Int. J. Engng Ed. Vol. 26, No. 4, pp. 807–819, 2010
Printed in Great Britain.

0949-149X/91 $3.00+0.00

Using Educational Research in the Design of Evaluation Tools for Open-Ended Problems*

HEIDI A. DIEFES-DUX[1], JUDITH S. ZAWOJEWSKI[2] and MARGRET A. HJALMARSON[3]
[1] *School of Engineering Education, Purdue University, 701 West Stadium Drive, West Lafayette, IN 47907-2045 USA. E-mail: hdiefes@purdue.edu*
[2] *Illinois Institute of Technology, MSED, 3424 S. State, South, Rm. 4007, Chicago, IL 60616-3893 USA. E-mail: zawojewski@iit.edu*
[3] *College of Education & Human Development, George Mason University, 4400 University Drive MSN 4C2, Fairfax, VA 22030-4422 USA. E-mail: mhjalmar@gmu.edu*

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 may be evaluating student work. The purpose of this paper is to present a concrete example of how educational design research, a models-and-modeling perspective from mathematics education, and multi-tiered teaching experiments are brought to bear in the design of valid and reliable evaluation tools for scoring team responses to complex problem-solving activities used in a large first-year engineering course in which teaching assistants evaluate student work. This on-going design study demonstrates how designing a package of evaluation tools (including rubrics, task-specific supports, and scorer training) based on the aforementioned educational research methods supports (1) sustained fidelity to engineering expert-identified characteristics of high performance across iterations of change to improve reliability, and (2) the implementation of planned iterations of the evaluation tools based on systematically collected data.

Keywords: design-based research; open-ended problem solving; evaluation tools

1. INTRODUCTION

OPEN-ENDED PROBLEMS are an important part of the engineering curriculum because, when well designed, they closely resemble problem-solving and design situations students will encounter as professional engineers. However, two major challenges exist in evaluating student work on such problems in engineering education classrooms. The first challenge is to identify criteria for evaluating students' solutions that reflect what would be valued in the professional engineering environment. The second challenge is to design valid and reliable tools for use by multiple instructors as they evaluate students' responses, especially when a variety of reasonable solutions may be produced whether assessing small-scale problem-solving or large-scale design tasks [1]. In order to motivate student learning, such assessment should provide formative feedback to students as well [2]. Thus, the purpose of this paper is to demonstrate how educational research methodology is used in the on-going design of evaluation packages (including rubrics, task-specific supports, and scorer training). In particular, the goal for the design of the evaluation tools is to maintain fidelity to characteristics of high performance as described by engineering experts and to obtain reliable scoring by graduate teaching assistants (TAs)—i.e., scoring sample work within one level of an expert's score 90% of the time.

In this study, the development of the evaluation tools is embedded in a larger system that involves course constraints, a large number of students, a large numbers of TAs who would be evaluating student products, and training of these TAs. Therefore, the researchers draw on three educational research perspectives. A models-and-modeling perspective [3] provides a framework for selecting problem-solving tasks that are specifically designed to simultaneously serve as sites for student assessment and for research. Design research methodology [4–6] provides guidance for planning iterations of designing evaluation tools, assessing them, and revising the tools based on information gathered in the assessment. The multi-tiered teaching experiment methodology [7] provides a framework for embracing the dynamic nature of educational research settings, where changing conditions, constraints, and perspectives of various constituencies (students, TAs, instructors, researchers) are common. Using these three

* Accepted 15 October 2009.

educational research perspectives, the iterations of designing evaluation tools produce a trail of evidence and documentation that is available for study and making subsequent design decisions.

1.1 Selection and design of open-ended problems

Open-ended problems are selected for a required first-year engineering course that emphasizes problem solving and computer tools. These problems share characteristics with those encountered in professional engineering: problems that exhibit a high level of challenge; require problem formulation; and can be solved in a number of ways [8]. Further, since the course requires team problem solving, the problems need to have enough challenge to require collaboration. Tapping the models and modeling literature [9], model-eliciting activities (MEAs) have been selected to fulfill the above purpose. In addition to the above characteristics, MEAs require teams to mathematize (e.g., quantify, organize, dimensionalize) information in an engineering context, and the solutions are mathematical models that reveal information about teams' approaches to solving the problem. MEAs are carefully designed to adhere to six design principles [3, 9–12]. The model-construction principle means that the problem requires students to create a mathematical system (i.e., model) to address the needs of a given client. The reality principle requires that the problem be based in a realistic engineering situation for which a mathematical model needs to be created. The self-assessment principle means the problem must contain information or data that can assist the team in ongoing evaluation of their progress. The model-documentation principle requires that the problem's solutions will be the teams' models. The generalizability principle requires that the model produced can be shared with others (i.e., clearly communicated to other users) and re-usable (i.e., articulates rationales and assumptions that facilitate revising the model for use in similar, although somewhat different, situations). Finally, the effective prototype or simplicity principle requires that the problem and solution to the problem provide powerful metaphors to students for interpreting future situations, and thus has clear educational value. For the first-year engineering instantiations of MEAs, the problem is conveyed to the student teams via a memo from a fictitious supervisor. The student team responses are in the form of a memo directed to the supervisor and contain their procedures for solving the problem with results from applying their procedures to a given data set.

The MEA used to illustrate the design of evaluation tools is titled *Just-in-Time Manufacturing* MEA. This MEA requires students use their knowledge of statistics to develop a procedure to rank potential shipping companies to meet the delivery needs of a client (here, Devon Dalton) [13]. Using the number of minutes late for a set of past deliveries, the students should use statistical measures beyond the mean, such as standard deviation, frequency analysis, and range to generate a procedure that quantifies the potential for arriving on time. For instance, the students have to consider consistency and reliability in order to determine whether a company is more likely to deliver on time. In particular, the *Just-in-Time* MEA requires students to develop a procedure to rank shipping companies in order of most likely to least likely able to meet a client's delivery timing needs. The motivation for developing the procedure is established by using a realistic context in which D. Dalton Technologies, a manufacturer of advanced piezoceramics and custom-made ultrasonic transducers, is unsatisfied with their current shipping service. The manufacturer operates in a just-in-time manufacturing mode and requires a shipping service to move materials between two subsidiary companies. Student teams of four are required to establish a procedure to rank a number of alternative shipping companies using a small subset of a large historical data set. Teams are provided with data for eight shipping companies in terms of number of minutes late a shipment arrived at its destination (Table 1). Students are instructed to address ways to break ties in company rankings.

A prototypical student team response to this problem is provided in Fig. 1 to highlight the characteristics of a solution that are considered during the evaluation of student work. One strength of this team's response is that the team has explicitly articulated a generalized procedure for rank-ordering the companies. A second strength is that the team has addressed the complexity of the problem; they have realized that *just* ranking on the basis of mean is inadequate as the means are quite close to each other

Table 1. Number of Minutes Late for Shipping Runs from Noblesville, IN to Delphi, IN (sample data set)

FPS	UE	BF	SC	LL	NPS	SS	HC
6	11	15	10	11	24	0	12
11	10	2	8	8	0	6	0
3	18	0	0	6	27	19	5
10	0	16	11	13	5	0	4
17	12	15	8	11	1	33	40
14	14	13	25	15	5	3	2

Note: FPS = Federal Parcel Service; UE = United Express; BF = Blue Freight; SC = ShipCorp; LL = LandLine; NPS = National Package Service; SS = Swift Star; HC = Highway Carriers

To: Devon Dalton
From: Team 5
RE: Procedure to rank shipping companies

Our task is to develop a procedure that ranks potential shipping companies from best to least able to meet the needs of DDT. Historical data for each shipping company that consists of lateness of arrival between subsidiaries is used. It is assumed that time is the most important factor in determining the best shipping company. It is also assumed that DDT has access to the program Excel and has the knowledge to use the histogram, skewdness, and standard deviation functions. There should be adequate data for each shipping company on which to base a decision. No other limitations apply.

Procedure

1. Make a chart of all the shipping times for each company considered. The time entered should be in minutes and it should be the amount of time each shipment is late.
2. Calculate the standard deviation of each company. The standard deviation is a measurement of the spread of the data. A higher standard deviation means the data is more spread out. We want to calculate the standard deviation because it shows us how close each company is to the mean. A lower standard deviation is better because it means the company is making deliveries at consistent times.
   a. To calculate the standard deviation, use Excel.
   b. Assign each company a value between one and the number of companies. The company with the lowest standard deviation is assigned a value of one, the company with second lowest standard deviation is assigned a value of two, etc. until each company is assigned a value.
   c. If two companies have the same standard deviation, assign them the same value.
   d. Multiply each company's assigned value by four. We do this because the standard deviation should be weighted higher than the other factors being used to rank the companies.
3. Calculate the mean of the times for each company. The mean is the average. We want to calculate the mean of each company's delivery time to see approximately how many minutes late each delivery is. The lower the mean, the better, because that means the company is close to being on time.
   a. To calculate the mean, add all of the data points, then divide by the number of data points.
   b. Assign each company a value between one and the number of companies. The company with the lowest mean is assigned a value of one, the company with second lowest mean is assigned a value of two, etc. until each company is assigned a value.
   c. If two companies have the same mean, assign them the same value.
   d. Multiply each company's assigned value by three. We do this because the mean should be weighted higher than the other factors being used to rank the companies, but lower than the standard deviation.
4. Create a histogram for each company. A histogram is a chart that displays the frequency of data. We want to create histograms because we want the company to have the highest frequency of data within the lowest time ranges.
   a. To create a histogram, use Excel. After creating the histogram, calculate skewdness using Excel (use descriptive statistics to get the skewdness value).
   b. Assign each company a value between one and the number of companies. The company with the highest value for skewdness is assigned a value of one, the company with second highest value for skewdness is assigned a value of two, etc. until each company is assigned a value.
   c. If two companies have the same skewdness value, assign them the same value.
   d. Multiply each company's assigned value by two. We do this because the histogram data should be weighted higher than the other factors being used to rank the companies, but lower than the standard deviation and the mean.
5. Calculate minimum shipping time for each company. We want to calculate the minimum shipping time because this is the closest the company came to delivering on time. If the company was on time the minimum would be zero, which is ideal.
   a. To get the minimum shipping time, select the lowest time value for each company.
   b. Assign each company a value between one and the number of companies. The company with the lowest minimum is assigned a value of one, the company with second lowest minimum is assigned a value of two, etc. until each company is assigned a value.
   c. If two companies have the same minimum, assign them the same value.
6. Calculate the maximum shipping time for each company. We want to calculate the maximum shipping time because this is the latest the company was from delivering on time. We want the company to have a low value for maximum shipping time because this means the company was not very late.
   a. To calculate the maximum shipping time, select the highest time value for each company.
   b. Assign each company a value between one and the number of companies. The company with the lowest maximum is assigned a value of one, the company with second lowest maximum is assigned a value of two, etc. until each company is assigned a value.
   c. If two companies have the same maximum, assign them the same value.
7. Add all of the assigned values given to each of the companies. The company with the lowest value is ranked highest, while the company with the highest value is ranked lowest.

Results (highest ranked to lowest ranked)
Federal Parcel Service, rank sum 31
Blue Freight, rs 36
Swift Star, rs 43
ShipCorp, rs 44
LandLine,rs 46
National Package Service, rs 51
United Express, rs 57
Highway Carriers, rs 61

Fig. 1. Sample Good First Draft Response (Prototypical Student Work).

and do address variability and distribution of the data. Further, the procedure the team has produced is supported with rationales for the steps and explanations for implementing the steps. In addition, the team provides results for the given set of data, which illustrates how to apply this procedure to a set of data. The team could improve this response by imagining they are someone else applying this procedure and revising difficult to interpret steps accordingly.

1.2 Research goal

A curriculum reform effort in the first-year engineering course at Purdue University has brought to the surface the problem of developing valid and reliable evaluation tools for MEAs [14]. Since 2002 over 20 different MEAs have been used in the first-year engineering course on problem solving and computer tools. This course serves 1200–1600 students each fall semester, and 300 to 400 teams produce solutions to multiple MEAs in each course offering. These team solutions are evaluated by the 15 to 20 TAs employed by the course. Initially, an assessment guide adapted from a models-and-modeling perspective [15, 16] was provided to student teams and TAs to assist them in understanding what constitutes a high quality solution. The Quality Assurance Guide (QAG) is a five-point holistic scoring rubric that focuses on the usefulness of the product for the client and has been adapted to the first-year engineering course for scoring team responses. Similarly structured versions of the QAG have been used in other settings to assess students' work [e.g., 15]. A team solution is considered to be of high quality if it meets the client's needs, has justified procedural steps, articulates underlying assumptions, and indicates awareness of limitations [16]. It was found that the TAs scored team responses inconsistently and that the QAG pays no attention to conceptual understanding specific to a given MEA. The goal of the research, therefore, has been to develop evaluation tools that would be valid (i.e., maintained fidelity to what is valued by professional engineers) and reliable (i.e., used consistently by TAs in scoring teamwork).

2. METHODS, ANALYSIS AND FINDINGS

Four stages of design of the evaluation package are described in this section. Using a design research approach [4–6], the researchers have planned for iterative cycles of producing (and revising) evaluation tools, implementing them, and gathering data to inform the next iteration. Stage I was designed to establish valid criteria by tapping the expertise of practicing engineers: the experts identified the characteristics of responses that would be valued in a professional engineering setting. Stage II was designed to build towards reliability by producing initial evaluation tools that would be usable by TAs while maintaining fidelity to the expert-identified criteria. Stages III & IV were designed to begin the iterative process of implementing, assessing, and revising the evaluation tools as they were used in full-scale implementation. What is presented here is an illustration of the design research methodology in action; the evaluation package that marks the end of Stage IV is a snapshot in time. These tools continue to be used and revised using design research methodology.

2.1 Building validity (Stage I)

The validity of the evaluation tool was established by asking a panel of four engineering experts to identify characteristics of high performance that should be used for evaluating team responses to the *Just-in-Time* MEA. While the panel was small, it was comprised of engineers representing a range of fields, types of practical and educational experience, and familiarity with MEAs so as to incorporate multiple perspectives and minimize disciplinary bias. The experts on the panel first became familiar with the *Just-in-Time* MEA by solving it prior to attending a workshop. Activities at the workshop, and after the workshop, were planned to ensure that the experts would reveal what they valued in a good response. The initial activity required the experts to share their solutions, to compare and contrast their solution with the others, and to discuss the strengths and weaknesses that were noticeable across solutions. The second activity asked experts to use an adaptation of the QAG as a starting point for scoring prototypical teamwork, along with their own professional judgment if the QAG did not capture important characteristics for high quality responses. Experts individually scored five prototypical team responses to the MEA and wrote notes on the responses indicating why they gave the score they did. Then, experts were asked to share their scores and to come to consensus on each sample solution. The researchers were free to intervene with questions that would help bring to the surface what the experts were valuing as they negotiated a score. This second activity was repeated on a second set of teamwork, based on the assumption that the experts would have revised their thinking in response to interacting about the first set of teamwork.

After the workshop, the researchers used field notes and the handwritten notes of the experts on the teamwork to produce a first draft evaluation package comprised of a statement of the important characteristics for high quality work along with a rubric that captured the general criteria the experts were apparently using. Three areas emerged: appropriateness of the mathematical model, attention to audience, and generalizability of the product. For the appropriateness of the mathematical model, experts wanted the complexity of the problem to be addressed and rationalized, and in particular for the *Just-in-Time* MEA wanted the model produced by the team to go beyond using

just a measure of central tendency—thus addressing the conceptual understandings imbedded in this MEA. A high quality solution that attended to the audience was described as a product that clearly and effectively communicates the model to the client. A high quality solution that demonstrated a generalizable model was described as a product that goes beyond being useful to its creators (student team) to being useful for others (client) and usable on a variety of data sets. These characteristics were articulated in an independent statement and then mapped onto the adapted QAG. A revised evaluation tool was produced and the most significant change was that it included MEA-specific criteria related to the conceptual understandings imbedded in a given MEA.

The third expert panel activity, completed by individuals after the workshop, involved grading 10 pieces of prototypical teamwork, writing notes on the products to explain the score given, and overall approval and/or feedback on the proposed evaluation package. The evaluation tools were approved 'as is' by the experts.

2.2 Building reliability (Stage II)

The goal of Stage II was to involve experienced teaching assistants in applying and revising the evaluation package, since they could anticipate issues that might arise with the general TA population. Those selected to serve on this TA panel were graduate students who had experience in the first-year engineering course and a range of experience with MEAs. The session began with an overview of the evaluation package, and then the TAs individually applied the evaluation tools to five pieces of prototypical teamwork. They compared their evaluation to that of the experts, discussed their scores, came to consensus, and repeated the process with five additional team products. Throughout, the TA conversation flowed freely back and forth with the researchers and an external evaluator, discussing the practicalities of scoring, raising questions for clarification, and making suggestions for revision.

The practical perspective of the experienced TAs provided information that led to the greatest revisions throughout the study. Their comments and perspectives revealed the need to simplify the general rubric for application to all MEAs. However, they also said the general rubric needs task-specific tools that explain and illustrate how the general rubric should be applied to a particular MEA. The input from the experienced TAs, researcher observational field notes, and review by the external evaluator (an assessment expert) led to the development of two evaluation tools: Instructors' MEA Assessment/Evaluation Package (Instructor's Package) and the MEA Feedback and Assessment Rubric (Rubric). The major change in the Rubric was a move from a holistic to a dimensionalized (more analytical) evaluation tool, separately addressing the appropriateness of the mathematical model, the attention to audience, and the generalizability of the model. The Instructor's Package provided MEA-specific guidance for applying the Rubric to a particular MEA. The Rubric items represent the learning outcomes for MEAs as used in the first-year course, whereas the Instructor's Package articulates the instantiation of the learning outcomes with respect to the specific MEA.

Given the useful feedback by the experienced TAs, the format of the Instructor's Package and the Rubric, as developed at this stage, remained relatively constant throughout the rest of the study. (See Appendix A and B for the evaluation tools resulting after Stage IV.) Additional evaluation tools were designed to enhance TA understanding of assessment and evaluation of team work products, including expert evaluations of prototypical student work.

2.3 Assessing reliability (Stages III & IV)

In the move to full-scale course implementation, assessment training was provided for all TAs; this training was in addition to university and course-specific training (as described in [17]). The 2007 (Stage III) and 2008 (Stage IV) training sessions had similar structure: an initial block of training emphasizing the Instructor's Package and the Rubric as they pertained to the first MEA implemented in the semester (4 hours in 2007 [17], 8 hours in 2008); additional one-hour sessions emphasizing the evaluation tools for each additional MEA implemented; and TA assessment of prototypical team responses. The Stage IV increase in the initial block of training was in response to faculty concerns about the TAs' ability to assess teams' mathematical models. Stage IV also included a requirement that TAs apply the students' models to the given data sets, summarize mathematical approaches used by the teams in their models, and summarize the rationalizations and assumptions supporting the teams' solutions. These techniques were intended to help TAs identify real problems with the teams' mathematical approaches, rationales, and assumptions.

The assessment of the quality of TA scoring for the purpose of improving the evaluation package has been based on two sources of data. A measure of inter-rater reliability involves comparing TA scores to an expert score on five pieces of prototypical teamwork for a given MEA (in this paper, the *Just-in-Time* MEA). In this case, the expert was a member of the research team, a long-time instructor for the first-year engineering course, a member of the original expert panel, and the one conducting the TA training with MEAs. Inter-rater reliability data, along with observational data gathered informally during training sessions, and an examination of TAs written feedback on prototypical teamwork has been useful for making decisions about the nature of the revisions to the evaluation package.

Inter-Rater Reliability. Measuring reliability in

Table 2. Graduate Teaching Assistant Demographics

Semester	TA Status	Total	Domestic/International	Male/Female
Fall 2007	New	9	4/5	7/2
	Returning	10	6/4	9/1
Fall 2008	New	13	1/12	8/5
	Returning	7	2/5	7/0

scoring involves selecting a situation where all TAs score the same set of student work. This situation only happens during TA training while the TAs are in the process of learning to use the evaluation tools. Thus, the reliability results gathered are likely to be an underestimate of the quality of the tools and TAs' ability to score consistently compared to what might be found after the TAs receive formative feedback on their scoring. Note that between Stage III and Stage IV, the sets of prototypical work given to the TAs to grade were different, because the returning TAs had graded the selected teamwork in the previous year. The TAs involved also varied in their experience as TAs in the course, their status as domestic or international students, and gender (See Table 2).

To analyze inter-rater reliability, the expert score for each sample of teamwork is assigned the label; 'E'. Then, each of the TAs scores are compared to the expert score, and labeled in terms of its difference from the expert score. For example, if the expert scored Team A Response as a '2' and a TA scored it as a '3,' the TA data point is labeled 'E+1' (the TA is an 'easier' grader than the expert in this case). Similarly, if another TA scored the work with a '1,' that TAs data point is labeled 'E-1' (indicating that the TA is a 'harder' grader than the expert). An end-in-view [18] is critical in design research in order to know when the object under design meets criteria. In this case, the end-in-view is to have at least 90% of the TAs scores for each dimension within one point of the expert's score, a criteria adapted from Herman, J., Aschbacker, P. & Winter, L. [19] on each of the dimensions scored: appropriateness of mathematical model, attention to audience and generalizability.

Scoring for the first dimension, *appropriateness of mathematical model*, requires TAs to score each sample response from 0 to 4. A score of 4 is only attainable if the student work is deemed to address the complexity of the problem (demonstrates conceptual understandings imbedded in the specific MEA—see Appendix A), use all data types or justify not using certain data types, and contain rationales for critical steps. For example, in the sample response in Fig. 2, a score of 4 would apply because the students went beyond using a measure of central tendency to appropriately using mean, standard deviation, and the distribution of the data, and justifications were articulated. The TA scores from Fall 2007 (Stage III) and Fall 2008 (Stage IV) for mathematical model are represented in Fig. 2. From this graph, a shift toward easier grading is notable over the two stages (the mean difference of the TA scores from the expert scores makes a statistically significant shift from –0.705 in 2007 to 0.180 in 2008, using a two-tailed t-test $p < 0.001$) and the grading is more accurate (since the mean score is closer to 0 in 2008). The spread of the scores (standard deviation) did not change in statistically significant ways (0.846 in 2007 to 1.009 in 2008, using a two-tailed F-test $p = 0.09$). The consistency in grading (defined as TA scores being within 1 point of the E grade) stayed the same (86.4% in 2007 and 85.0% in 2008), still short of the 90% goal.

TAs scored each sample response for the second dimension, *attention to audience,* from 1 to 4 using the Instructor's Package. A score of 4 is only attainable if the model produced by the team addresses three aspects: (1) the team's results from applying the procedure can be replicated, (2) the procedure is easy for the client to understand and replicate, and (3) the description of the procedure contains no extraneous information. A miscommunication and an inadvertent change in the task used during the Stage IV TA training led to results that were not interpretable. Thus, the TA sub-scores for 'results from applying the procedure' are not included in this analysis. As a result, the possible scores for 'results from applying the procedure' range from 2 to 4. As such, for the sample team response in Fig. 1, a score of 4 would apply because the procedure is well articulated and devoid of extraneous information.

The difference in TA scores as compared to the

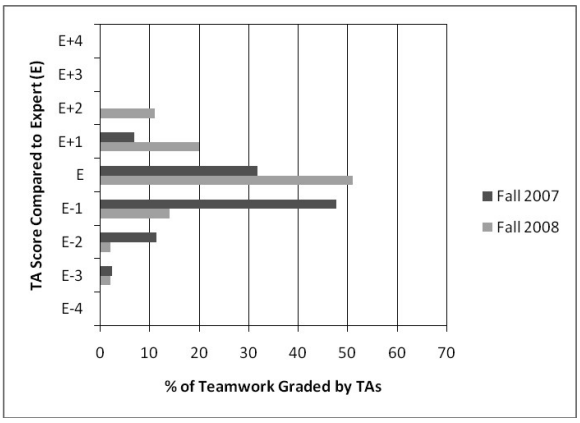

Fig. 2. Mathematical Model Score: TA Score Compared to Expert Score.

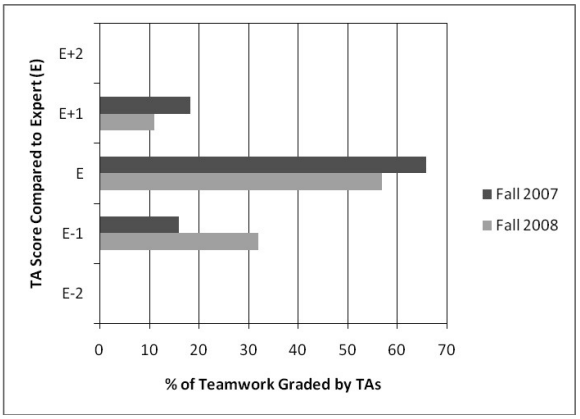

Fig. 3. Audience Score: TA Score Compared to Expert Score.

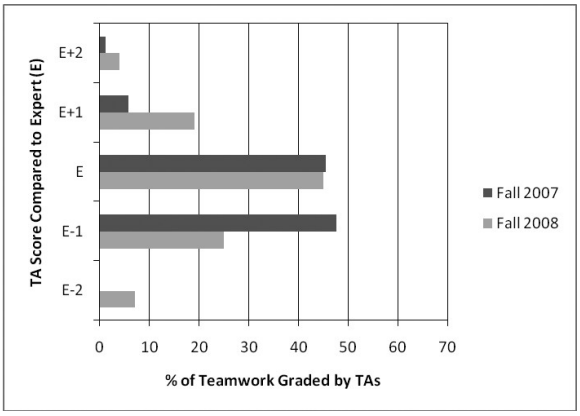

Fig. 4. Generalizability: TA Score Compared to Expert Score.

Expert for 2007 and 2008 for the audience dimension are represented in Fig. 3. From this graph, a shift from easier to harder grading is notable over the two stages (the mean TA scores make a statistically significant shift from 0.023 in 2007 to –0.210 in 2008, using a two-tailed t-test $p < 0.01$) meaning the grading is less accurate (since the mean score is farther from 0 in 2008). The spread of the scores did not change over the two stages as reflected in the standard deviation (0.587 in 2007 to 0.624 in 2008, using a two-tailed F-test $p = 0.56$). The end-in-view, 90% of the scores within one point, could not be applied to this dimension because the range of 2 to 4 points falls outside the required 4-point range. Therefore, considering the exact agreement between TAs and expert scores revealed some decrease in inter-rater reliability (65.9% in 2007 and 57.0% in 2008).

For the *generalizability* dimension, the possible scores on using the Instructor's Package ranged from 2 to 4. The difference in the TA scores as compared to the Expert for 2007 and 2008 concerning aspects of generalizability are shown in Fig. 4. From this graph, a shift toward the mean expert score is notable over the two stages (the difference in the TA scores makes a statistically significant shift from –0.398 in 2007 to –0.120 in 2008, using a two-tailed t-test $p < 0.05$) meaning the grading is more accurate (since the mean score is closer to 0 in 2008). However, the spread of the scores increases statistically significantly (0.653 in 2007 to 0.935 in 2008, using a two-tailed F-test $p < 0.001$). Hence, the consistency in grading (defined as TA scores being within 1 point of the E grade) declined over the two stages (98.9% in 2007 and 89.0% in 2008) due to the increased spread.

Revisions to the evaluation tools. Using design research methodology, information about measures of reliability over the two stages is combined with informally gathered information from observations during training sessions, examination of written comments by the TAs, and professional judgment to make revisions to the evaluation package. Identification and examination of, and reflection on difficulties that TAs encountered in differentiating among the subscore aspects of the dimensions has led to a series of informed revisions. The revisions to the Instructor's Package and Rubric over Stages III and IV can be summarized as no change to the *appropriateness of the mathematical model* dimension, change in language used related to *generalizability,* and shifting language from one dimension to another. Revisions to the TA training included a change in how TAs were directed to interpret students responses and a doubling of initial training time.

No changes were made to the Instructor's Package and Rubric for the *appropriateness of mathematical model* dimension between Stages III and IV, though more discussion of statistical concepts embedded in the *Just-In-Time* MEA occurred during TA training. This was not sufficient to improve the reliability of TA scores between Stages III and IV. Future work is planned to improve the reliability of TA scoring in response to evidence that some of the TAs struggle with the statistical content embedded in the *Just-in-Time* MEA. For example, TA written (or lack of written) feedback on the prototypical teamwork sometimes fails to (appropriately) comment on teams' statistical conceptual misunderstandings. Some TAs have also verbalized a lack of confidence in their own ability to interpret the quality of teamwork in this regard. This information suggests the need for more work on this dimension, including an investigation into TAs statistical preparedness, which may be followed by revisions to the TA training or the evaluation tools.

The *generalizability* and *attention to audience* dimensions for scoring were difficult for the researchers to differentiate, articulate, illustrate and communicate to the TAs in Stages III and IV. The characteristics of these two expert-identified dimensions are confounded because a good solution (i.e., a model) goes beyond producing something usable for one's self to producing something usable by others—meaning the solution is simultaneously generalizable *and* well articulated. One of the design principles for developing MEAs described in the literature [3] is that the solutions to

the problems under design must be s*hare-able* and *re-usable.* In Stage III, these terms were adopted from the literature to characterize the *generalizability* dimension. A model was described as share-able if the client was able to modify the model for slightly different situations, and a model was described as re-usable if the client was able to use the model for new but similar situations.

Implementation during Stage III indicated problems existed with the use of these terms. The TAs pointed out that they were giving similar comments in their feedback to students on both dimensions, suggesting that there was overlap between the two dimensions. In Stage IV, the term *modifiability* (also a term used to describe generalizability in the literature [20]) was introduced. The Rubric (see Appendix B) was revised to combine re-usability and modifiability to characterize the *generalizability* dimension. The term *share-ability* was then explained in terms of the model being used by the client to reproduce results, and thus became synonymous with the *attention to audience* dimension. This attention to language marks the beginning of a journey through iterations of using and describing these terms. This work to articulate the aspects of the *generalizability* dimension continues.

TA training for Stage IV was revised to address concerns about their interpretation of teams' mathematical models. A push was made to get the TAs to actually attempt to apply the teams' procedures to the data provided, assuming that this technique would help TAs understand more deeply the teams' solutions. While this revision was intended to improve scoring and feedback on team work in the *appropriate mathematical models* dimension, there seemed to be a negative impact on the *attention to audience* dimension. Specifically, the TAs began grading harder than the expert in this dimension perhaps because careful interpretation of teams' solutions revealed more errors, omissions and miscommunications. Another possible explanation for the harder grading could be due to the use of prototypical teamwork that contained more rationales and explanations in Fall 2008 as compared to Fall 2007. Perhaps the work was more challenging to interpret, and TAs translated that challenge into identifying more errors, omissions and miscommunications. Another explanation may be related to the larger number of international TAs during Stage IV compared to Stage III. Could it be that language difficulties lent to less lenient grading of complex answers? Could it be that the international students come with different standards of what constitutes a well-written response? All of these questions remain open for future research.

3. CONCLUSIONS AND IMPLICATIONS

The challenge of designing evaluation tools for open-ended problems embedded in a larger educational system raises issues that can be addressed by tapping various educational research methods. Open-ended problems that reveal authentic insights into what students know and can do are necessary for the development of a valid evaluation system; the selection of such problems is addressed by a models-and-modeling perspective [9]. Given that in practice the improvement of educational programs evolves over time, based on reflection and revision of current practice, a research methodology formalizes the iterative process of testing and revising desired evaluation tools. Thus, design research methodology [4–6] is used to plan for stages of work, assess outcomes at each stage, and inform subsequent stages of work. Finally, because the educational endeavor simultaneously involved students, TAs, and researchers in the context of their work, multi-tiered teaching experiment methodology [7] ensures the production of an informative trail of documentation that can be used to guide subsequent revisions to the evaluation tools. As a result of combining these methodologies, the development of evaluation tools and training for TA use has grown from an initial grounding in expert engineers' perceptions of important criteria for interpreting work, has engaged experienced TAs in the design process, and has led to large scale implementation that generates information for on-going fine-tuning in subsequent implementations.

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.

REFERENCES

1. D. C. Davis, K. L., Gentili, M. S. Trevisan and D. E. Calkins. Engineering design assessment process and scoring scales for program improvement and accountability. *Journal for Engineering Education*, **91**(2), 2002, pp. 211–221.
2. P. Black, C. Harrison, C. Lee, B. Marshall and D. William, *Assessment for learning: Putting it into practice*, 2003. Berkshire, England: Open University Press.
3. 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.

4. 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.
5. B. Bannan-Ritland, The role of design in research: The integrative learning design framework. *Educational Researcher*, **32**, 2003, pp. 21–24.
6. T. D. Lamberg and J. A. Middleton, Design research perspectives on transitioning from individual microgenetic interviews to a whole-class teaching experiment. **38**, 2009, pp. 233–245.
7. R. Lesh and A. Kelly, Multitiered teaching experiments. 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. 197–230.
8. J. Gainsburg, The mathematical disposition of structural engineers. *Journal for Research in Mathematics Education*, **38**(5), 2007, pp. 477–506.
9. 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.
10. H. A. Diefes-Dux, M. A. Hjalmarson, T. K. Miller and R. Lesh, Chapter 2: Model-Eliciting Activities for Engineering Education. In J. S. Zawojewski, H. A. Diefes-Dux, and K. J. Bowman (Eds.) *Models and Modeling in Engineering Education: Designing Experiences for All Students*. Sense Publishers, Rotterdam, the Netherlands, 2008, pp. 17–35.
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.
13. M. Hjalmarson. Engineering students designing a statistical procedure. *Journal of Mathematical Behavior*, **26**(2), 2007, pp. 178–188.
14. H. A. Diefes-Dux and P. K. Imbrie, Chapter 4: Modeling Activities in a First-Year Engineering Course. In J. S. Zawojewski, H. A. Diefes-Dux and K. J. Bowman (eds) *Models and Modeling in Engineering Education: Designing Experiences for All Students*. Sense Publishers, Rotterdam, the Netherlands, 2008, pp. 55–92.
15. G. Carmona-Dominguez. *Designing and assessment tool to assess students' mathematical knowledge,* 2004, Dissertation presented at Purdue University.
16. K. K. Clark and R. Lesh, Whodunit? Exploring proportional reasoning through the footprint problem. *School Science and Mathematics*, **103**(2), 2003, pp. 92–99.
17. H. A. Diefes-Dux, K. Osburn, B. Capobianco and T. Wood, Chapter 12: On the Front Line—Learning from the Teaching Assistants. in J. S. Zawojewski, H. A. Diefes-Dux, and K. J. Bowman (eds) *Models and Modeling in Engineering Education: Designing Experiences for All Students*. Sense Publishers, Rotterdam, the Netherlands, 2008, pp. 225–255.
18. L. English and R. Lesh, Ends-in-view problems. In R. Lesh and H. Doerr (eds), *Beyond Constructivism: Models and Modeling Perspectives on Mathematics Problem Solving, Learning, and Teaching*. Lawrence Erlbaum, Mahwah, New Jersey, 2003, pp. 297–316.
19. J. L. Herman, P. R. Aschbacker and L. Winter, *A Practical Guide to Alternative Assessment.* Association for Supervision and Curriculum Development (ASCD), Arlington, VA, 1992.
20. R. Lesh, K. Cramer, H. M. Doerr, T. Post and J. S. Zawojewski. *Model development sequences.* in R. Lesh and H. M. Doerr (eds) *Beyond Constructivism.* Lawrence Erlbaum Associates, Mahwah, NJ, 2003, pp. 35–58.

APPENDIX A
INTRUCTORS' MEA ASSESSMENT/EVALUATION PACKAGE (I-MAP)
JUST-IN-TIME MANUFACTURING MEA
(Fall 2008 Implementation)

Appropriateness of the Mathematical Model

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 defined. Completes the sentence, the ranking procedure is based on . . .*
  - *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)

- *Critical steps that needs 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.*

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 the variability, but not the distribution, of these data.
- 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 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 statement of what defines the overall ranking.
- Mathematical detail should be clear from start to finish.
- Mathematical errors should be eliminated.
- Additional but separate LEVEL 4 criteria:
- Rationales for the critical steps in the procedure must be provided. (If the rationales provided are not correct, this is FALSE. If they just need minor clean-up/clarification this is TRUE.)
- If all data provided is not used in the mathematical model, this must be explained or justified. (If the justifications are not correct this is FALSE. If they just need minor clean-up/clarification this is TRUE.)

Generalizability of the Model: Re-Usability and Modifiability

The mathematical model produced must be *Re-usable* (the client can use it for new but similar situations) and *Modifiability* (the client can modify it easily for slightly different situations). 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 a variety of data sets. The model may be in the form of a procedure or explanation that accomplishes a task, makes a decision, or fills a need for a client.

Further, 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. 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.

At a minimum, the mathematical model should include assumptions about the situation and the types of data to which the procedure can be applied. Hard-coded quantitative values imbedded in a procedure require explicit assumptions or explanations.

If the mathematical model is not developed in enough detail to clearly demonstrate that it works on the data provided, it cannot be considered re-usable and modifiable.

Student teams should state that the procedure 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.

Students should also indicate limitations of their procedure. Limitations might be centered around hard-coded quantitative values imbedded in a procedure. These require explicit assumptions or explanations. Hard-coded values might include: an indication of what is considered late, ways to parse the data (related to degree of lateness), weighting factors.

Level 2—Missing all or most of the standard Introduction parts A & C. (Part B is part of the mathematical model criteria)

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.

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.

Level 3— ~2–3 things need work, typically from the standard introduction (criteria for success, constraints, assumptions, or limitations) or implicit assumptions.

Level 4— ~1 thing needs work.

Audience (Share-ability)

***Effectively communicating to the client*: The mathematical model is share-able—the client can use it to reproduce results.**

Although the client (or an intermediary) has 'hired' the consultant team to construct a mathematical model, the client (or the intermediary) needs and wants to understand what the model accomplishes, what trade-offs were involved in creating the model, and how the model works. A high quality product (i.e., model communicated to the client) will clearly, efficiently and completely articulate the steps of the procedure. A high quality product will also illustrate how the model is used on the 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. Given this type of information, the client will be able to intelligently use and/or modify the model for new situations. At a minimum, the results from applying the procedure to the data provided must be presented in the form requested.

RESULTS: Results of applying the procedure MUST be included in the memo. This must include a ranking of all shipping companies (or listing of those discarded prior to ranking) and quantitative (possibly intermediate) results. If results are missing students will receive a Level 1 (D grade) for the MEA.

PROCEDURE: The client 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 client, cannot replicate or generate results, the solution is not Level 3 work.

Memos left in outline form may only receive a maximum Level 3 audience rating.

EXTRANEOUS INFORMATION might include mentions of specific tools (MATLAB or Excel) to complete computations or overly describing how to compute basic statistical measures (e.g. mean, standard deviation).

APPENDIX B
MEA FEEDBACK AND ASSESSMENT RUBRIC
(Fall 2008 Implementation)

Overriding Option

☐ 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. (Level 0)

Mathematical Model

☐ The procedure fully addresses the complexity of the problem. (Level 4)

☐ A procedure moderately addresses the complexity of the problem or contains embedded errors. (Level 3)

☐ A procedure somewhat addresses the complexity of the problem or contains embedded errors. (Level 2)
☐ Does not achieve Level 2. The procedure does not meet minimum requirements for addressing the complexity of the problem or meeting the clients' needs. (Level 1)

The procedure takes into account all types of data provided to generate results OR *reasonably* justifies not using some of the data types provided. (Level 4)

TRUE ☐
FALSE ☐

The procedure is supported with *acceptable* rationales for critical steps in the procedure. (Level 4)

TRUE ☐
FALSE ☐

Provide Written Feedback About the Mathematical Model Here:

Re-Usability and Modifiability

Re-usability = can be used by the client for new but similar situations.
Modifiability = can be modified easily by the client for slightly different situations.

☐ The procedure not only works for the data provided but is clearly re-usable and modifiable. Re-usability and modifiability are made clear by well articulated steps and clearly discussed assumptions about the situation and the types of data to which the procedure can be applied. (Level 4)
☐ The procedure works for the data provided and *might* be re-usable and modifiable, but it is unclear whether the procedure is re-usable and modifiable because assumptions about the situation and/or the types of data that the procedure can be applied to are not clear or not provided. (Level 3)
☐ Does not achieve Level 3. (Level 2)

Provide Written Feedback about Re-Usability and Modifiability Here:

Audience (Share-ability)

Share-ability = can be used by the client to reproduce results

Results from applying the procedure to the data provided are presented in the form requested.

TRUE ☐ (Level 4)
FALSE ☐ (Level 1)

☐ The procedure is easy for the client to understand and replicate. All steps in the procedure are clearly and completely articulated. (Level 4)
☐ The procedure is relatively easy for the client to understand and replicate. One or more of the following are needed to improve the procedure: (1) two or more steps must be written more clearly and/or (2) additional description, example calculations using the data provided, or intermediate results from the data provided are needed to clarify the steps. (Level 3)
☐ Does not achieve Level 3. (Level 2)

There is no extraneous information in the response.

TRUE ☐ (Level 4)
FALSE ☐ (Level 3)

Provide Written Feedback About Audience (Share-ability) Here:

Heidi A. Diefes-Dux is an Associate Professor in the School of Engineering Education at Purdue University. She received her B.S. and M.S. in Food Science from Cornell University and her Ph.D. in Food Process Engineering from the Department of Agricultural and Biological Engineering at Purdue University. Since 1999, she has been a faculty member within the First-Year Engineering Program at Purdue. 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.

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 of a models-and-modeling perspective in the development of problem-solving experiences as sites for research and assessment in the context of program improvement.

Margret A. 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.