DEVELOPMENT OF A RUBRIC FOR USE IN ASSESSING TRANSFER OF LEARNING IN MIDDLE GRADES ENGINEERING PROGRAM PARTICIPANTS

Jessica M. Harlan
Department of Professional Studies
University of South Alabama
jmh1324@jagmail.southalabama.edu

Melissa Dean
Mobile Area Education Foundation
mdean@maef.net

James P. Van Haneghan
Department of Professional Studies
University of South Alabama
jvanhane@southalabama.edu

Program Background

The present study examines the development of a rubric used to score an assessment tool utilized in the evaluation of Engaging Youth through Engineering (EYE) Modules. The EYE Modules are a set of eight comprehensive and extensive instructional units for middle grades math and science teachers to implement through collaboration in both mathematics and science classes. Each Module provides students with opportunities to engineer solutions to interesting problems relevant today through hands-on and practical applications. They address science, technology, engineering, and mathematics (STEM) content and practices that fill gaps between state-mandated and tested content and what business and industry say they need, including innovative problem solving, communication, and teamwork skills. Module specific professional development and implementation kits accompany each Module. The set of 8 Modules, along with their grade level "Launcher" lessons, involve about 50 hours of STEM exposure. Each EYE Module requires a combination of 6 to 8 hours of class time and 1) addresses an engineering design challenge around issues related to National Academy of Engineering's (NAE) Grand Challenges for Engineering[1]; 2) fosters the development of an "engineering habit of mind;" 3) integrates technology and other resources to engage and meet the needs of diverse middle grades students, and 4) deepens understanding of mathematics and science content, with an emphasis on mathematics. The Modules *are not* a complete engineering, technology, or STEM curriculum; rather they are a supplement to and in support of the existing mathematics and science curriculum. They *are* a set of comprehensive and extensive instructional guides that use design challenges and the engineering design process to engage middle grades students in pursuing STEM careers and academics.

Engineering Design Process

Throughout the Modules, we have emphasized the engineering design process. Because there were few measures related to engineering design developed for middle school students, we used the work of Bailey and Szabo[2] on evaluating design processes and Atman, et al.[3], to design an exercise that we believe addresses elements of the design process. Bailey and Szabo[2] focus on how students evaluate design processes. Our assessment includes such an evaluation. Atman, et al.[3] focus on the breadth and depth of thinking surrounding a design problem. Our assumption is that that participating in the Modules should change how students look at and think about engineering problems. The 6th grade problem addresses a civil engineering problem related to trash found in a tidal river after rainstorms. The 7th grade problem involves a situation where two individuals attempt to use algae to make biofuel. The 8th grade problem (currently in development) involves modification of seatbelts to decrease force-related injuries in elderly adults. The problems provide students with a short overview of the problem, followed by 4 sets of questions. The first question asks the students what they would need to think about as they considered the problem. The second set of questions asked about teaming and expertise. The third set asked the students to critique the design process reported. Finally, the last set involved presentation of graphs or tables with data relevant to the problem.

Rubric Development

Initially, we scored the tasks by conducting content analyses to determine whether students mentioned particular themes. For example, did the students mention the need to revise the design plan presented? This methodology provided rich information, but limited our ability to use multiple raters because of concerns about inter-rater consistency. To address issues related to inter-rater consistency, and more formally assess learning outcomes, we developed a rubric for analyzing the assessments. The first step in the development of the rubric was to determine indicators of the outcome this assessment was intended to evaluate: a difference in thinking about and solving engineering design problems. We developed four indicators that parallel the four dimensions noted in the introduction: a) depth and breadth of thinking about the problem, b) identification of the team skills and expertise needed, c) critical analysis of another's application of the design process, and d) ability to use and interpret data to solve an engineering design problem.

Next, we defined the levels of responses for each dimension; we decided to use a scale ranging from 0 to 3. The lowest level, Level 0, indicated either irrelevant responses or no response. At Level 1 were responses that addressed the general problem, but which failed to connect multiple aspects of the problem or to address the system as a whole. At Level 2 were responses that showed an ability to view the problem as a whole and to understand its complexity, but did not demonstrate a strong integration and application of engineering design principles. At Level 3, the highest level, were responses that demonstrated an ability to integrate and apply engineering design principles. For each level on each indicator, a descriptive phrase was provided to assist the rater in identifying at which level a response should be scored. For example, on the dimension "teams, skills, expertise," a Level 2 response would be one that meets the criteria "Mentions two or more areas of content expertise (teaming may be one)." On the same dimension, a Level 3 response would be one that "Describes expertise in specific terms and addresses specific teaming skills."

Rubric Testing and Inter-Rater Reliability

Following the development of the rubric, the rubrics needed testing for clarity and for inter-rater reliability. Ideally, the rubric would have such clarity that all independent raters would assign the same scores on a given assessment every time. To evaluate whether the rubric was successful at meeting this goal, 30 student assessments were copied and distributed to two independent raters. Of these assessments, 15 were from the treatment school and 15 were from the comparison school. Both raters assigned a score of 0, 1, 2, or 3 for each response to each dimension.

Version 1

The percent agreement and Cohen's Kappa for this version of the rubric (see Table 1) showed fair to moderate agreement, but did not reach desired levels. Further examination showed that the raters scored the items within one level of each other between 94 and 100% of the time. This indicated to us that our rubric was approaching reliability, but needed improvement.

The raters were interviewed after initial analyses to discuss areas of discrepancy. It became evident that there was a lack of clarity about the levels on each dimension. Upon further review, we determined that the language on some dimensions was vague, and could cause confusion for raters trying to identify the correct level at which a response should be rated. Additionally, it became clear that the raters needed additional training to orient them to the subject matter and the rubric itself. For example, on the "critical evaluation" dimension of the rubric, a Level 2 score would indicate an answer that "Sees need for improvement, but focus is on a single detail or two and not steps of design process". Because the raters did not have prior experience in engineering or in the program being evaluated, they were unfamiliar with the steps of the engineering design process. This created difficulty when they were trying to determine whether a response met Level 2 criteria.

Table 1. Percent Agreement and Cohen's Kappa

		Depth/Breadth	Teams, Skills, Expertise	Critical Analysis	Data Use/Interpretation
Version 1	κ	.40*	.53**	.42*	.68**
	% agreement	63.6%	69.7%	63.6%	75.8%
Version 2	κ	.72**	.84**	.64**	.71**
	% agreement	83.3%	90.0%	80.0%	83.3%

*$p < .05$, **$p < .001$

Version 2

Based on the feedback from the raters, the rubric was modified to create greater clarity about the various response levels. It was determined that to be properly oriented to the subject and rubric, raters would need to receive training on the engineering design process. Additionally, raters needed to be provided with verbal examples of the types of responses that might appear at each level.

Two independent raters (different from the version 1 raters) were provided with 30 student assessments (also different from those used for version 1 analyses). Of these assessments, 15 were from the treatment school and 15 were from the comparison school. Both raters assigned a score of 0, 1, 2, or 3 for each response to each dimension.

We calculated interrater agreement and Cohen's Kappa on the rubric. We found Kappas ranging from .64 to .83 and interrater percent agreement ranging from 80% to 90% (see Table 1).

The results indicated that we are beginning to see evidence of moderate to substantial agreement. We continue to work on improving reliability, and used feedback from the second set of raters to identify possible areas of improvement. One suggestion is to have engineering professors complete the tasks as subject matter experts. Their responses would serve as the “ideal” responses, and would provide raters with an anchor from which to base their scoring of student responses.

While we are performing continuous improvement on the rubric, we believe we have sufficient evidence of reliability to begin analyzing the assessments. Our application of the rubric to score assessments has led us to rethink the process of using the rubric as a standalone scoring tool. We have begun to go back to capture some nuances of the responses that we felt were not captured in a numeric score. The findings suggest that while the rubric creates consistency in scoring, it does not capture response details or trends. Content analyses of specific responses indicated some areas of learning seemed to transfer at greater rates, and provided context to the scores generated using the rubric. For example, it was noted that EYE students often listed teaming skills as something team members needed. This same pattern of responses was not noted for students in the comparison school. The researchers suggest using the rubric as part of a mixed-methods approach to assessing student transfer of learning.

References

[1] National Academy of Engineering, “NAE Grand Challenges for Engineering”, (n.d.). Retrieved from http://www.engineeringchallenges.org/cms/challenges.aspx

[2] Bailey, R., & Szabo, Z., “Assessing Engineering Design Process Knowledge,” International *Journal of Engineering Education*, **22**(3), 508-518, (2006).

[3] Atman, C. J., Adams, R. S., Cardella, M. E., Turns, J., Mosborg, S., & Saleem, J., “Engineering Design Processes: A Comparison of Students and Expert Practitioners,” *Journal Of Engineering Education*, **96**(4), 359-379, (2007).