If student work were to include more rationales, it would not be surprising to see the agreement between the GTAs and the Expert decrease.

Accounting for Data Types

As part of each MEA, teams are provided with a sample set of data that allows the student team to self-assess their procedure. Student teams are required to either use or justify not using all of the data types provided. For the *Paper Plane Challenge* MEA, teams are provided with three data types for each of the two throw paths (straight and boomerang). GTAs assess this item through the true-false statement, "The procedure takes into account all types of data provided to generate results OR justifies not using some of the data types provided", where true corresponded to a level 4 and false corresponded to a level 3. During training, the TAs were instructed to view the data set as including six data types (three types per throw path) and that all six must be used or justified not being used at some point in the overall the procedure. The Expert assessed this data types as a set of 24, three data types corresponding to Time in Air, Length of Throw, and Distance from Target for each of the two paths for each of the four competition award categories. For each competition award category it was noted whether each all data types were used, some were used or not used, as well as if there was a justification associated with its not being used. The Expert mapped the 24 type indicator to the 6 type indicator used by the GTAs

In the sample student work (Appendix A), the team explicitly used all of the data types except the Straight Throw – Length of Throw measurement (as indicated in Appendix B). However, this measurement would likely be needed to compute the "velocity" used in the tie-breaker for "Best Overall", though this is never stated.

For 71 out of 105 (67.6%) pieces of student team work, the GTAs assigned the same level as the Expert (Table 8). The average level of achievement with regards to accounting for data types as

assigned by the Expert was 3.27; for the GTAs, it was 3.30. There is no significant difference ($p = 0.49$) between the GTAs and the Expert.

The 19 pieces of student team work for which the GTA felt that data types were all used that the Expert did not were distributed over 8 GTAs (5 returning, 3 new). Five (3 returning, 2 new) GTAs assigned 3 pieces of student work a level 4 when the Expert assigned a level 3. The 15 pieces of student team work for which the Expert felt all data accounted but the GTAs did not were distributed across 8 new GTAs. Two of these new GTAs assigned 3 pieces of students work a level 3 when the Expert assigned a level 4. It would seem that returning GTAs better align with the Expert; new GTAs are more inclined to assign a lower level to student work.

Table 8. Accounting for Data Types Level Assigned by Expert versus GTAs

		GTA Level		
		3	**4**	**Total**
Expert Level	**3**	58	19	77
	4	15	13	28
	Total	73	32	105

As shown in Table 9, the mean difference between returning GTAs' assessments of accounting for data types in student work and the Expert's was 0.34, while the difference between new GTAs' assessments and the Expert's was -0.11. There is a significant difference ($p < 0.001$) between the returning and new GTAs. The standard deviation of the difference between the returning and new GTAs' assessments versus the Expert's was 0.48 and 0.55, respectively. The difference in variance is not significantly different for the returning and new GTAs ($p = 0.38$). It is evident that returning GTAs often assign a level 4 when not all data types are accounted for, while new GTAs seem unsure about whether data types are accounted for or not. It may be that the GTAs do not understand how to use the sub-rubric or are not using the sub-rubric to determine whether data types are present or not.

Table 9. Percentage of Accounting for Data Types Levels Assigned Relative to the Expert

GTA Status	**Number of Pieces of Student Work Assessed**	**GTA Easier than Expert**		**GTA Harder than Expert**
		E+1	**0**	**E-1**
Returning	35/105	34%	66%	0%
New	70/105	10%	69%	21%

B. Audience (Share-ability)

The Audience (Share-ability) dimension of the *MEA Rubric* is broken down into three items: results presented, readability, and no extraneous information. Each is discussed below.

Results Presented

Teams are required to present the results of applying their procedure on the data set provided. For the *Paper Plane Challenge* MEA, the teams were expected to present not only which competing team wins in each award category, but also the quantitative values associated with those winners. GTAs assess this item through the true-false statement, "Results from applying the procedure to the data provided are presented in the form requested", where true corresponded to a level 4 and false corresponded to a level 1. In contrast to most of the questions where level differences are typically only 1 level apart (e.g., levels 4 and 3 for most true-false questions), the results item heavily penalizes teams who do not include results.. The importance of providing results as evidence that the procedure works is emphasized numerous times throughout MEA instruction. In the sample student work (Appendix A), the team winner in each competition category is provided with a quantitative result derived from applying the procedure.

The Expert assessed this item using an 8-point accounting (two points for each of the four competition award categories) of the presence of results. One point was assigned if a winner was identified, the second was assessed if quantitative results were present. These were then translated to the same scale used by the GTAs; all 8 points had to be accrued for a level 4 assignment. Generally, the GTAs aligned well with the Expert, as can be seen in Table 10 and Table 11

Table 10. Results Presented Level Assigned by Expert versus GTAs

		GTA Level		
		1	**4**	**Total**
Expert Level	**1**	26	26	52
	4	2	51	53
	Total	28	77	105

Table 11. Percentage of Results Presented Levels Assigned Relative to the Expert

GTA Status	**Number of Pieces of Student Work Assessed**	**GTA Easier than Expert**		**GTA Harder than Expert**
		E+3	**0**	**E-3**
Returning	35/105	29%	71%	0%
New	70/105	23%	74%	3%

The Expert and the GTAs assigned the same Level to 77 out of 105 (73.3%) pieces of student team work. The mean level assigned by the Expert was 2.51; for the GTAs is was 3.20. There was a significance difference ($p < 0.001$) between the GTAs and the Expert, possibly due to the level 1 or 4 assignment structure for this item. There were 26 out of 105 instances of GTAs indicating that teams had presented results in the form requested when the Expert identified incomplete results. Ten TAs (6 new, 4 returning) were involved in this misidentification of results. Of the 26 pieces of student work, 17 had all the team winners identified but 10 were missing all of the quantitative results and 7 only had partial quantitative results. Six GTAs (4

new, 2 returning) account for 21 of the misidentified presence of complete results, meaning they each had 3 or 4 out of their 7 teams being incorrectly assessed. A more detailed look at the assessments of six GTAs is shown in Table 12.

Table 12. GTAs Accounting for the Most Differences with Expert on Results Presented

Pseudonym	GTA Status	Number of Instances Out of 7 Pieces of Student Work	Trends Seen in Responses
Joshua	Returning	4 higher than Expert	Any partial results were assigned a a level 4
Emily	New	4 higher than Expert	No quantitative results were necessary to be assigned a level 4
		1 lower than Expert	All results were present (3 similar pieces of student work were scored correctly)
David	New	4 higher than Expert	No quantitative results were necessary to be assigned a level 4
Joseph	New	3 higher than Expert	No quantitative results were necessary to be assigned a level 4
Alexander	New	3 higher than Expert	Partial quantitative results scored a level 4 (1 identical piece of student work was graded correctly)
Ethan	Returning	3 higher than Expert	Any partial results were assigned a level 4.

There were two instances of GTAs indicating incomplete results when the Expert found them to be complete. In each instance, the GTAs made no comment to the student teams about what was missing. These are more than likely situations of the GTA accidentally selecting the incorrect response bubble. Emily (in Table 10) was one of these GTAs; she had assessed three other similar pieces of results correctly.

There were no differences found between the returning and new GTAs' assessments of the student work as compared to the Expert. The mean difference between the returning and new GTAs' assessments and the Expert's were 0.60 and 0.86, respectively ($p = 0.37$). The standard deviation of the difference between the returning and new GTAs' assessments and the Expert were 1.38 and 1.41, respectively. There was no significant difference in the variance ($p = 0.90$).

Readability

Students are asked to create a clear, coherent, concise memo to the client describing their procedure. As part of developing a clear memo, the client should be able to understand the procedure and replicate the results using the data provided. GTAs assess this item through the three level rubric item shown in Table13.

Table 13. Readability Rubric Item

Level	Description
4	The procedure is easy for the client to understand and replicate. All steps in the procedure are clearly and completely articulated.
3	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.
2	Does not achieve the above level.

The Expert assessed this item using 6 weighted sub-items. Four sub-items, worth 0.5 points each, assessed if it was clear which data was being used (e.g., straight versus boomerang data) in the determination of the winners of each of the four competitions. The other two sub-items, worth 2 points each, assessed whether 1) the steps in the procedure were easy to understand in terms of what generally the steps were supposed to accomplish and 2) the steps were replicable, as noted by the absence of subjective language. Each of these two sub-items was rated on a "Yes-Sort of-No" scale, weighted as indicated and a sum score was taken for a total of 6 possible points. Teams with 5 points or more received a level 4; teams with more than 2 points earned a level 3, and teams with less than or equal to 2 points earned a level 2.

The GTAs and Expert assigned the same level to 49 of 105 (46.7%) pieces of student team work (Table 14). The Expert assigned a higher level to 48 (45.7%) pieces of student team work. The mean level assigned by the Expert was 3.29, while the mean level assigned by the GTAs was 2.81. This significant difference $p < 0.001$ was found between the GTAs and Expert.

Table 14. Readability Level Assigned by Expert versus GTAs

		GTA Level			
		2	**3**	**4**	**Total**
Expert Level	**2**	5	7	0	12
	3	13	37	1	51
	4	10	25	7	42
	Total	28	69	8	105

New GTAs seem to have poor alignment with the Expert (Table 15), while returning GTAs had a stronger alignment with the Expert. The mean difference between the returning and new GTAs' assessments and the Expert's were -0.23 and -0.60, respectively. A slight significant difference ($p = 0.02$) was found between the returning and new GTAs. The standard deviation of the difference between the returning and new GTAs' assessments and the Expert's were 0.69 and 0.79, respectively. No significant difference ($p = 0.40$) was found in the variance between the returning and new GTAs. The lower scores provided by the new GTAs is most likely attributed to the fact that new GTAs are not familiar with MEAs and have to work at understanding student responses. They have also not read as many responses as experienced GTAs and may not fully understand the breadth of response quality. As such, they may hold a higher standard for what they consider "understandable". While it is clear that both new and returning GTAs hold

students to a higher standard of understandability than the Expert, it is unclear what this standard entails. More work needs to be done to identify the unspecified issues GTAs are using to evaluate the readability of the procedures.

Table 15. Percentage of Readability Levels Assigned Relative to the Expert

GTA Status	Number of Pieces of Student Work Assessed	GTA Easier than Expert			GTA Harder than Expert	
		E+2	E+1	0	E-1	E-2
Returning	35/105	0%	9%	66%	20%	6%
New	70/105	0%	7%	37%	44%	11%

Regardless of their standard, GTAs avoid identifying work as having achieved a level 4 readability. GTAs only assigned 8 responses a level 4 as compared to the Expert who assigned 42 responses a level 4. As this is feedback on the students' first draft, it is speculated that this has more to do with setting a high standard to encourage improvement than it does with the actual quality of the work. At this stage, students generally are producing fragmented responses that lack a sense of flow. Though these responses may not be as polished as the final products they will later produce, they are generally complete and understandable, but their lack of polish gives them an initial outward appearance of not being level 4 work, and thus GTAs may not be assigning level 4s. It is also possible that GTAs may be assessing this rubric item last (it is physically item 6 out of 7 on the actual interface) and using it as a means to adjust the grade for problems that should have been noted with other rubric items.

No Extraneous Information

Students are asked to create a clear, coherent, concise memo to the client describing their procedure. As part of this, they are told to avoid including extraneous information in their memo. GTAs assess this item through the true-false statement, "There is no extraneous information in the response" where true corresponded to a level 4 and false corresponded to a level 3. The Expert assessed the presence or absence of extraneous information using four "Yes-Sort of-No" sub-rubric items. The four items indicated if responses: 1) retained an outline format in the response, 2) mentioned MATLAB, Excel, or other software package, 3) told how to calculate a basic statistic (e.g., "To find the average distance, take the three distances and divide by three."), and 4) contained other unnecessary text (e.g. reiterating, providing details, or changing the rules of the competition). While brevity is encouraged, the Expert did not assess whether responses were overly wordy. Only one team made reference to a software package, and their GTA correctly identified this as extraneous information, providing the feedback, "References to software tools are extraneous information."

The GTAs assigned the same level as the Expert to 66 out 105 (62.9%) pieces of student team work (Table 16). The mean Expert level assigned was 3.61; for the GTAs, it was 3.66. The difference between the GTAs and Expert was not significant ($p = 0.43$). Again, there was no difference between the new and returning GTAs (Table 17). The mean difference between the returning and new GTAs' assessments and the Expert were 0.00 and 0.07, respectively ($p = 0.58$). The standard deviation of the difference between the returning and new GTAs'

assessments and the Expert's were 0.64 and 0.60, respectively. There was no significant difference in the variance ($p = 0.61$). There were 22 instances identified by 13 TAs as having no extraneous information when the expert identified the presence of extraneous information. Two TAs accounted for 7 of these instances (3 or 4 of their 7 responses each). While there were a large number of pieces of student work being assessed differently by the GTAs and Expert, there was no discernable pattern to indicate what these GTAs had difficulty identifying. For as many instances of any particular combination of extraneous information present, there were almost always that many GTAs not detecting that extraneous information was present (Table 18). Of the seven possible combinations of extraneous information (excluding software tools, as this was not present in more than one piece of student work), the two GTAs with the most difficulties detecting extraneous information had instances that fell in five of the combinations.

Table 16. Extraneous Information Level Assigned by Expert versus GTAs

		GTA Level		
		3	**4**	**Total**
Expert Level	**3**	19	22	41
	4	17	47	64
	Total	36	69	105

Table 17. Percentage of Extraneous Information Levels Assigned Relative to the Expert

GTA Status	**Number of Pieces of Student Work Assessed**	**GTA Easier than Expert**		**GTA Harder than Expert**
		E+1	**0**	**E-1**
Returning	35/105	20%	60%	20%
New	70/105	21%	64%	14%

Table 18. Combinations of Extraneous Information and Instances of Expert and GTA Misalignments

Extraneous Information Type[a]			**Number of Total Instances**	**Number of Instances of Misalignments**	**Number of Different GTAs with Misalignment**
Outline Format	**Described Basic Statistics**	**Other**			
		X	9	4	3
	X		10	7	7
	X	X	4	1	1
X			7	4	4
X		X	4	2	2
X	X		4	3	3
X	X	X	2	1	1
17/105	20/105	19/105	40	22	

[a] An X in the Extraneous Information Types columns indicates the presence of this type in a given combination. So, in the first row, the combination contains only Other.

C. Re-usability & Modifiability

A re-usable and modifiable procedure is robust. It can be used by the client for new but similar situations (i.e. it is r*e-usable*); it can be modified easily by the client for slightly different situations (i.e. it is modifiable). The Re-usability and Modifiability of a procedure is primarily assessed through a team's introductory paragraph(s). It is here that teams are instructed to explain who the client is and what the client needs and state the assumptions and limitations of their procedure (Figure 1). This dimension is assessed by GTAs through a single rubric item, shown in Table 19.

Table 19. Re-usability & Modifiability Rubric Item

Level	Description
4	The procedure not only works for the data provided but is clearly re-usable and share-able. Re-usability and share-ability 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.
3	The procedure works for the data provided and *might* be re-usable and share-able, but it is unclear whether the procedure is re-usable and share-able because assumptions about the situation and/or the types of data that the procedure can be applied to are not clear or not provided.
2	Does not achieve the above level.

The Expert assessed this item by identifying the presence or absence of the following sub-items: 1) identification of the client, 2) identification of the client's need (e.g. a procedure as opposed to just rankings of the provided dataset), 3) stating that the procedure is designed to identify award winners in four categories, 4) stating that these awards are based on data from two throw paths, 5) indicating that there are multiple throws per path, and 6) stating that the data consists of three measurements. Collectively, each of these items contributes to the client understanding not only what the procedure can and cannot be used for, but also to understanding what other situations the procedure can be adapted to fit. The Expert assigned up to 2 points for each sub-item, for a score of up to 12 points. Teams with more than 8 points were assigned a level 4, 6-8 points earned a level 3, and fewer than 6 points received a level 2. In training, the GTAs were instructed to look for these items in the context of the outline items I.A-C (Figure 1) and were shown an example in the I-MAP of how this might be written in a student team solution. However, the GTAs were not explicitly told how to translate the presence or absence of these items into a level of achievement.

In the sample student work (Appendix A), the client (AIAA judges) and the client's needs for a procedure to determine the winners in the four award categories are clearly identified. However, the student team does not provide any detail about the data that must be available to use the procedure.

The GTAs and the Expert assigned the same level to 54 out 105 (51.4%) pieces of student work (Table 20). The mean level assigned by the Expert was 2.68; the mean level assigned by the GTAs was 2.69. No significant difference was found between the GTAs and Expert ($p = 0.90$). Assigned levels of achievement are understandably low for this *MEA Rubric* item – the Expert only assigned a level 4 to only 8 pieces of student work. Again, on a first draft solution to a

MEA, most student teams focus on the development of their procedure at the expense of a proper introduction designed to frame the scope of the problem. Of the 55 student team responses the Expert assessed to be of level 3 quality, 37 accurately identifying the client, the client's need for a procedure, and that the procedure was designed to find the winners of the four competition award categories.

Table 20. Re-usability & Modifiability Level Assigned by Expert versus GTAs

		GTA Level			
		2	**3**	**4**	**Total**
Expert Level	**2**	18	21	3	42
	3	20	34	1	55
	4	1	5	2	8
	Total	39	60	6	105

The difference between the returning and new GTAs' assessments and the Expert's assessments appear quite similar (Table 21). The mean difference between returning and new GTAs' assessments of the re-usability and modifiability of student team work and the Expert's was 0. 11 and -0.04, respectively. No significant difference was found between returning and new GTAs ($p = 0.37$). The standard deviation of the difference between the returning and new GTAs' assessments versus the Expert's was 0.90 and 0.71, respectively. No significant difference ($p = 0.10$) in the variance of the returning and new GTAs assessments was found.

Table 21. Percentage of Re-usability & Modifiability Levels Assigned Relative to the Expert

GTA Status	**Number of Pieces of Student Work Assessed**	**GTA Easier than Expert**			**GTA Harder than Expert**	
		E+2	**E+1**	**0**	**E-1**	**E-2**
Returning	35/105	6%	26%	46%	20%	3%
New	70/105	1%	19%	54%	26%	0%

Of the 15 GTAs in this study, 13 had assessed one of their seven pieces of student team work at one level higher than the Expert. Three (3) of those GTAs accounted for 9 of the 22 pieces of student work assessed at a the E+1 level. All 15 TAs had at least 1 piece of student work align with the Expert's assessment. Eleven (11) GTAs scored at least 1 piece of student work one level lower than the Expert, with 4 of those GTAs accounting for 16 of the 25 total pieces of student work assessed at the E-1 level.

While there were only 8 instances of a level 4 response identified by the Expert, it seems that GTAs do not tend to select a level 4. Only 6 total responses were marked as level 4 by the GTAs and of those only 2 were deemed to be at a level 4 by the Expert.

V. Summary of Findings

Table 22 summarizes all of the findings from this study. A significant difference was found between the mean level assigned by the GTAs and the Expert for the *MEA Rubric* items Mathematical Model – Complexity, Audience – Results, and Audience – Readability.

Table 22. Summary of Findings

Item	Participants	Mean	p-value	Standard Deviation	p-value
Mathematical Model					
Complexity (level 1 – 4)	**Expert**	1.21	$p < 0.001$	---	---
	All GTAs	1.92		---	
	GTA_{ret} – Expert	0.74	$p = 0.54$	0.82	$p = 0.02$
	GTA_{new} - Expert	0.70		1.03	
Rationales (level 3 or 4)	**Expert**	3.20	$p = 0.45$	---	---
	All GTAs	3.24		---	
	GTA_{ret} – Expert	0.20	$p = 0.03$	0.53	$p = 0.60$
	GTA_{new} - Expert	-0.04		0.49	
Accounting for Data Types (level 2 – 4)	**Expert**	3.27	$p = 0.49$	---	---
	All GTAs	3.30		---	
	GTA_{ret} – Expert	0.34	$p < 0.001$	0.48	$p = 0.38$
	GTA_{new} - Expert	-0.11		0.55	
Audience					
Results (level 1or 4)	**Expert**	2.51	$p < 0.001$	---	---
	All GTAs	3.20		---	
	GTA_{ret} – Expert	0.60	$p = 0.37$	1.38	$p = 0.90$
	GTA_{new} - Expert	0.86		1.41	
Readability (level 3 or 4)	**Expert**	3.29	$p < 0.001$	---	
	All GTAs	2.81		---	
	GTA_{ret} – Expert	-0.23	$p = 0.02$	0.69	$p = 0.40$
	GTA_{new} - Expert	-0.60		0.79	
No Extraneous Information (level 3 or 4)	**Expert**	3.61	$p = 0.43$	---	---
	All GTAs	3.66		---	
	GTA_{ret} – Expert	0.00	$p = 0.58$	0.64	$p = 0.61$
	GTA_{new} - Expert	0.07		0.60	
Re-usability & Modifiability					
Re-usability & Modifiability (level 2 – 4)	**Expert**	2.68	$p = 0.90$	---	---
	All GTAs	2.69		---	
	GTA_{ret} – Expert	0.11	$p = 0.37$	0.90	$p = 0.10$
	GTA_{new} - Expert	0.04		0.71	

A significant difference was found between the mean difference between returning and new GTAs and the Expert for the *MEA Rubric* item Mathematical Model – Accounting for Data Types. A slight significant difference was found between the mean difference to Expert level

assigned by returning and new GTAs for Mathematical Model – Rationales and Audience – Readability. The slight difference was also found in the variance of the difference between returning and new GTAs' and the Expert's assessment for Mathematical Model – Complexity.

VI. Conclusions and Next Steps

In this study, we investigated the extent to which GTAs in a first-year engineering course could reliably apply a rubric that was designed to articulate three core elements of performance that were identified as being valued in engineering practice. Across the three dimensions of the *MEA Rubric* with its 7 items, the GTAs and Expert assessments aligned for 33.3 to 73.3% of the pieces of students work analyzed. Statically significant differences between the GTAs and Expert's assessments were found for 3 of the 7 items. Statically significant differences between the returning and new GTAs were found for 1 of the 7 items; a slight significant difference was found for 2 additional items. To some extent, the nature of differences between the GTAs' and the Experts' assessments were identifiable, particularly for the Mathematical Model - Complexity and Audience - Results Presented items.

Clearly more needs to be done to improve the GTA's reliability including revisiting the GTA training and revising the *MEA Rubric* and MEA specific I-MAPs. This requires further investigations into the lack of reliability. For instance, it is unclear what their internal criteria are for Audience - Readability. In a parallel study, it was found that GTAs have difficulty differentiating the Audience (Share-ability) and Re-usability & Modifiability dimensions[10]. Confusion over what these dimensions mean and how they are different is likely leading to reliability issues within these dimensions. As a result, the *MEA Rubric* and MEA specific I-MAPs are being revised.

Some issues specifically about the use of the *MEA Rubric* as it is applied to the *Paper Plane Challenge* MEA can be addressed in the GTAs training. The GTAs need more training with the point system used to assess Mathematical Model - Complexity. They also need more guidance in identifying the presence or absence of quantitative results. The I-MAP for the *Paper Plane Challenge* MEA could also be revised to include the sub-rubrics used by the Expert, such as the list of the things that are considered extraneous information and can be expected to appear in students' work on this MEA. Further investigations are needed to understand how the GTAs apply the *MEA Rubric*.

Acknowledgements

This work was made possible by a grant from the National Science Foundation (DUE 0535678). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

Bibliography

1. National Academy of Engineering. (2004). *The engineering of 2020: Visions of engineering in the new century.* Washington, DC: The National Academies Press.
2. Accreditation Board for Engineering and Technology (ABET) (2007). *Criteria for Accrediting Programs in Engineering*. Accreditation Board for Engineering and Technology, Baltimore, MD. Retrieved April 25, 2008, from http://www.abet.org/forms.shtml#For_Engineering_Programs_Only
3. Zawojewski, J. S., Hjalmarson, M.A., Bowman, K, & Lesh, R. (2008). "Chapter 1: A Modeling Perspective on Learning and Teaching in Engineering Education." In Zawojewski, J. S., Diefes-Dux, H., & Bowman, K. (Eds.) (in press). *Models and modeling in Engineering Education: Designing experiences for all students.* Rotterdam, the Netherlands: Sense Publishers.
4. Diefes-Dux, H.A., Moore, T., Zawojewski, J., Imbrie, P.K., and Follman, D. (2004). "A Framework for Posing Open-Ended Engineering Problems: Model Eliciting Activities," Frontiers in Education Conference, Savannah, GA.
5. Diefes-Dux, H.A. and Imbrie, P.K. (2008). "Chapter 4: Modeling Activities in a First-Year Engineering Course" In Zawojewski, J. S., Diefes-Dux, H., & Bowman, K. (Eds.) *Models and modeling in Engineering Education: Designing experiences for all students.* Rotterdam, the Netherlands: Sense Publishers.
6. Diefes-Dux, H.A., Hjalmarson, M., Zawojewski, J., and Bowman, K. (2006). "Quantifying Aluminum Crystal Size Part 1: The Model-Eliciting Activity," *Journal of STEM Education: Innovations and Research*, 7(1&2):51-63.
7. Lesh, R., Hoover, M., Hole, B., Kelly, A., and Post, T., "Principles for developing thought-revealing activities for students and teachers," *Handbook of Research Design in Mathematics and Science Education*, Mahwah, NJ: Lawrence Erlbaum, pp. 591-645, 2000.
8. Zawojewski, J. S., Diefes-Dux, H., & Bowman, K. (Eds.) (2008). *Models and Modeling in Engineering Education: Designing Experiences for All Students.* Rotterdam, the Netherlands: Sense Publishers.
9. Diefes-Dux, H.A., Osburn, K., Capobianco, B., and Wood, T. (2008). "Chapter 12: On the Front Line – Learning from the Teaching Assistants" In Zawojewski, J. S., Diefes-Dux, H., & Bowman, K. (Eds.) (in press). *Models and modeling in Engineering Education: Designing experiences for all students.* Rotterdam, the Netherlands: Sense Publishers
10. Cardella, M.E., Diefes-Dux, H.A., Verleger, M.A., Oliver, A. (2009). "Insights Into The Process of Providing Feedback to Students on Open-Ended Problems," 2009 ASEE National Conference Proceedings, Austin, TX.

Appendix A

SAMPLE STUDENT TEAM WORK

TO: Mandi Conner; Education Chair, American Institute of Aeronautics and Astronautics
From: Team 9
Subject: Purdue Paper Airplane Competition

We have found a procedure to determine a winner in each category for the judges of the AIAA competition. Our procedure will determine a winner in most accurate, best floater, best boomerang and best overall, objectively using specific data. This method is only useful given this is the only data that the judges are provided with.

In order to find most accurate in each team, find the least distance from the target in the straight throw and the boomerang throw. Average the two numbers, and the lowest resulting number is the winner. This is because the lowest distance from the target in both throws would mean the plane is most accurate. By averaging the distances from targets, both types of throws are taken into account.

In order to determine the best floater, take the greatest amount of time in the air for the straight path and the boomerang path for each team and average the two numbers. The greatest resulting number is the winner. This is because the greatest amount of time in the air would mean that the plain had floated for the longest period of time.

In order to determine the winner of the best boomerang you should find the lowest distance from the target in the boomerang throw. All teams who throw below four meters are eliminated from contention. We arrived at this conclusion for boomerang because the closest flight to the target would mean the best boomerang throws. Throws must at least reach the chair, so a minimum of four meters is necessary.

In order to find the winner of the best overall team, you should take the winner of the most accurate throw. If this results in a tie, you should determine which plane acquired the highest velocity to determine the tie breaker. We came to this conclusion because the most accurate throw would be the best overall since accuracy is the most important aspect of design for a airplane. The tie-breaker being the greatest velocity is because a greater velocity would mean a more efficient flight (once accuracy is already taken into account).

The most accurate is team 1 with an average distance to target being 3.1 meters. Best floater is team 5 with an average flight time of 4.15 seconds. The best boomerang is team 6 with 2.7 meters from target. The best overall is team 1 with 3.1 meters average distance to target (there were no ties so tie breaker method of employing greatest velocity to determine most efficient flight was not necessary).

Appendix B

APPLICATION OF THE *MEA FEEDBACK AND ASSESSMENT RUBRIC* TO THE SAMPLE STUDENT TEAM WORK IN APPENDIX A

Item: Mathematical Model Complexity

GTA *MEA Rubric*

4	The procedure fully addresses the complexity of the problem.
3	The procedure moderately addresses the complexity of the problem and/or contains embedded errors.
2	The procedure only somewhat addresses the complexity of the problem and/or contains embedded errors.
1	The procedure does not address the complexity of the problem and/or contains significant errors.
0	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 direct user does not constitute turning in a product.

Expert & GTA Sub-Rubrics

Expert and GTA Sub-Rubric

		Yes	No
Most Accurate	Model Present	**1**	0
	Definition of Most	1	**0**
	Model Aligns with Definition	1	**0**
Best Floater	Model Present	**1**	0
	Definition of Best	**1**	0
	Model Aligns with Definition	**1**	0
Best Boomerang	Model Present	**1**	0
	Definition of Best	**1**	0
	Model Aligns with Definition	**1**	0
Best Overall	Model Present	**1**	0
	Definition of Best	1	**0**
	Model Aligns with Definition	1	**0**

Translation from Expert to GTA *MEA Rubric* Item:
Points = 0 → Level 0
Points <= 4 → Level 1
Points <= 8 → Level 2
Points <= 10 → Level 3
Points > 10 → Level 4

Rationales	4	True	The procedure is supported with rationales for critical steps in the procedure.	**Expert Sub Rubric** 4 The procedure is well rationalized for each of the model steps, making it very reusable/modifiable 4 The procedure contains some rationales for some of the steps, but is missing enough of them to not make the procedure truly re-useable/modifiable **3** Does not achieve the level above.
	3	False		

Accounting for Data Types

4	True	The procedure takes into account all types of data provided to generate results OR justifies not using some of the data types provided.
3	False	

Expert Sub-Rubric

		Used/Justified Straight Data		Used/Justified Boomerang Data	
		Yes	No	Yes	No
Most Accurate	Length of Throw	1	**0**	1	**0**
	Time in Air	1	**0**	1	**0**
	Distance from Target	**1**	0	**1**	0
Best Floater	Length of Throw	1	**0**	1	**0**
	Time in Air	**1**	0	**1**	0
	Distance from Target	1	**0**	1	**0**
Best Boomerang	Length of Throw	1	**0**	**1**	0
	Time in Air	1	**0**	1	**0**
	Distance from Target	1	**0**	**1**	0
Best Overall	Length of Throw	1	**0**	1	**0**
	Time in Air	1	**0**	1	**0**
	Distance from Target	1	**0**	1	**0**

GTA Sub-Rubric

		Used/Justified Data	
		Yes	No
Straight Throw	Length of Throw	1	**0**
	Time in Air	**1**	0
	Distance from Target	**1**	0
Boomerang Throw	Length of Throw	**1**	0
	Time in Air	**1**	0
	Distance from Target	**1**	0

Translation from Expert to GTA *MEA Rubric* Item:
Marks are tallied to insure that all 6 data types are used or justified as not being used at some point throughout the procedure. If all data types are either used or justified, the team is given a Level 4 mark, otherwise they receive a level 3.

Results Presented

4	True	Results from applying the procedure to the data provided are presented in the form requested.
1	False	

Expert Sub-Rubric

	All Teams	Winning Team	None
Most Accurate – Rankings	1	**1**	0
Most Accurate – Quantitative Results	1	**1**	0
Best Floater – Rankings	1	**1**	0
Best Floater – Quantitative Results	1	**1**	0
Best Boomerang – Rankings	1	**1**	0
Best Boomerang – Quantitative Results	1	**1**	0
Best Overall – Rankings	1	**1**	0
Best Overall – Quantitative Results	1	**1**	0

Translation from Expert to GTA *MEA Rubric* Item:
Marks are tallied to insure that the top winners and their associated quantitative results are all present within the procedure. If all results are present, the team is given a level 4, otherwise they receive a level 1.

Readability

4	The procedure is easy for the client to understand and replicate. All steps in the procedure are clearly and completely articulated.
3	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.
2	Does not achieve the above level.

Expert Sub-Rubric

	Yes	Sort of	No
Most Accurate - Clear which data is being used	**0.5**	0.25	0
Best Floater - Clear which data is being used	**0.5**	0.25	0
Best Boomerang - Clear which data is being used	**0.5**	0.25	0
Best Overall - Clear which data is being used	**0.5**	0.25	0
Easy to Understand	**2**	1	0
Repeatable	**2**	1	0

Translation from Expert to GTA *MEA Rubric* Item:
Points <= 2 → Level 2
Points <= 5 → Level 3
Points > 5 → Level 4

No Extraneous Information

4	True	There is no extraneous information in the response.
3	False	

Expert Sub-Rubric

	Yes	Sort of	No
Used an outline format	2	1	**0**
Referred to Excel, MATLAB, or other software tools	2	1	**0**
Described how to calculate a basic statistic	2	1	**0**
Included other extraneous information	2	1	**0**

Translation from Expert to GTA *MEA Rubric* Item:
Points > 0 → Level 3
Points = 0 → Level 4

Reusability & Modifiability

4	The procedure not only works for the data provided but is clearly re-usable and share-able. Re-usability and share-ability 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.
3	The procedure works for the data provided and *might* be re-usable and share-able, but it is unclear whether the procedure is re-usable and share-able because assumptions about the situation and/or the types of data that the procedure can be applied to are not clear or not provided.
2	Does not achieve the above level.

Expert Sub-Rubric

	Yes	Sort of	No
Identified the correct client	**2**	1	0
Identified that the client needs a procedure	**2**	1	0
Identified that the procedure need to be designed to find winner of Most Accurate, Best Floater, Best Boomerang, and Best Overall competitions	**2**	1	0
Data consists of straight and boomerang throws	2	1	**0**
Data consists of multiple throws	2	1	**0**
Data contains Time in Air, Length of Throw, & Distance from Target	2	1	**0**

Translation from Expert to GTA *MEA Rubric* Item:
Points <= 5 → Level 2
Points <= 8 → Level 3
Points > 8 → Level 4