AC 2011-1010: EVALUATING STUDENT RESPONSES IN OPEN-ENDED PROBLEMS INVOLVING ITERATIVE SOLUTION DEVELOPMENT IN MODEL-ELICITING ACTIVITIES

Mark T Carnes, Purdue University, West Lafayette

Mark Carnes is a licensed Professional Engineer (PE) and is currently a doctoral student and a future faculty fellow in the School of Engineering Education at Purdue University. Before coming to Purdue, he spent over 30 years as an electronics designer of control and power conversion circuits. He received an MS from the University of Michigan (1982) and a BS from the University of Notre Dame (1975), both in Electrical Engineering.

Heidi A. Diefes-Dux, Purdue University, West Lafayette

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 in Purdue's First-Year Engineering Program, the gateway for all first-year students entering the College of Engineering. She is currently the Director of Teacher Professional Development for the Institute for P-12 Engineering Research and Learning (INSPIRE). Her research interests center on implementation and assessment of mathematical modeling problems.

Monica E Cardella, Purdue University, West Lafayette

Monica E. Cardella is an Assistant Professor of Engineering Education and is the Co-Director of Assessment Research for the Institute for P-12 Engineering Research and Learning (INSPIRE) at Purdue University. Dr. Cardella earned a B.Sc. in Mathematics from the University of Puget Sound and an M.S. and Ph.D. in Industrial Engineering at the University of Washington. At the University of Washington she worked with the Center for Engineering Learning and Teaching (CELT) and the LIFE Center (Learning in Informal and Formal Environments). She was a CASEE Postdoctoral Engineering Education Researcher at the Center for Design Research at Stanford before beginning her appointment at Purdue. Her research interests include: learning in informal and out-of-school time settings, pre-college engineering education, design thinking, mathematical thinking, and assessment research.

Evaluating Student Responses in Open-Ended Problems Involving Iterative Solution Development in Model Eliciting Activities

Abstract

Open-ended problems are an important part of the engineering curriculum because, when well designed, they closely resemble problem-solving situations students will encounter as professional engineers. However, valid and reliable evaluation of student performance on open-ended problems is a challenge given that numerous reasonable responses are likely to exist for a given problem and multiple instructors and peers may be evaluating student work. In previous work, evaluation tools for open-ended problems, specifically Model-Eliciting Activities (MEAs), were rigorously developed to ensure that the evaluation tools evolved with fidelity to characteristics of high performance and with increased reliability. As part of an on-going process of tool development, this study presents an expert evaluation of student work using the Spring 2009 version of assessment tools. The *Just-in-Time Manufacturing MEA* was implemented in Spring 2009 in a large first-year engineering course. A sample of 50 teams was selected for this study. Each of the teams' three iterative solutions was rigorously scored and coded by an engineering expert using the *MEA Rubric* and *JIT MEA* specific assessment supports. The expert scores were then compared across the three iterations, showing a positive trend of improvement in student performance across some dimensions, but little change in others. These findings have implications for instruction along each dimension. These findings also provide opportunities to investigate the nature of peer and GTA feedback that may (or may not) have resulted in change on a given iteration.

I. Introduction

An ability to apply mathematics in solving problems has been identified as an essential engineering skill[1]. A critical element in solving problems is the ability to make decisions concerning what mathematical knowledge should be applied in a given context. To develop these skills, students need the opportunity to practice applying their knowledge to open-ended problems. They need to be able to recognize that these problems may have multiple legitimate approaches that will lead to many possible acceptable answers[2]. In addition to developing solutions to open-ended problems, students also need to develop the ability to effectively communicate their solutions to those who will use them. In the *Just-In-Time Manufacturing MEA (JIT MEA)* used in this study, the solution is a procedure. The students are required to communicate in writing their procedure and results of applying it clearly enough that the direct users will be able to apply the procedure themselves and get the same results (share-ability). The written solution should help the user to understand the context in which the model can be used (re-usability), as well as provide rationales that went into the design of the model (modifiability).

A significant challenge in implementing open-ended problems in a larger educational setting is the development of valid and reliable tools for use by multiple instructors and Graduate Teaching Assistants (GTAs) to evaluate the quality of students' responses along these dimensions and motivate student learning by providing formative feedback[3]. This challenge has resulted in an on-going effort to develop such tools for implementing and assessing student work on Model-Eliciting Activities (MEAs) in a large first-year engineering course at Purdue University.

Because of the range of student approaches and solutions to these problems, it was found that consistency of GTA scoring of students' work was a significant problem (Zawojewski, et al, 2008, ch. 4)[5]. Further, the potential for guiding student learning along multiple learning objectives through the use of these problems was not being realized. The researchers discovered that a generic holistic rubric, used prior to Fall 2007, was not sufficient to address these problems. It was necessary to include MEA-specific criteria that were related to the specific issues and conceptual understandings involved in a given MEA[3]. Field observations, input from GTAs, and review by the external evaluator led to the development of a set of two evaluation tools to be used together: the MEA Feedback and Assessment Rubric (*MEA Rubric*), and the Instructors' MEA Assessment/Evaluation Package (I-MAP). The *MEA Rubric* moved from a holistic to a dimensionalized evaluation tool that separately addresses the mathematical model and the dimensions of generalizability, giving guidance for the assignment of scoring levels for each dimension. The I-MAP gives additional guidance on how to apply the rubric to the specific MEA, focusing on the special features of the MEA that apply to each dimension.

In this study, to understand the quality of student work across the dimensions, an engineering expert who is familiar with MEAs and their implementation applied the *MEA Rubric* and the I-MAP to 50 samples of student work selected from an implementation of the *JIT MEA* in Spring 2009. By doing an expert application of the *MEA Rubric* and the I-MAP of a particular MEA, it will be possible to reveal student thinking on this MEA as they iteratively develop their models in response to feedback and changing data sets. This will show students' levels of understanding of subject matter (in the case of the *JIT MEA*, descriptive statistics) as well as their ability to present and justify their models. This would include their ability to not only develop a solution, but to present it to the user in a clear and understandable way, considering the various dimensions of generalizability. By highlighting areas of student difficulty, these results could be used to make recommendations for improvement to the instructional system in general by changing or refocusing student learning objectives in the areas of concern.

The expert application can also serve as an evaluation of the feedback and assessment system. It can help identify areas of the *MEA Rubric*, I-MAP, and GTA training that may need modification or clarification to improve their effectiveness in enabling GTAs to accurately evaluate students' work.

Research Questions

1. What is the quality of student work within each dimension of the *MEA Rubric* as they iterate their solutions?

2. What parts of the feedback and assessment system (*MEA Rubric*, I-MAP, GTA training) need improvement to better enable instructors and GTAs to accurately evaluate student work?

3. What areas of the broader instructional system need improvement to better develop the knowledge and abilities being developed and assessed by this MEA?

II. Methods

A. Setting & Participants

In Spring 2009, the *JIT MEA* was implemented in a required first-year engineering problem solving and computer tools course with an enrollment of approximately 550 students. This was the second MEA implemented in that semester. The implementation was conducted over a four week period. The MEA was launched in the laboratory setting which was facilitated by two GTAs supported by four undergraduate assistants. Student teams of 3-4 students developed DRAFT 1 of their memo with procedure and results. This draft entered a double-blind peer review process. In preparation for the peer review, students participated in a calibration exercise in which they practiced giving feedback on one prototypical piece of student work using the *MEA Rubric*, were provided an expert's review of that student work, and reflected on what they needed to do differently to improve their ability to give a peer review. For the actual peer review, each student reviewed one other team's solution to the MEA. Each team was assigned at least 3 peer reviewers. Each student team actually received 2-4 peer reviews (6 teams had 2 reviews, 50 had 3, and 85 had 4), which they used to revise their solutions, creating DRAFT 2. A GTA individually assessed DRAFT 2 using the *MEA Rubric* and supporting I-MAP. Teams then revised their work and submitted TEAM FINAL, which was then assessed by the GTA for a final grade.

All students enrolled in the first-year engineering course in Spring 2009 were considered eligible for this study (N ~ 550). Fifty student teams were randomly selected for inclusion in this study. This study was approved by the human subjects protection program (i.e. IRB)"

B. The *Just-in-Time (JIT) Manufacturing MEA*

The *JIT MEA*, the focus of this study, requires student teams to use their knowledge of mathematics and statistics to develop a procedure (mathematical model) to rank shipping companies in order of most likely to least likely to be able to meet a company's delivery timing needs[4]. The motivation for developing the procedure is established by using a realistic context in which D. Dalton Technologies (DDT), a manufacturer of advanced piezoceramics and custom-made ultrasonic transducers, is unsatisfied with their current shipping service. DDT operates in a JIT manufacturing mode and requires a shipping service to move materials between two subsidiary companies. For DRAFT 1, student teams are required to establish a procedure to rank a number of alternative shipping companies using a small subset of a larger historical data set. Students are provided with data for 8 shipping companies in terms of number of minutes late a shipment arrived at its destination (Table 1 shows a subset of the data). Students are instructed to address ways to break ties in company rankings. For DRAFT 2 and TEAM FINAL, teams revisit their procedure (using GTA and peer feedback) and work with larger historical data sets. A high quality solution to this MEA would look past measures of central tendency and variation to look at the actual distribution of the data. Attention would be drawn to the frequency of values, minimum and maximum values, probabilities of values within certain ranges, etc.

The JIT MEA has been the subject of ongoing research into the development of MEAs in general, as well as the iterative development of tools for assessment of student work[3-6].

Table 1. Number of minutes late for shipping runs from Noblesville, In. to Delphi, In. (subset of sample data set)

FPS	UE	BF	SC
6	11	15	10
11	10	2	8
3	18	0	0
10	0	16	11
17	12	15	8
14	14	13	25

Note: FPS = Federal Parcel Service; UE = United Express; BF = Blue Freight; SC = ShipCorp

C. Implementation & Data Collection

The focus of this paper is the evaluation of the quality of student work on the *JIT MEA*, using the Spring 2009 version of the *MEA Rubric* and I-MAP. An earlier version of these tools was implemented in 2007 and 2008. A process of continuous improvement by implementation and iteration was envisioned. Evaluation of the results by Diefes-Dux et al.[3] led to changes to both tools to better distinguish the four primary dimensions that comprise the *MEA Rubric* used in the assessment of Spring 2009 student work:

- <u>Mathematical Model</u>: The mathematical model adequately addresses the complexity of the problem.
- <u>Share-ability</u>: The direct user/client can apply the procedure and replicate results.
- <u>Re-usability</u>: The procedure can be used by the direct user in new but similar situations.
- <u>Modifiability</u>: The procedure can be modified easily by the direct user for use in different situations.

The new version of the *MEA Rubric* was implemented for the first time in Spring 2009. The implementation occurred over a three week period as follows:

1. **FIRST MEMO**: The students individually read the company profile and the memo from Devon Dalton giving them the assignment. Also included is a set of sample data to be used by the students in developing their model and presenting results for DRAFT 1.
2. **INDIVIDUAL QUESTIONS**: Students answer individual questions about who is the direct user, what does the user need, and what are some issues that need to be considered.
3. **DRAFT 1**: Student teams come to consensus on answers to the individual questions, work as a team to develop a proposed solution to the problem, and then create DRAFT 1, a memo responding to Devon Dalton's request:

   > "In a memo to my attention, please include your team's procedure and the rank order of the shipping companies generated by applying your procedure to the sampling of data. Be sure to include additional quantitative results as appropriate to demonstrate the functioning of your

procedure. Please be sure to include your team's reasoning for each step, heuristic (i.e. rule), or consideration in your team's procedure."

4. **CALIBRATION:** Students use the *MEA Rubric* to evaluate a sample piece of student work on this MEA, and then compare their feedback to that of an expert in a reflection.
5. **PEER FEEDBACK:** A double-blind peer review[6] is conducted so that each student has an opportunity to review one other team's work, and each team is reviewed by up to 4 individuals on other teams.
6. **DRAFT 2:** The teams receive their peer feedback as well as a second memo from Devon Dalton, asking them to test their procedure on a larger data set and revise it as needed. He also reminds them of the importance of re-usability and modifiability. They are then to create DRAFT 2, a second memo to Devon Dalton.
7. **REFLECTION ON PEER FEEDBACK:** Teams evaluate the peer feedback they receive for its usefulness and then reflect on their own experience of giving feedback.
8. **GTA FEEDBACK:** The teams receive feedback from their GTA, who scores their DRAFT 2 based on the *MEA Rubric* and the I-MAP, and provides written feedback to guide them to improve their solutions.
9. **RESPONSE TO GTA FEEDBACK:** The teams respond to the GTA feedback in a guided reflection.
10. **TEAM FINAL:** The teams receive a third memo from Devon Dalton, providing a new data set and asking for a final version of their solution. They submit TEAM FINAL for grading.
11. **GRADING:** The GTA scores and provides feedback on each team's TEAM FINAL.

D. Data Analysis

Based on this implementation of the *JIT MEA* in Spring 2009, Carnes et al.[6] analyzed this same sample of student responses independently of the *MEA Rubric* and I-MAP with a focus specifically on the development of the mathematical model. Specifically, the progression of development in student team responses was gauged by the particular statistical measures that were used and the changes in those measures at each iteration. However, the quality and generalizability of students' models was not assessed. This requires an expert application of the *MEA Rubric* and *JIT MEA* I-MAP.

To address this need, the goal of the present study is to rigorously analyze the quality of the student work on the same sample of student responses used in Carnes[6]. The sample consisted of the work of 50 student teams selected from across the first year engineering course to be evenly distributed across the sections and the GTAs involved. All three iterations of each team were then rigorously scored by a single engineering expert according to the *MEA Rubric* and the *JIT MEA* I-MAP. To minimize the effect of awareness of a single team's progression across the iterations, the student responses were analyzed by sets, that is, all of the DRAFT 2 responses were analyzed first. After a lapse of time (2-3 weeks), the TEAM FINAL responses were analyzed in the same way. Then, after a similar lapse of time, the DRAFT 1 responses were analyzed. The first step of the analysis process was to separate each student response memo into its component parts: Re-statement of problem, Assumptions, Overarching description, Procedure, Rationales, and Results. Having each response organized in this way facilitated the application of the *MEA Rubric* and I-MAP criteria in a consistent manner. A single table containing the parts of the memo, the scoring criteria and the scores for each of the 50 memos made it possible to regularly check for consistency of application across the student responses. A separate table using the same format was created for each of the three iterations. Once all of the scoring was completed, the scores were then tabulated in two ways. First, the aggregate scores were tabulated to determine the overall progression on each dimension and sub-dimension; and then the number of teams changing between iterations, either positive or negative, was compared.

III. Results & Discussion

The purpose of this paper is to evaluate the quality of student work on the *JIT MEA*, with respect to the four dimensions of the *MEA Rubric* over three iterations. Based on this evaluation, recommendations will be made to improve the feedback and assessment system and the instructional system more generally.

A. What is the quality of student work within each dimension of the MEA Rubric as they iterate their solutions?

The *MEA Rubric in Four Dimensions* (Appendix A) focuses on the dimensions of Mathematical Model and Generalizability, specifically Re-Usability, Modifiability, and Share-ability. To address multiple aspects of these dimensions, Mathematical Model is subdivided into Mathematical Model Complexity and Data Usage, and Share-ability is sub-divided into Results, Audience Readability, and Extraneous Information. The result of this sub-division is that there are actually seven (sub)dimensions that are scored. The *MEA Rubric* provides generic guidelines to assigning quality Levels (scores) to each of the dimensions, while the I-MAP (Appendix B) provides guidance specific to the *JIT MEA.* The quality Levels possible vary from one (sub)dimension to another, but the values of the Levels are the same. Specifically, a quality Level 4 corresponds to a letter grade of "A", Level 3 = "B", Level 2 = "C", and Level 1 = "D".

Mathematical Model

<u>Mathematical Model Complexity</u>. In Carnes, et al.[6], the Mathematical Model Complexity was assessed based on the progression of the number of statistical measures used. For example, a team that began by using Mean alone to rank the shipping companies in DRAFT 1 was judged to have made progress if they used both Mean and Standard Deviation in DRAFT 2. Similarly, the addition of a measure of distribution, such as range, percentage of on-time deliveries, etc., was seen as a sign of progress towards a higher quality solution. For a more rigorous analysis, this study applies the definitions and criteria prescribed by the *MEA Rubric* and I-MAP to differentiate student solutions into four distinct levels of quality. The use of the *MEA Rubric* alone is not sufficient to make the necessary distinctions. Guidance provided by the I-MAP (Appendix B) helps to remove much of the subjective judgment that the grader (whether the expert for this analysis or the GTA during implementation) would need to employ if using the

MEA Rubric alone. While grading the three responses of each of the 50 teams in the sample, the expert noticed that additional guidance was needed for consistency in evaluating what constituted “accounts for how the data is distributed” at a quality Level 3. These clarifications are as follows:

> **Level 3**: To reach this level, some attempt to go beyond MEAN and STDEV must be evident. Examples include:
> 1. When MEAN and/or STDEV are used, an additional measure is included to address distribution, such as IQR, Range, count of 't=0', or outliers.
> 2. Point systems which account directly for distribution.
> 3. Count of 't=0' as a primary measure.
> 4. Sum of Squares as a primary measure.
>
> Note: Using a particular measure as a tie-breaker alone is not sufficient, since in many procedures, tiebreakers rarely, if ever, come into play.

For the sample of 50 student responses, there is a clear upward progression for student scores over the three iterations: DRAFT 1, DRAFT 2, and TEAM FINAL. As shown in Table 2, while 24 teams began at Level 1, only 8 teams were still at that level by TEAM FINAL. Only 6 began at Level 3 and 4 combined, while 17 ended there. The mean Level score showed a positive increase as well. Paired t-tests between the iterations showed statistical significance between all levels. The improvement from DRAFT 1 to DRAFT 2, while smaller, was significant at a level of $p=0.018$. Greater improvement was seen going from DRAFT 2 to TEAM FINAL ($p=0.008$).

Table 2. Student team scores for Mathematical Model Complexity on the three iterations of the *JIT MEA* (*n* = 50).

	Score Frequency		
Score	**DRAFT 1**	**DRAFT 2**	**TEAM FINAL**
4	2	2	7
3	4	10	10
2	20	21	25
1	24	17	8
Mean	**1.68**	**1.94**	**2.32**

Looking at the direction of change in Table 3, over half of the teams showed no change in their Mathematical Model Complexity score over the course of the MEA. But for those who did change, it was overwhelmingly positive. Twenty-two (22) teams improved their score from DRAFT 1 to TEAM FINAL, while only one team actually declined. A closer look at this one team revealed that this team had a reasonable point system in their first two drafts (Level 3), but switched to a Mean and Standard Deviation only method (Level 2) for TEAM FINAL, based largely on GTA feedback they received.

The primary differentiator between Level 2 and Level 3 is the recognition that solutions must go beyond mean and standard deviation to address the distribution of the data in some way. Of the 50 teams evaluated, only 17 (34%) were able to do this, leaving 34 teams (66%) who did not make this connection by TEAM FINAL.

Table 3. Direction of score change for Mathematical Model Complexity between the three iterations of the *JIT MEA* (*n* = 50).

	Frequency of Score Change		
Direction of Change	**DRAFT 1 → DRAFT 2**	**DRAFT 2 → TEAM FINAL**	**Overall: DRAFT 1 → TEAM FINAL**
Positive	11	18	22
No Change	36	29	27
Negative	3	3	1

<u>Data Types</u>. Student teams' choice to use or not use data is assessed within the Mathematical Model Complexity dimension. The *MEA Rubric* makes this a binary choice (Appendix A), either all data types are used, or they are not. When they are not, justifications are necessary. Since this MEA only has one data type (late delivery times for the shipping companies), the I-MAP focuses on the arbitrary or unjustified removal of some of the data such as removal of outliers or even some of the companies from consideration. Such removals, without reasonable justification, resulted in dropping to a quality Level 3 for this sub-dimension. Results are shown in Table 4. Since there were only 6 data points for each company in the DRAFT 1 data set, several teams chose to remove the worst delivery performances to make it "fair." There was less of this in DRAFT 2, perhaps because there were so many data points in the second data set that the student teams did not think it as necessary to remove them.

Table 4. Student team scores for Data Type on the three iterations of the *JIT MEA* (*n* = 50).

	Score Frequency		
Score	**DRAFT 1**	**DRAFT 2**	**TEAM FINAL**
4	39	46	41
3	11	4	9
Mean	**3.78**	**3.92**	**3.82**

Table 5 shows the direction of score change between iterations. The reason for the reversal going to TEAM FINAL is unclear, but it may have been an effort to "improve" their models by removing outliers. It is not clear from the *MEA Rubric* or the I-MAP what a "reasonable" justification for removal of data would be. The only cases in which the expert accepted

Table 5. Direction of score change for Data Type between the three iterations of the *JIT MEA* (*n* = 50).

	Frequency Of Score Change		
Direction Of Change	**DRAFT 1 → DRAFT 2**	**DRAFT 2 → TEAM FINAL**	**Overall: DRAFT 1 → TEAM FINAL**
Positive	8	3	6
No Change	41	39	40
Negative	1	8	4

rationales as reasonable occurred in DRAFT 1. One student team chose to remove a company completely from consideration because of one excessively late time. They reasoned that for a small sample of 6, if one delivery is excessively late, there is a likelihood that this will occur again. A second team eliminated companies from consideration if they delivered on time ('t=0') less than 5% of the time, or had deliveries that were over 50 minutes late more than 5% of the time. Their rationale was that "These eliminations remove companies that are never on time or are often extremely late. Both situations would cause a glitch in production." They maintained this idea in DRAFT 2, but removed it in TEAM FINAL, in which they ranked all of the companies without eliminating any of them.

Re-Usability

The *MEA Rubric* definition of Re-usability is:

> "***Re-usability*** means that the procedure can be used by the direct user in new but similar situations. A re-usable procedure: (1) identifies who the direct user is and what the direct user needs in terms of the deliverable, criteria for success, and constraints, (2) provides an overarching description of the procedure, and (3) clarifies assumptions and limitations concerning the use of procedure."

To implement this definition, the I-MAP uses a quantitative scoring system assigning 0-2 (0=No, 1= Sort of, 2=Yes) points for each of the 6 items identified in the definition above. These are:

1. Identification of the Direct User
2. Identification of the deliverable as a procedure
3. Criteria for success
4. Constraints
5. Overarching Description
6. Assumptions and Limitations

The sum of these scores then determines the Level score (see Appendix B) for this dimension.

Table 6. Student team scores for Re-usability on the three iterations of the *JIT MEA* (*n* = 50).

	Score Frequency		
Score	**DRAFT 1**	**DRAFT 2**	**TEAM FINAL**
4	16	12	30
3	27	36	18
2	7	2	2
Mean	**3.18**	**3.20**	**3.56**

As shown in Table 6, there was little change in the Re-usability scores from DRAFT 1 to DRAFT 2. Table 7 shows that this consists of about as many improvers as decliners. This could indicate that peer review does not seem to contribute significantly to improvement. Much more significant improvement occurred between DRAFT 2 and TEAM FINAL, possibly due to more focused feedback from the GTAs on this dimensions' items. The largest improvement was in the quality of the overarching descriptions, followed by explicit identification of the direct user.

Both of these items were emphasized in the GTA feedback. As a result, 60% of the student teams are including the appropriate information in their memo by TEAM FINAL.

Table 7. Direction of score change for Re-usability between the three iterations of the *JIT MEA* (*n* = 50).

Direction Of Change	Frequency Of Score Change DRAFT 1 → DRAFT 2	DRAFT 2 → TEAM FINAL	Overall: DRAFT 1 → TEAM FINAL
Positive	10	21	22
No Change	31	25	22
Negative	9	4	6

Modifiability

The *MEA Rubric* definition of Modifiability is:

> **"*Modifiability*** means that the procedure can be modified easily by the direct user for use in different situations. A modifiable procedure (1) contains acceptable rationales for critical steps in the procedure and (2) clearly states assumptions associated with individual procedural steps."

As shown in Table 8, more than half of the sample (28 of 50) received the top score on this dimension in DRAFT 1, which required that they have "acceptable rationales" for the steps of their procedure. For the 22 who did not, little progress is evident for the Modifiability scores. Table 9 shows a small negative trend after peer review, and a slight positive trend after GTA feedback.

Table 8. Student team scores for Modifiability on the three iterations of the *JIT MEA* (*n* = 50).

Score	Score Frequency DRAFT 1	DRAFT 2	TEAM FINAL
4	28	28	33
3.5	4	3	2
3	18	19	15
Mean	**3.60**	**3.59**	**3.68**

Table 9. Direction of score change for Modifiability between the three iterations of the *JIT MEA* (*n* = 50).

Direction Of Change	Frequency Of Score Change DRAFT 1 → DRAFT 2	DRAFT 2 → TEAM FINAL	Overall: DRAFT 1 → TEAM FINAL
Positive	10	10	13
No Change	28	34	28
Negative	12	6	9

To score this dimension, the expert looked for a clear and understandable attempt to explain why particular measures, calculations, or weighting factors were used. Teams should try to explain what these measures tell the user. When developing intermediate ranking or weighting methods, these must be justified.

Share-ability

The *MEA Rubric* definition of Share-ability is:

> ***Share-ability*** means that the direct user can apply the procedure and replicate results.

To score this dimension, it is evaluated in three parts: the presentation of results, the clarity and ease of use of the procedure, and the presence of extraneous information.

Share-Ability: Results. To satisfy this requirement, the student teams must apply their procedure *as written* to the data provided and include their results in the memo. All of the results from applying the procedure to the data provided should be presented in the form requested. The task was to provide a procedure for ranking the shipping companies, so the results reported must include a final ranking with quantitative results. Students frequently fail to present results as requested resulting in low scores on this sub-dimension. Many will pick a single "winner" instead of showing a complete ranking of all companies. As shown in Table 10, there was a decrease in scores move between DRAFT 1 and DRAFT 2. The DRAFT 2 instructions presented a new data set that the teams were to use. For five of the teams who moved to a lower level, it was due to ignoring the new data set and not presenting results for this data set. If not for this, the number of teams with an increase in score would have equaled the number with a decrease in score. There was a general improvement in the presentation of rankings and quantitative data, but little change in units and significant figures.

Table 10. Student team scores for Share-ability: Results on the three iterations of the *JIT MEA* (*n* = 50).

	Score Frequency		
Score	**DRAFT 1**	**DRAFT 2**	**TEAM FINAL**
4	5	5	10
2	36	28	35
1	9	17	5
Mean	**2.02**	**1.86**	**2.30**

Table 11. Direction of score change for Share-ability: Results between the three iterations of the *JIT MEA* (*n* = 50).

	Frequency of Score Change		
Direction Of Change	**DRAFT 1 → DRAFT 2**	**DRAFT 2 → TEAM FINAL**	**Overall: DRAFT 1 → TEAM FINAL**
Positive	7	20	17
No Change	29	25	25
Negative	14	5	8

The improvement in scores (Table 11) between DRAFT 2 and TEAM FINAL can be largely attributed to the correct use of units and significant figures. Neither of these had been mentioned in the instructions for DRAFT 1 and DRAFT 2. The TEAM FINAL instructions stated for the first time that "**Quantitative results should be presented with appropriate units and significant figures.**" With it bolded and underlined, several teams seem to have taken it seriously and improved their scores in these areas.

Share-Ability: Apply/Replicate. The I-MAP text that applies to this sub-dimension describes it as follows:

> "Share-ability means that the direct user can apply the procedure and replicate results. A high quality product (i.e., model communicated to the direct user) will clearly, efficiently and completely articulate the steps of the procedure. The description will be clear and easy to follow; it must enable the results of the test case to be reproduced."

The best way for the grader to evaluate this dimension is to follow the procedure as written and to verify that results can be obtained that either match the results presented in the student memo, or show errors in either the procedure or the student calculations. Any ambiguities found in the procedure steps can be fed back to the students for clarification in the next iteration. Table 12 shows that scores on this sub-dimension actually declined after peer review, but then recovered after GTA feedback, finishing with a small net improvement. Nearly half of the teams had no net change in their scores on this dimension over the course of the MEA (Table 13). For those who did change, it is likely that the clarity of their procedures is weakest at DRAFT 2, as this is their first attempt at developing a procedure to handle the larger data set. DRAFT 1 texts are typically shorter and more direct as they are working with only the small data set and the mathematics they are applying is simpler to articulate to the audience.

Table 12. Student team scores for Share-ability: Apply/Replicate on the three iterations of the *JIT MEA* (*n* = 50).

	Score Frequency		
Score	**DRAFT 1**	**DRAFT 2**	**TEAM FINAL**
4	32	21	34
3	7	20	10
2	11	9	6
Mean	**3.42**	**3.24**	**3.56**

Table 13. Direction of score change for Share-ability: Apply/Replicate between the three iterations of the *JIT MEA* (*n* = 50).

	Frequency of Score Change		
Direction of Change	**DRAFT 1 → DRAFT 2**	**DRAFT 2 → TEAM FINAL**	**Overall: DRAFT 1 → TEAM FINAL**
Positive	9	19	16
No Change	24	25	24
Negative	17	6	10

Share-Ability: Extraneous Information. The mathematical model should be free of distracting and unnecessary text. This includes such things as unnecessary detail on how to calculate standard statistical measures, reference to details of computer tools used, or addressing issues outside the scope of the problem.

There is no evident progression on this dimension. As shown in Table 14 and Table 15, after peer review the scores declined, but then after GTA review they improved back to the same level as in DRAFT 1. The net effect was no overall change. Thirty teams (60%) did not change the quality of their work along this dimension from start to finish. The remaining twenty are evenly split between improving and declining.

Table 14. Student team scores for Share-ability: Extraneous Information on the three iterations of the *JIT MEA* (*n* = 50).

	Score Frequency		
Score	**DRAFT 1**	**DRAFT 2**	**TEAM FINAL**
4	22	13	22
3	28	37	28
Mean	**3.44**	**3.26**	**3.44**

Table 15. Direction of score change for Share-ability: Extraneous Information between the three iterations of the *JIT MEA* (*n* = 50).

	Frequency of Score Change		
Direction of Change	**DRAFT 1 → DRAFT 2**	**DRAFT 2 → TEAM FINAL**	**Overall: DRAFT 1 → TEAM FINAL**
Positive	2	14	10
No Change	37	31	30
Negative	11	5	10

Final Score

In the first-year engineering course, the Final Score on the MEA is set to the lowest score in any of the seven scored (sub)dimensions. The philosophy being that a student team's work is only as good as the weakest element. A good model alone is not sufficient if someone else does not know when or how to apply it with equal success or there is inadequate demonstration that it works. Or a clear narration of the model and presentation of results is inadequate if the model does not address the complexity of the problem.

The fact that the Share-ability: Results sub-dimension has no Level 3 and requires perfect performance to score above Level 2, makes it extremely difficult to achieve a Final Score above Level 2, and in fact very few student teams were able to do it. To achieve Level 4 on this dimension, teams needed to show rankings of all companies, along with quantitative results with correct units and significant figures. As such, there was no improvement between DRAFT 1 and DRAFT 2, but there was significant gain between DRAFT 2 and TEAM FINAL, where GTA feedback may have helped some teams fix enough of their problems to achieve a higher score.

By TEAM FINAL, of the seventeen teams who had models at Level 3 or Level 4, twelve of them dropped back to Level 2 because of this dimension. Three presented inadequate rankings; five gave insufficient quantitative or interim results; and four failed to show appropriate units or included too many significant figures. Of the 22 who improved, 15 were due to improvements in the Share-Ability: Results sub-dimension, 6 were due to improved Mathematical Model Complexity, and 1 was due to improvement in Re-Usability.

Table 16. Final student team scores on the three iterations of the *JIT MEA* (*n* = 50).

	Score Frequency		
Score	**DRAFT 1**	**DRAFT 2**	**TEAM FINAL**
4	0	2	2
3	1	3	2
2	22	18	35
1	27	29	11
Mean	**1.48**	**1.48**	**1.90**

Table 17. Direction of change of Final Score between the three iterations of the *JIT MEA* (*n* = 50).

	Frequency of Score Change		
Direction of Change	**DRAFT 1 → DRAFT 2**	**DRAFT 2 → TEAM FINAL**	**Overall: DRAFT 1 → TEAM FINAL**
Positive	9	22	21
No Change	31	25	25
Negative	10	3	4

B. What parts of the feedback and assessment system (*MEA Rubric*, I-MAP, TA training) need improvement to better enable instructors and GTAs to accurately evaluate student work?

Table 18. Mean scores on each of the seven *MEA Rubric* (sub)dimensions for *JIT MEA* (*n*=50).

Dimension	**DRAFT 1 → DRAFT 2**	**DRAFT 2 → TEAM FINAL**	**Overall DRAFT 1 → TEAM FINAL**
Model Complexity	1.68	1.94	2.32
Data Types	3.78	3.92	3.82
Re-usability	3.18	3.20	3.56
Modifiability	3.60	3.59	3.68
Share-ability: Results	2.02	1.86	2.30
Share-ability: Apply	3.42	3.24	3.56
Share-ability: Extraneous	3.44	3.26	3.44
Final Score	1.48	1.48	1.90

Table 18 summarizes the scores for all seven of the (sub)dimensions scored on the *JIT MEA*. For all *MEA Rubric* (sub)dimensions except for Data types and Modifiability, positive gains in team scores were evident between DRAFT 2 and TEAM FINAL. This has implications for both tool development and instructional approaches. As described in Diefes-Dux et al.[3], this study is part of an on-going process to develop the *MEA Rubric* and I-MAPs for MEA implementation. By examining the results of this implementation, several areas of possible improvement have been identified; particularly the need to explain further what constitutes "reasonable" responses for several of the (sub)dimensions.

Mathematical Model

Mathematical Model Complexity. In their training, the GTAs were instructed to interpret students' work and guide them towards a mathematical approach that addresses the complexity of the problem, rather than tell the students what math to use. Since the development of the procedure was the students' primary goal, it would be expected that there should be some improvement in the procedures through the process of feedback and iteration. Between DRAFT 1 and DRAFT 2, the peer calibration exercises and the giving and receiving of peer feedback seems to have resulted in positive movement for some teams. The GTA feedback received between DRAFT 2 and TEAM FINAL seems to have resulted in even more teams making improvements. However, only 20% of the teams made positive improvements at addressing the complexity of the problem from DRAFT 1 to TEAM FINAL. Better feedback strategies are needed for GTAs to use to get students past looking at only mean and standard deviation.

Data Types**.** The GTAs need some guidance for identifying what a "reasonable" justification for removal of data would be. Then, the GTAs also need better feedback strategies that would enable them to discuss the appropriateness of getting rid of shipping companies outright or eliminating "outliers." Sample student work and suggested feedback strategies could be added to training and the I-MAP about this.

Re-usability

Improvements to the peer calibration activity are needed so that peers are better able to review this dimension. This is the only dimension where students are told explicitly in the *MEA Rubric* what is expected. They should be able to tell whether the appropriate items are present or not in another team's work. Students may need additional clarification of the terms used in the *MEA Rubric*, such as criteria for success, constraints, and "overarching."

Modifiability

GTAs and students alike find this one of the more difficult dimensions to address[3,7]. In scoring, it is easy to tell when modifiability is completely not addressed (i.e. there are no rationales present in the procedure), but it is not clear how a GTA could prompt students to improve their model for modifiability if the presence and quality of the rationales are mixed. A common confusion is over the meaning of the phrase "acceptable rationales." Future versions of the I-MAP should include some examples of acceptable rationales. GTAs also need better feedback strategies for recognizing when rationales are needed and how to prompt the students for any or better rationales. Further investigation is needed into ways of presenting Modifiability to enhance the GTAs' understanding of this dimension and their ability to provide better feedback on this dimension.

Share-ability

Share-ability: Results. The *MEA Rubric* was intentionally designed to be harsh on the aspect of presentation of results because students have had a history of not reading directions and presenting any results at all. Now that we seem to be past that, reconsideration of how minor presentation problems impact the overall score may be necessary.

Share-ability: Apply/Replicate**.** The best way to assess this sub-dimension is for the GTA to actually implement the procedure and compare the results obtained to the students' results, as the I-MAP requires. This will allow the GTA to easily assess both the accuracy of the model and its ease of use. This is feasible for the *JIT MEA*, since most student solutions are relatively easy to replicate, but it may not be true for other MEAs which have more complex solutions. This sub-dimension can become difficult to assess if the GTA does not actually try to apply the procedure. In a large class with many student solutions to grade, GTAs may not have sufficient time to replicate each procedure, so assessment of this sub-dimension can become somewhat subjective.

Share-ability: Extraneous Information. Students often have difficulty determining how much detail is too much[7]. Some feedback may ask for more explanation and justification, but then this becomes a negative for this sub-dimension when the students explain the wrong things. For instance, the GTA may ask a team to explain their use of mean; the student team's reaction is to describe how to take an average rather than why average is used in their model. The description of how to take a mean becomes extraneous information.

C. What areas of the broader instructional system need improvement to better develop the knowledge and abilities being developed and assessed by this MEA?

Mathematical Model

Mathematical Model Complexity. The amount of time devoted to and the emphasis placed on definitions when teaching descriptive statistics may not be sufficient to develop sufficient statistical conceptual understanding. It may be necessary to revisit how we are teaching descriptive statistics, to think about how we talk about using descriptive measures, particularly distribution of data, in practice to make decisions.

Data Types**.** Dealing with outliers is an important topic in statistical conceptual understanding. We need to revisit our instruction on outliers and when data can and cannot be justifiably removed from a data set.

Re-usability

Given the low impact of peer review between DRAFT 1 and DRAFT 2 on the scores for this dimension, the need for documenting the team's problem formulation step needs to be more strongly emphasized to the students in general. They need to better understand that the user needs to know what this procedure is and when to use it.

Modifiability

Students' rationales are tied to their existing procedure. Over 30% of the students are not writing adequate rationales. Some instruction needs to be developed to help them develop good rationale statements. Often when they are nudged to explain something better, the GTA is after a better rationale. The student often adds something that is really extraneous information – so the students do not know how to respond to the feedback they are currently receiving on this dimension.

Share-ability

Share-ability: Results**.** Consistent reminders about the use of proper units and significant figures on numerical results in the MEA instructions for all iterations would be helpful for students.

Share-ability: Apply/Replicate. It needs to be emphasized to the students that the procedure must be written in such a way that the user can apply it without asking the students for clarification. Students often fail to recognize that the procedure must stand alone. On the assessment side, this sub-dimension may suffer from an amount of subjectivity, even within a single instructor or expert, if they do not take the time to test the procedure. With a recent reduction in student teams to evaluate per GTA, there may be better assessment along this sub-dimension.

Share-ability: Extraneous Information. Student difficulties with this sub-dimension may relate back to their difficulties with Modifiability, which may result in calls for additional explanation of rationales for procedural steps. Students may be confused as to what explanations they are expected to give.

IV. Conclusion

The *MEA Rubric* serves as a framework for evaluating the quality of student work on authentic mathematical modeling problems. It emphasizes the mathematics of the model and various aspects of communication of the model. Through an expert application of the *MEA Rubric* to a set of student teams' iterative solutions to the problem, we can see how students' solutions are progressing along these (sub)dimensions as a result of peer review, GTA feedback and team revisions. This analysis exposes weaknesses in students' understanding of descriptive statistics concepts. It also exposes weaknesses in students' ability to rationalize their models (Modifiability) and realize the importance of articulating their understanding of the problem (Re-usability). These weaknesses have implications for more traditional instruction on statistics and problem solving for all instructors who teach these subjects.
For this research effort, a valuable next step would be to correlate the results of this expert application of the *MEA Rubric* and I-MAP with the GTA scores and feedback on the same student work sample. Such a comparison could identify the types of feedback that seem to be most effective in improving student performance.

This analysis also shows peer review has low impact on the quality of student work. As a result of peer review occurring between DRAFT 1 and DRAFT 2, positive gains were seen only in the Mathematical Model; little change occurred along the Modifiability and Re-usability dimensions, and negative changes were seen in all three sub-dimensions of Share-ability. Additional study is needed to determine why this should be. One possibility is that since these dimensions are new to the students and not well understood by them, they are not adequately equipped to provide formative feedback to others. But there may be other learning value in participating in peer review[7]. Peer review, its impact and learning value, needs more exploration.

Acknowledgements

This work was made possible by grants from the National Science Foundation (DUE 0717508 and EEC 0835873). 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. The authors would also like to thank Amanda Fry and Amani Salim for their feedback on this work.

Bibliography

1. ABET Engineering Accreditation Commission. "Criteria for Accrediting Programs in Engineering". 2010, ABET, Inc.: Baltimore, MD, Retrieved March 22, 2010 from http://www.abet.org/forms.shtml#For_Engineering_Programs_Only
2. Schoenfeld, A. H. "Learning to Think Mathematically: Problem Solving, Metacognition, and sense-making in mathematics." In Grouws, D. (ed.), *Handbook for Research on Mathematics Teaching and Learning*, 1992, New York: MacMillan, pp. 334-370.
3. Diefes-Dux, H.A., Zawojewski, J.S., and Hjalmarson, M.A. "Using Educational Research in the Design of Evaluation Tools for Open-Ended Problems." *Int. J. Engng Ed.* Vol. 26, No. 4, pp.807-819, 2010
4. Hjalmarson M. "Engineering Students Designing a Statistical Procedure." *Journal of Mathematical Behavior*, **26**(2), 2007, pp.178-188.
5. Zawojewski, J.S., Diefes-Dux, H.A., Bowman, K.J., *Models and Modeling in Engineering Education: Designing Experiences for All Students*, 2008, Sense Publishers, Rotterdam, The Netherlands.
6. Carnes, M.T., Diefes-Dux, H.A., Cardella, M., "Progression of Student Solutions over the Course of A Model-Eliciting Activity (MEA)". *40th ASEE/IEEE Frontiers in Education Conference, 2010.*
7. Fry, A., Diefes-Dux, H.A., "Student Responses to and Perceptions of Feedback Received on a Series of Model-Eliciting Activities: A Complete Case Study". *ASEE Annual Conference, 2011.*
8. R. Lesh, M. Hoover, B. Hole, A. Kelly, A., and T. Post, Principles for developing thought-revealing activities for students and teachers. in A.E. Kelly and R.A. Lesh (Eds.), *Handbook of Research Design in Mathematics and Science Education.* Lawrence Erlbaum, Mahwah, New Jersey, (2000), pp. 591-645.
9. E. A. Kelly, R. A. Lesh, and J. Y. Baek (Eds.), *Handbook of design research methods in education: Innovations in science, technology, engineering and mathematics learning and teaching*. Routledge, London, (2008).
10. R. Lesh and H. M. Doerr (Eds.), *Beyond Constructivism: Models and Modeling Perspectives on Mathematics Problem Solving, Learning, and Teaching*. Lawrence Erlbaum, Mahwah, New Jersey, (2003).
11. H. A. Diefes-Dux, T. Moore, J. Zawojewski, P. K. Imbrie, and D. Follman, A Framework for Posing Open-Ended Engineering Problems: Model-Eliciting Activities. in *Frontiers in Education Conference*, Savannah, GA. (2004).
12. T. Moore and H. A. Diefes-Dux, Developing Model-Eliciting Activities for Undergraduate Students Based on Advanced Engineering Content. in *Frontiers in Education Conference*, Savannah, GA. (2004).

APPENDIX A: *MEA Rubric* with Four Dimensions

<table>
<tr><th>Dim.</th><th>Item Label</th><th>Full Item Wording</th><th colspan="2">Level</th></tr>
<tr><td rowspan="8">Mathematical Model</td><td rowspan="5">Mathematical Model Complexity</td><td>The procedure fully addresses the complexity of the problem.</td><td colspan="2">4</td></tr>
<tr><td>A procedure moderately addresses the complexity of the problem or contains embedded errors.</td><td colspan="2">3</td></tr>
<tr><td>A procedure somewhat addresses the complexity of the problem or contains embedded errors.</td><td colspan="2">2</td></tr>
<tr><td>The procedure does not address the complexity of the problem and/or contains significant errors.</td><td colspan="2">1</td></tr>
<tr><td>No progress has been made in developing a model. Nothing has been produced that even resembles a poor mathematical model. For example, simply rewriting the question or writing a "chatty" letter to the client does not constitute turning in a product.</td><td colspan="2">0</td></tr>
<tr><td rowspan="2">Data Usage</td><td rowspan="2">The procedure takes into account all types of data provided to generate results OR reasonably justifies not using some of the data types provided.</td><td>True</td><td>4</td></tr>
<tr><td>False</td><td>3</td></tr>
<tr><td colspan="4">Provide Written Feedback About the Mathematical Model Here:</td></tr>
<tr><td rowspan="6">Re-Usability</td><td colspan="4">Re-usability means that the procedure can be used by the direct user in new but similar situations. A re-usable procedure: (1) identifies who the direct user is and what the direct user needs in terms of the deliverable, criteria for success, and constraints, (2) provides an overarching description of the procedure, and (3) clarifies assumptions and limitations concerning the use of procedure.</td></tr>
<tr><td rowspan="3">Re-Usability</td><td>The procedure is clearly re-usable.</td><td colspan="2">4</td></tr>
<tr><td>The procedure might be re-usable, but it is unclear whether the procedure is re-usable because a few pieces are missing or need clarification.</td><td colspan="2">3</td></tr>
<tr><td>The procedure is not re-usable because multiple pieces are missing or need clarification.</td><td colspan="2">2</td></tr>
<tr><td colspan="4">Provide Written Feedback about Re-Usability Here:</td></tr>
</table>

<table>
<tr><td rowspan="5">Modifiability</td><td colspan="3">Modifiability means that the procedure can be modified easily by the direct user for use in different situations. A modifiable procedure (1) contains acceptable rationales for critical steps in the procedure and (2) clearly states assumptions associated with individual procedural steps.</td></tr>
<tr><td rowspan="3">Modifiability</td><td>The procedure is clearly modifiable.</td><td>4</td></tr>
<tr><td>The procedure is lacking acceptable rationales for a few critical steps in the procedure, and/or a few assumptions are missing or need clarification.</td><td>3.5</td></tr>
<tr><td>The procedure is lacking acceptable rationales for multiple critical steps in the procedure, and/or multiple assumptions are missing or need clarification.</td><td>3</td></tr>
<tr><td colspan="3">Provide Written Feedback about Modifiability Here:</td></tr>
<tr><td rowspan="12">Share-ability</td><td colspan="3">Share-ability means that the direct user can apply the procedure and replicate results.</td></tr>
<tr><td rowspan="4">Results</td><td>All of the results from applying the procedure to the data provided are presented in the form requested.</td><td>4</td></tr>
<tr><td>Results from applying the procedure to the data provided are presented, but additional results are still required, they are not presented in the form requested, or they are not consistent with the procedure.</td><td>2</td></tr>
<tr><td>No results are provided and therefore this procedure does not meet the minimum requirements requested by the direct user.</td><td>1</td></tr>
<tr><td colspan="2">List any missing results. If no results are missing, type "No results are missing."</td></tr>
<tr><td rowspan="3">Audience Readability</td><td>The procedure is easy for the client to understand and replicate. All steps in the procedure are clearly and completely articulated.</td><td>4</td></tr>
<tr><td>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.</td><td>3</td></tr>
<tr><td>Does not achieve the above level.</td><td>2</td></tr>
<tr><td rowspan="2">Extraneous Information</td><td rowspan="2">There is no extraneous information in the response.</td><td>True 4</td></tr>
<tr><td>False 3</td></tr>
<tr><td colspan="3">Provide Written Feedback About Share-ability Here:</td></tr>
</table>

APPENDIX B: Instructors' MEA Assessment/Evaluation Package (I-MAP)
(for Team Solution Components Only)

Just-In-Time Manufacturing
MEA Feedback and Assessment
Core Elements of Performance on an MEA

Mathematical Model

A mathematical model may be in the form of a procedure or explanation that accomplishes a 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 errors.

Specific to Just-In-Time Manufacturing MEA

Complexity

Looking beyond a single measure of central tendency: This particular MEA is set in a context where patterns of late arrival are important. Therefore, the data sets are designed so that the differences in the mean are insignificant. This is intended to nudge students to look beyond measures of central tendency. Therefore, more than one statistical measure is needed. Teams might use a number of measures simultaneously, or one following the other. They might also use one measure to produce an answer and another to "check" how well the answer works, leading to a possible revision. Results from statistical procedures may be aggregated in some fashion using rankings, formulas, or other methods.

In a high quality model:

- The procedure looks past measures of central tendency and variation to look at the actual distribution of the data, where attention is drawn to the frequency of values, particularly minimum and maximum values.
- Final overall ranking measure or method must be clearly described in an overarching description statement. Completes the sentence, the ranking procedure is based on… (Overlap with Re-usability – but it must be present to obtain a Level 4 Math Model)
    - This is Part B of the standard introduction :
      B. Describe what the procedure below is designed to do or find – be specific (~1- 2 sentences)

LEVEL 1 -

- The procedure described does **not account** for both the variability or distribution of these data. Students cannot move past this level if only the mean of the data is used in their procedure.
- Merely computing a series of statistical measures without a coherent procedure to use the results fall into this level.

LEVEL 2 –

- The procedure described accounts for central tendency and variability, but not the distribution, of these data. Central tendency CANNOT be ignored as it provides the basis for looking at variability.
- Mathematical detail may be lacking or missing.
- Mathematical errors might be present.
- If the solution demonstrates lack of understanding of the context of the problem, this is the highest level achievable.
- If there is an indication that the team does not understand one or more statistical measures being used, drop to the next level

LEVEL 3 –

- The procedure described accounts for both the central tendency, variability, and distribution of these data. That is the procedure includes more than the mean and/or standard deviation. The ranking procedure accounts for how the data is distributed.
- The procedure provides a viable strategy for how to break tie.
- Some mathematical detail may be lacking or missing.
- Mathematical errors might be present.
- If there is an indication that the team does not understand one or more statistical measures being used, drop to the next level

LEVEL 4 –

- Clear overarching description of what of how the ranking is being determined (Overlap with Re-usability – but it must be present to obtain a Level 4 Math Model)
- Mathematical detail should be clear from start to finish.
- Mathematical errors should be eliminated.

Accounting for Data Types

It must be determined whether the mathematical model *takes into account all types of data provided* to generate results. If any shipping companies in their entirety or parts of a shipping company's data are not used in the mathematical model, a *reasonable* justification must be provided.

LEVEL 4 – All data types are used OR *reasonable* justifications are provided.

Justifications must be provided for things like:

- Removal of any part of the time data for any company
  - Removal of "outliers"
- Dropping shipping companies

Generalizability

Generally, one would not produce a mathematical model to solve a problem for a single situation. A mathematical model is produced when a situation will arise repeatedly, with different data sets. Therefore, the model needs to be able to work for the data set provided and a variety of other data sets. That is, a useful mathematical model is adaptable to similar, but slightly different, situations. For example, a novel data set may emerge that wasn't accounted for in the original model, and thus the user would need to revise the model to accommodate the new situation.

A mathematical model that is generalizable is share-able, re-usable, and modifiable. Thus, one should strive for clarity, efficiency and simplicity in mathematical models; as such models are the ones that are more readily modified for new situations. Although the student team has been "hired" as the consultant team to construct a mathematical model, direct user needs and wants to understand what the model accomplishes, what trade-offs were involved in creating the model, and how the model works.

Re-Usability

Re-usability means that the procedure can be used by the direct user in new but similar situations.
A *re-usable* procedure:

- Identifies who the direct user is and what the direct user needs in terms of the deliverable, criteria for success, and constraints
- Provides an overarching description of the procedure
- Clarifies assumptions and limitations concerning the use of procedure.

Constrains and limitations of the use to the procedure include assumptions about the situation and the types of data to which the procedure can be applied.

Student teams should state that the procedure is for DDT's Logistic Manager and is designed to rank shipping companies in order of best to least able to meet DDT's timing needs given historical data for multiple shipping companies of time late for shipping runs between two specified locations.

Limitations that may arise based on the nature of their procedure might include assumption of the number of teams in the competition, number of team throws, the presence of a complete set of data for each team, or other.

Re-Usability Item	**JIT Manufacturing**	**Yes (2 pts)**	**Sort Of (1 pt)**	**No (0 pt)**
Identification of direct user	DDT's Logistic Manager			
Deliverable	Procedure			
Criteria for success	Rank shipping companies in order of best to least able to meet DDT's timing needs			
Constraints	Given historical data for multiple shipping companies of time late for shipping runs between two specified locations.			
Overarching Description	Should provide an overview of how the ranking is to be determined			
Assumptions and limitations concerning the *use* of procedure	Anything not covered by Constraints? May depend on procedure.			

LEVEL 4: rubric score of 9-12
LEVEL 3: rubric score of 6-8
LEVEL 2: rubric score <= 5

Modifiability

Modifiability means that the procedure can be modified easily by the direct user for use in different situations.
A modifiable procedure:

- Contains acceptable rationales for critical steps in the procedure and
- Clearly states assumptions associated with individual procedural steps.

Given this type of information, the direct user will be able to modify the model for new situations.

Critical steps that need justification / rationale:

- When teams use any statistical measures, these measures must be justified – explain what these measures tells the user.
- When developing intermediate ranking or weighting methods, these must be justified.

Share-ability

Share-ability means that the direct user can apply the procedure 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 shareable.

Results

LEVEL 4 achievement requires that the mathematical model be applied to the data provided to generate results in the form requested. Quantitative results are to be provided.

Results of applying the procedure MUST be included in the memo.
LEVEL 1 – No results or ranking of shipping companies with no quantitative results or results do not seem to be those for the data set indicated. Ensure that the student teams are presenting results for the specified data set. Multiple data sets may have been made available to the students and the analysis of only the latest may have been requested in the current memo.
LEVEL 2 – Partial rankings or quantitative results. Units may be missing or contain errors. Significant figures are to more than one decimal place.
LEVEL 4 – Both rankings and quantitative results for each shipping company are provided. Units are given and are correct. . Significant figures are to one decimal place.

Apply and Replicate Results

A high quality product (i.e., model communicated to the direct user) will clearly, efficiently and completely articulate the steps of the procedure. A high quality product may also illustrate how the model is used on a given set of data. The description will be clear and easy to follow; it must enable the results of the test case to be reproduced. At a minimum, the results from applying the procedure to the data provided must be presented in the form requested.

The direct user requires a relatively easy-to-read-and-use procedure. If this has not been delivered, the solution is not LEVEL 3 work.
If you, as a representative of the direct user, cannot replicate or generate results, the solution is not LEVEL 3 work.

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 or 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, and (4) general rambling.

LEVEL 3 – If any of the following are present::

- *Outline formatting.*
- *Mentions of computer tools*
- *Descriptions of how to compute standard statistical measures (e.g. mean, standard deviation)*
- *Reiterating, providing details, or changing the rules of the competition. For instance, discussions centered on allowing for extra throws (for instance, to break ties); this issue is outside the scope of the problem.*