Student Experience and Learning with a Formative Sustainable Design Rubric

Dr. Elise Barrella, Wake Forest University

Dr. Elise Barrella is an Assistant Professor and Founding Faculty Member of the Department of Engineering at Wake Forest University. She is passionate about curriculum development, scholarship and student mentoring on transportation systems, sustainability, and engineering design. Dr. Barrella completed her Ph.D. in Civil Engineering at Georgia Tech where she conducted research in transportation and sustainability as part of the Infrastructure Research Group (IRG). In addition to the Ph.D. in Civil Engineering, Dr. Barrella holds a Master of City and Regional Planning (Transportation) from Georgia Institute of Technology and a B.S. in Civil Engineering from Bucknell University. Dr. Barrella has investigated best practices in engineering education since 2003 (at Bucknell University) and began collaborating on sustainable engineering design research while at Georgia Tech. Prior to joining the WFU faculty, she led the junior capstone design sequence at James Madison University, was the inaugural director of the NAE Grand Challenges Program at JMU, and developed first-year coursework.

Mr. Charles McDonald Cowan II, Wake Forest University

Mack Cowan is a recent graduate of James Madison University's Psychological Sciences M.A. program. His primary research interests are sleep and pharmacology using animal models, the psychology of learning, statistical analyses in behavioral research, and more recently, engineering education.

Mr. Justyn Daniel Girdner
Dr. Mary Katherine Watson, The Citadel

Dr. Mary Katherine Watson is currently an Assistant Professor of Civil and Environmental Engineering at The Citadel. Prior to joining the faculty at The Citadel, Dr. Watson earned her PhD in Civil and Environmental Engineering from The Georgia Institute of Technology. She also has BS and MS degrees in Biosystems Engineering from Clemson University. Dr. Watson's research interests are in the areas of engineering education and biological waste treatment.

Robin Anderson, James Madison University

Robin D. Anderson serves as the Academic Unit Head for the Department of Graduate Psychology at James Madison University. She holds a doctorate in Assessment and Measurement. She previously served as the Associate Director of the Center for Assessment and Research Studies at JMU. Her areas of research include assessment practice and engineering education research.

Student experience and learning with a formative sustainable design rubric

Introduction

Sustainable design is not an alternative to traditional engineering design; rather, it is a more holistic design paradigm. Engineering design itself is described as "a creative decision-making process that aims to find an optimal balance of trade-offs in the production of an artifact that best satisfies customer and other stakeholder preferences" [1]. Sustainable design only requires that sustainability principles be incorporated into this complex decision-making process to promote consideration of and balance between the economic, environmental, and social systems during project development [2]. Describing this innovative approach to design, Skerlos et.al. [1] states that sustainable design "brings focus" to the design process, while McLennan [3] describes that sustainable design "expand[s] the definition of good design to include a wider set of issues."

Undergraduate curricula need to be updated to train engineers to operate according to a sustainable design paradigm. Indeed, numerous educators and researchers have reported on efforts to incorporate sustainable design principles into design courses and projects [4, 5]. However, a systematic review of ASEE proceedings showed a lack of rigorously-developed assessment tools for capturing the efficacy of interventions on student sustainable design skills [6]. Rubrics in particular are a promising assessment tool because they can be used for students to scaffold application of sustainable design principles and also by instructors to quantify the impacts of their course innovations [7, 8]. Sustainability rating systems developed for infrastructure systems, like Leadership in Energy and Environmental Design (LEED) or Envision^{TM}, are essentially rubrics for professional projects and have been used to introduce civil engineering students about sustainable design and evaluate capstone projects [4, 9]. Although they are valuable learning tools, professional rating systems are often limited to a subset of project topics (i.e., infrastructure) and may be difficult for students to apply to their more narrowly scoped projects.

In order to address the assessment tool gap, we sought to develop a sustainable design rubric that could be applied to student projects across engineering disciplines and to employ a rigorous construct validation process for the rubric's development. Benson opens her article on construct validation with the statement "Validation is the most crucial step in test development and use because it is the process by which test scores take on meaning" [10]. While this is true for traditional tests, this is also true of performance measures. Benson lays out a multiple stage process for developing a strong program of construct validation. The substantive stage of construct validation involves defining the construct of interest in terms of both its theoretical and empirical domains. At this stage of construct validation, researchers work to define the construct, considering the boundaries of the construct so not to exclude elements of the construct (underrepresentation) and not to include elements not part of the construct (irrelevancy). At the substantive stage, researchers are often engaged with the relevant literature and are working to develop competing hypotheses regarding the construct. At the structural stage, researchers examine the internal consistency of a specific measure of the construct by examining the relationship among observed variables through a series of internal domain studies. Internal

domain studies often examine the intercorrelations among items or subscales. In addition to intercorrelations, researchers often use factor analysis to explore the structure of a measure. The external stage of construct validation involves examining whether the construct of interest relates to other constructs as expected. Is the construct related to constructs we believe it theoretically should be related to? Is it NOT related to constructs it should not be related to? Group differential studies and exploring relationships with other measures are often methods employed in the external stage of construct validation. While the example is often employed with the development of tests measures, the model also is relevant for the development of performance measures, such as the Sustainable Design Rubric [10]. By making slight modifications to the methods used, the researchers were able to employ the Benson model, working through the substantive stage of construct validation and beginning the structural stage.

This paper focuses on efforts in the substantive stage of construct validation for the Sustainable Design Rubric, seeking to answer three primary research questions. First, we wanted to determine to what extent our theoretically and empirically defined rubric criteria were accessible to student audiences. This question was to be answered by using self-report data from students, researcher review of students' scoring of their own projects, and by looking for any criteria or criteria categories which stood out as unusually low-scoring. Second, we wanted to determine the impacts of the use of the rubric as a formative assessment on student design knowledge. Third, we wanted to determine how the rubric criteria could be used for summative project evaluation and program assessment. Ultimately, the goal was for students to assist researchers in identifying criteria that were not relevant to their projects, criteria that were misinterpreted (irrelevancy), or areas that were not covered by the rubric that should have been (underrepresentation). Before describing a new study that invited engineering students to rate their capstone projects through a consensus process, the next section briefly summarizes prior validation efforts and introduces the current rubric we are testing.

Prior Work in Substantive Stage

Prior work to define the construct of Sustainable Design for the rubric included three main activities. First, in 2017 we conducted a systematic literature review of recent literature on sustainability/sustainable design instruction and evaluation to identify themes that were not reflected in an existing sustainable design rubric that had been used to evaluate student projects in civil and environmental engineering [11]. After identifying several themes missing from our original 16 criteria, we ultimately created a comprehensive 34-item rubric across four categories [6]. The expanded set of criteria was then tested against other sustainability frameworks and expert opinions.

Our next step in the substantive stage of validation was to compare the sustainable design rubric to established sustainability frameworks in the field of engineering. This was accomplished in three ways: (1) evaluated the extent to which the frameworks overlapped with our own rubric, (2) evaluated the extent to which overlaps between the published frameworks were not reflected in our 34-item rubric iteration, and (3) determined which individual criteria in our rubric were not reflected within the frameworks. We evaluated the draft criteria against three established sustainability frameworks: the ENVISION™ infrastructure rating system, the STAUNCH© higher education sustainability assessment, and the UN Sustainable Development Goals. As

expected, the evaluation revealed significant overlaps across the three frameworks and our set of criteria but also indicated a few key gaps that were addressed in a future version of the draft rubric [12].

The third step completed for substantive construct validation was to seek feedback from experts across varying engineering disciplines. We sought a ranking of how important each of our criteria was in the eyes of a multidisciplinary sample of engineering professionals both inside and outside of academia. Ultimately, 55 respondents replied to our survey. The importance at which they saw the varying sustainability criteria guided our further refinement of the rubric nearly up to its current state [13].

The Sustainable Design Rubric currently includes fourteen criteria loosely grouped into five categories as listed below.

Environmental Category
A1. Minimizes the use of non-replenishable raw materials; requires minimal energy input or uses renewable energy sources
A2. Minimizes quantity of consumable waste (e.g., water, materials) output; manages quantity and quality (benign, usefulness) of waste
A3. Protects or enhances natural ecosystems (water, air, soils, flora, fauna, etc.)

Social Category
B1. Identifies and engages stakeholders in the design process
B2. Addresses needs of diverse stakeholders, acknowledging culture and other differences among individuals and groups
B3. Protects human health and physical safety of users and society
B4. Promotes human well-being and enhances quality of life for users and society

Economic Category
C1. Evaluates economic impacts of environmental design criterion
C2. Evaluates economic impacts of a social design criterion
C3. Considers affordability for users and/or demonstrates cost competitiveness or cost reduction for client/sponsor
C4. Evaluates economic costs and benefits to inform decisions

Trade-off Category
T. Final design impacted by trade-offs among environmental, social, and economic criteria and reflects balance of dimensions

Bonus Category
X1. Uses and/or creates innovation(s) in its specific field to achieve sustainability
X2. Worked with experts from other disciplines to enhance process or final design

Methods

Institutional Context

The current study was run in a small engineering department at a large public university. Engineering faculty in the department represent a variety of engineering disciplines and support students on capstone projects that are interdisciplinary. The students' capstone projects begin in the fall semester of their junior year and take two full academic years to complete. At the end of the spring semester, both junior and senior teams present their progress on these projects at a public event. Examples of capstone projects include designing surgical instruments for external stakeholders and building a high speed fully electric motorcycle for educational demonstrations.

Study Design and Data Collection

In Spring 2018, all 67 junior engineering students from two course sections were given a homework assignment that included evaluating their capstone projects with our sustainable design rubric (see Appendix A for the rubric). Each student belonged to one of fifteen capstone teams (see Appendix B for topic list) and was assigned to evaluate their projects against a subset of rubric criteria (approximately two-thirds of the rubric) This was done to promote meaningful participation in the assignment by reducing total workload. A subset of criteria was assigned such that each team would review the entire rubric and at least two team members would review each criterion in the three main categories. For example, if a group had three members, one group member would have a rubric sheet that would ask them to score their project on categories Environmental and Social, one would be asked to evaluate categories Environmental and Economic, and one student would be asked to evaluate their project in categories Social and Economic. Additionally, each student would complete either the Bonus or Trade-offs category.

Students received a grade for their complete, honest evaluation of their capstone design projects; they were not graded based on their perceived performance via rubric scores. This evaluation was a required homework assignment for all students, but informed consent was collected from students willing to participate in the study. Fifty-one students agreed for us to analyze their data to improve the rubric. The study was approved by the University's institutional review board prior to the homework assignment being distributed to students.

We visited the courses near the end of the semester to go over the homework assignment and answer any questions. Students were instructed to score their draft capstone project report on a subset of criteria (as described above) such that the team as a whole would cover all of the categories and most criteria would be scored by at least 2 raters. Students were told that it would not be unusual to have few criteria receiving the maximum score of '3' and some criteria receiving the minimum score of '0' given that their projects were still in progress. The research team member also reiterated that students' project performance on the rubric would not affect their grade on the assignment or in the course overall.

Students scored each assigned criterion on a 0 to 3 scale, with 0 indicating that the criterion is not evident in project work and 3 indicating a strong consideration of the criterion as evidenced by project work. Students were provided with a variety of examples for each criterion to illustrate potential applications in design projects. In awarding points, students were instructed to consider the criterion's topic and three dimensions:

(1) did documentation of that criterion provide quantitative and qualitative evidence?

(2) did the team consider the entire lifecycle of their product, process, system, or service and reflect long-term thinking?
(3) did the team use a formal method, standard, or best practice for their analysis?

Students were provided with Table 1 to help with evaluating performance for each criterion against the three dimensions. A higher score indicates stronger evidence that a project fulfills a given criterion.

Table 1. *Scoring dimensions to guide evaluation of each criterion*.

Dimension	0 pt	1 pt	2 pt	3 pt
Project documentation provides ***Quantitative and Qualitative*** evidence for design decisions.	No formal quantitative or qualitative evidence is provided or analysis does not support the decisions	Quantitative *or* Qualitative evidence is provided but its impact on decisions is unclear	Quantitative *or* Qualitative evidence was used to clearly support decisions	Quantitative *and* Qualitative evidence was used and clearly supports decisions
Design decisions consider the entire ***Lifecycle*** of a product, process, system, or service and reflect ***long-term thinking***.	Considers only one lifecycle phase in design process and only reflects short-term factors/impacts.	Decision considers more than one phase (partially) or considers a few variables over entire lifecycle; reflects short- and mid-term factors/impacts	Decision partially considers multiple variables over most lifecycle phases; reflects a few long-term factors/impacts	Decision considers all phases and multiple variables; decision reflects long-term thinking and adaptability
Analysis uses a ***formal method***, standard, or best practice	No documentation of formal methods or approaches for analysis	Analysis followed a best practice or formal method and is partially documented	Analysis followed a best practice or formal method and is documented well	More than one formal method or approach was used for analysis and is fully documented.

Students were asked to provide specific examples to support the score that their project received, whether exemplary or minimal (see example rubric in Appendix A).

After individual scoring, the project team was asked to meet and discuss final consensus scores for each of the 14 criteria. Similar to individual scoring, this consensus score required teams to provide a single integer value and a descriptive support for the final score, but it also allowed for students to leave individual feedback regarding the consensus process. This was done to give a voice to dissenting members within the scoring process (though few instances of this seemed to occur).

To complete the assignment, individual students responded to a series of open-ended questions related to their experience using the rubric, including ease/difficulty interpreting criteria, time required to complete the assignment, and thoughts on the consensus process. Students were also asked to complete six Likert scale questions on topics varying from criteria scoring difficulty to their perceived value of the entire exercise. For a full list of free response and Likert scale questions, please see appendix C.

Analysis of Quantitative Ratings

The researchers started data analysis by calculating basic descriptive statistics for all individual and consensus scores that students provided. Average scores for individual and consensus scores were prepared for comparisons. Likert scale reflection questions were aggregated for all participants. Overall responses for each ranking question were plotted on bar graphs to aid in visual analysis.

Analysis of Qualitative Reflection Questions

The qualitative reflection questions were coded using various internally-generated codes. Evidence of immediate agreement versus engagement in a consensus process was noted. If a consensus process occurred, the extent to which an individual gained greater perspective on their project was noted. We went through each free response question related to difficulty and coded the response according to which criteria, or criteria categories, were specified as most and least difficult. We then coded the reason a particular criterion or category was identified as a 'most difficult criteria', if an explanation was given. The qualitative codes for those reasons included "minimal consideration", "difficulty understanding", and "student deemed criterion inapplicable to their project".

Analysis of Student Ratings Quality

In addition to analyzing the students' self-reported scores, we also reviewed the responses from all consenting participants' rubrics and evaluated the quality of their work. For each criterion, two independent raters reviewed each participant's score and evidence and provided a rating of 0 to 2 as follows: (0) justification does not support score/relate to criterion, (1) justification is related to the criterion and somewhat supports score, or (2) justification sufficiently supports the score. This rating was intended to denote the quality of the individual's justification of their final score, rather than to evaluate the performance of a participant's project. For example, if a student gave their project a score of '0' for a specific criterion and then gave detailed reasoning for how and why their project did not yet address that specific sustainability concern, then the response would receive a '2' for a quality score. This task was performed with the goal of gaining a greater understanding of the accuracy with which students evaluated their projects to ensure students were interpreting the criteria and rating scale correctly. Secondarily, we wanted to investigate if student scores were inflated for any criteria or categories.

Results and Findings

Individual Ratings

The first data that we looked at was students' individual ratings to their assigned criteria. Table 2 shows how students overall scored each criterion. Participants (N = 51) scored a total of 406 criteria, with individuals evaluating their projects with a score of 2 most often at a frequency of 49.79%. Participants assigned their projects a 0 least often with a frequency of 9.05%, and scores of 1 and 3 were assigned at nearly identical rates of 23.73% and 23.43%, respectively.

Table 2. *Descriptive statistics for individual and consensus scores for each criterion.*

Criterion	Individual *M* (*SD*)	Consensus *M*
A1 Non-replenishable resources	1.53 (.78)	1.73
A2 Waste	1.56 (.86)	1.73
A3 Ecosystem protection	1.43 (.82)	1.33
B1 Stakeholder engagement	2.56 (.50)	2.67
B2 Diverse cultures and needs	1.68 (.75)	1.67
B3 Human health/safety	2.09 (.69)	2.40
B4 Quality of life	2.41 (.71)	2.67
C1 Economic/environment	1.55 (.95)	1.53
C2 Economic/social	1.62 (.90)	1.93
C3 Affordability, cost competitiveness	2.10 (.82)	2.13
C4 Costs and benefits	1.61 (.89)	1.73
T1 Trade-offs	1.57 (1.08)	1.73
X1 Innovations in field	1.97 (.96)	1.87
X2 Interdisciplinary experts	1.77 (.94)	1.80

The research team members reviewing participants' individual scores and their scores' justifications found that students generally gave themselves relatively appropriate scores, with an average reviewer score of 1.48 on the 0-2 scale (where 0 indicated inadequate evidence and 2 indicated a well-justified score). With that said, there were a few general trends noticed by the reviewers when they evaluated participants' justifications for their scores. Specifically, students would occasionally give their projects credit for addressing two separate criteria using the same support. For example, students were generally too generous in their scoring of the criterion "Worked with experts from other disciplines to enhance process or final design", applying a very liberal interpretation of "other disciplines". Furthermore, many teams also provided collaboration with other disciplines as evidence of "Identifies and engages stakeholders in the design process."

For a few criterion scores, (n = 13 out of 406) participants explicitly scored all three dimensions (i.e., quantitative/qualitative evidence, lifecycle, formal analysis) in their evaluation for each of the criteria. However, most students did not directly cite dimensions when they analyzed their performance relative to the different criteria. The instructions for applying the dimensions were intentionally somewhat open-ended as to whether each dimension should be individually scored, or merely considered for criterion scoring. For this study, the research team wanted to see how students might apply the scoring dimensions when they were not instructed to provide a score for each of the three dimensions on each criterion. Explicitly scoring each dimension on a 0-3 scale and aggregating the dimensions per criteria for a 0-9 scale would be a more direct way to measure project completeness within the sustainability framework, however it would be very time intensive. Other scoring methods were discussed such as normalizing the dimensional

scores for each criterion, using the lowest of the three scores, or averaging the scores. All these, however, involve the participant explicitly scoring each of the criteria within the three dimensions.

The social category was the most robust and had no blank justifications, as opposed to six justifications left completely blank in environmental and three in the economic category. Comparing economic responses with social and environmental, students tended to give themselves zeros more and generally felt less comfortable in the economic category. When responding to the economic criteria, many students used social topics in an attempt to address economic criteria. This may be due to a combination of lack of economic focus in sustainable design teaching, general disregard for economic concerns in sustainability discussions, or that economic considerations are innately harder to tie into a sustainability framework as undergraduate engineering student understand it. The social category tended to have the best observations and score justifications. This could be due to the department's curricular emphasis on stakeholder and user involvement in the design process. In spite of this, students were liberal in their descriptions of the "Working with professionals from other disciplines" criterion, often referencing professors within the engineering department who had experience in other fields.

Team Consensus Ratings

Next, we examined the consensus ratings and justifications provided by each team. The fifteen capstone design teams ultimately provided 210 consensus scores for their projects. Of these 210 responses, '2' was the most common score with a frequency of 50%, with '3' and '1' being a close second and third in frequency at 23% and 21%, respectively. A score of '0' was the least commonly reported score at 5%.

These frequency scores and average ratings for criteria closely reflect the individual scores with a few minor differences. Though the most frequent score was 2 for both categories, only 44% of the individually scored criteria received a 2 while 50% of the consensus scores were 2. The least frequent score was, again, 0 however at an individual score frequency of 9% and a consensus score frequency of 5%, students were almost twice as likely to individually assign their projects a score of 0 on a particular criterion compared to the consensus scores. This may be due to justification for minor credit being brought up by another group member during the consensus process. Overall, there seemed to be most agreement for the bonus categories. For a full breakdown of each criterion's average individual and average consensus score, see Table 2 above.

The construct validation process prioritizes unique criteria that both encompass sustainable design holistically while minimizing redundancy between criteria. To address this concern, the research team correlated all average consensus scores for each category of criteria. Out of all comparisons, only the *Social* and *Economic* categories were moderately correlated (r = .52). The researchers then correlated each individual criterion with every other criterion in the same category. Only "Minimizes the use of non-replenishable raw materials and requires minimal energy input or uses renewable energy sources" and "Protects or enhances natural ecosystems (water, air, soils, flora, fauna, etc.)" were strongly correlated (r = .69). This suggests that most of our criteria are performing a unique function and are therefore not redundant.

Reflection Questions

Participants were given a series of Likert-type and free-response questions at the end of the assignment (see Appendix C for the questions). The majority of participants (n = 47) completed these questions, with at least one participant from each of the 15 teams providing responses. The responses were analyzed via visual analysis of the responses, with all but one question having a skewed distribution with a mode of 5 (see Figure 1, right side). The only question with response data that stood out to us was "I was able to understand each criterion" with a mode of 5 (n = 13) on a 1-7 Likert scale, in which 1 would indicate great difficulty understanding and 7 would indicate great ease of understanding. However, our second most common response was a 3 (n = 11) and the frequency of responses greater than 5 was noticeably smaller than all other Likert based responses.

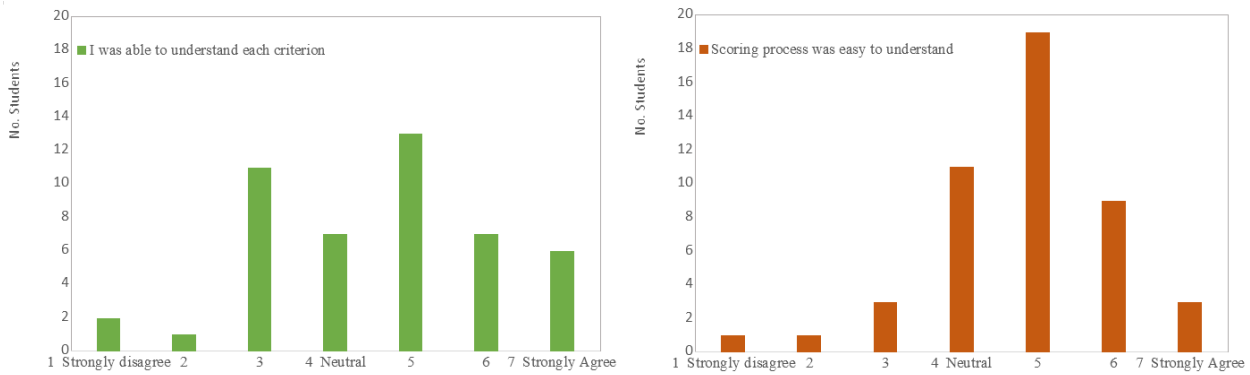

Figure 1. *Frequency of responses for "I was able to understand each criterion" compared to a typical question's distribution.*

In two separate free response questions participants were asked to indicate the criteria, or criteria categories, which were most and least difficult for them to score. These two free response questions were qualitatively coded for criteria or categories and for possible explanations of difficulty. Many participants reported the social (n = 19) and environmental (n = 13) categories as the least difficult. In contrast, few participants reported the economic category as the least difficult (n = 4). Many participants considered a category to be easier to address if they viewed it to be a natural focus of their projects. Specifically, teams reported their projects as socially- (n = 19), environmentally- (n = 9), or economically- (n = 2) focused.

Participants reported the economic (n = 18) category as their most difficult, followed by environmental (n = 13) and social (n = 2). As expected, this is the inverse of responses to the "least difficult" question. For this question, two codes were created to qualify individuals' responses, "Minimal Consideration" and "Difficulty Understanding". We had many responses indicate that "Minimal Consideration" was the reason they noted the economic category as their most difficult (n = 16), compared to far smaller numbers for environmental (n = 5) and social (n = 2). "Difficulty Understanding" was coded less frequently in general, but economic (n = 5) still led over the environmental (n = 2) and social (n = 1) categories.

Qualitative data analysis concluded with multiple passes of coding the following open-ended free response prompt: "What did you learn from discussing scores with your teammates and trying to

reach agreement? Have you identified any areas for improvement or future work?" The first portion of this prompt was coded as either "CONSENSUS" or "AGREEMENT" to denote whether an individual's group went through a process to reach consensus or seemed to simply agree on their scores with limited discussion. The responses we received were roughly split down the middle with "AGREEMENT" (n = 16) having a slightly greater presence than "CONSENSUS" (n = 14). The question was then coded with multiple, non-exclusive codes including "AoI" (n = 7), which denoted that a participant had specified a particular area of improvement to work on in their project, "METACHANGE" (n = 6), which denoted a change in that participant and/or their team's perspective on the project, and "CHANGED" (n = 2), which denoted a specific and direct change to the project as a result of this rubric exercise.

Though there was great variability in the responses to the last reflection question, both in terms of content and response length, student responses indicated that the scoring process would have a positive impact on their final projects. For example, one student stated that "From discussing scores with teammates, I learned that the project has a chance to be sustainable, but that's far down the road. I also learned that we could better advertise our project as socially and economically friendly, which is something we never considered before" was one response indicating that a project team was now considering two previously ignored aspects of sustainability. Another student observed that "...we clearly need to evaluate the economic impacts of social design more than we have. We also need to factor in the environmental designs of a biopharma system...". This statement, and other similar responses, demonstrated to the research team that the scoring process had encouraged greater sustainability considerations in these student projects.

Discussion

Overall, students were able to interpret the sustainable design criteria and provide evidence for how their project decisions reflected the criteria (or not). Scores may have been slightly inflated for projects that were still in-progress, however, this seemed to be consistent across teams, with the exception of a couple of outliers. That said, student scores and the evaluation of their justifications suggested that some criteria could use additional definitions or clarifications to help students distinguish between criteria. For example, the social criteria seemed to overlap with each other (B1 and B2) and with criteria in other categories (B1 and X2) based on how students justified their scores for those criteria. Students also had difficulty with the economic criteria. While this seemed in large part due to a lack of consideration of economic impacts of their projects (based on reflection questions), they also seemed to frequently double count environmental decisions as economic decisions without providing separate evidence. Do we need to add additional descriptors to distinguish criteria? Based on other analysis, we may not need to revise criteria but rather add more detailed instructions for students/professors using the rubric. For example, preventing students from using a single action/decision to satisfy two criteria might resolve issues with "double counting". The finding that only two criteria within the same category were significantly correlated in consensus scores and that only two criteria categories' scores were correlated supports the idea that our criteria are each performing a different function.

In addition to difficulty with specific criteria, students also appeared to have difficulty in considering or explaining all three "dimensions" of evidence (quant/qual evidence, lifecycle,

formal analysis) for each criterion. Instead of complicating the scoring, we may instead ask student or faculty raters to apply these dimensions to the overall project performance or, at most, to each category. The more holistic rating could then be used as a weighting factor for the project's final sustainable design score.

In terms of process, there seemed to be value in having students complete both individual and team scoring in a step-wise fashion. As a formative tool, students found value in discussing scores with each other and in some cases expanded their view of how criteria applied to their projects. The consensus process also generally led to better evidence supporting their ratings.

In order to better understand how the rubric performs and aids student learning, testing the rubric in a different institutional context or comparing faculty and student reviews would provide new insights and are planned for future work. In particular, testing in a different institutional context will provide insights on whether this study's institutional context (large university with a small, multi-disciplinary engineering department) leads to unique student and faculty experiences with the rubric. A different future direction could be summative assessment of the fifteen capstone projects, in which we could evaluate the extent to which teams implemented sustainability measures and the extent to which they performed the project changes that they mentioned in this study. Apart from testing different implementation methods for the construct validation process, the rubric results were used by the study institution to support ABET accreditation efforts. Descriptive statistics summarizing team consensus scores and students' identified areas for improvement were provided as assessment evidence for continuously improving student achievement of sustainable design learning outcomes and identifying curricular areas that could be strengthened.

Moving forward, we will begin to address the structural stage of construct validation. In particular, future research will focus on whether scores on criteria grouped under specific domains are more closely related to one another than to criteria in different domains. It is important to note that while the Benson model is presented as a series of stages, researchers often move back and forth among the stages. What one learns in one stage may prompt the researchers to return to a previous stage for additional study. The researchers, while exploring whether the internal structure of the rubric fits with the theoretical and empirical models, will also continue to refine those models based on what is learned in both the structural and external stages of construct validation.

Conclusions

Based on the scoring results and the reflection questions, students had the most difficulty rating and justifying the economic criteria, usually because they had not yet considered economic costs and benefits of their project. In some cases, students had difficulty understanding a criterion and how it applied to their project. The social criteria were deemed easiest to apply because students saw direct connection to project work they had already completed. That said, high ratings were often not strongly justified, indicating room for continued improvement in engaging stakeholders and considering their needs. Environmental criteria earned mixed results, with most students finding the criteria relevant to their project but with little direct evidence at the mid-point in their projects. Most students identified areas for additional learning or project improvement as a result

of completing the individual scoring and consensus process, which supports using the rubric for formative assessment. As a result of the students' performance and reflection, the research team will re-evaluate the criteria and application of the rubric to support construct validation.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant No. 1811170 Developing and Assessing Engineering Students' Cognitive Flexibility in the Domain of Sustainable Design. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] Skerlos, S.J., W.R. Morrow, and J.J. Michalek, "Sustainable Design Engineering and Science: Selected Challenges and Case Studies." *Sustainability Science and Engineering: Defining Principles*, M.A. Abraham, Editor. 2006, Elsevier, B.V.: Amsterdam, The Netherlands. p. 467-515.

[2] Mihelcic, J.R., et al., *Sustainability Science and Engineering: The Emergence of a New Metadiscipline*. Environmental Science & Technology, 2003. 37(23): p. 5314-5324.

[3] McLennan, J.F., *The Philosophy of Sustainable Design: The Future of Architecture.* 2004, Bainbridge Island, WA: Ecotone Publishing Company LLC.

[4] Burian, S. J. "Using a sustainable infrastructure rating system in the civil engineering capstone design course." in *Proceedings of the 2014 ASEE Annual Conference & Exposition. June, 2014*, Available: https://peer.asee.org/23281

[5] Cecere, J. "Integrating Sustainability in an Engineering Capstone Course." in *Proceedings of the 2018 ASEE Conference for Industry and Education Collaboration. August, 2018, San Antonio, TX.* Available: https://peer.asee.org/31342

[6] Watson, M. K., and E. Barrella. "A systematic review of sustainability assessments in ASEE proceedings." in *Proceedings of the 2017 ASEE Annual Conference and Exposition, June 2017, Columbus, OI, USA*, pp. 25-28.

[7] Estell, J. K., and J. Hurtig. "Using rubrics for the assessment of senior design projects." in *Proceedings of the 2006 ASEE Annual Conference & Exposition: Excellence in Education*. 2006.

[8] Watson, M. K., and E. Barrella. "Using concept maps to explore the impacts of a learning-cycle-based sustainability module implemented in two institutional contexts." *Journal of Professional Issues in Engineering Education and Practice* 143, no. 2 (2016): D4016001.

[9] Sisiopiku, V., Peters, R.W., and O. E. Ramadan. "Introducing sustainability into the civil engineering curriculum." in *2015 ASEE Annual Conference and Exhibition*. *June, 2018, Seattle, WA.*

[10] Benson, J. "Developing a strong program of construct validation: A test anxiety example." *Educational Measurement: Issues and Practice* 17, no. 1 (1998): 10-17.

[11] Watson, M. K., and E. Barrella. "A systematic review of sustainability assessments in ASEE proceedings." in *Proceedings of the 2017 ASEE Annual Conference and Exposition, Columbus, OH, USA*, pp. 25-28. 2017.

[12] Cowan, C.M., Barrella, E., Watson, M.K., and Anderson, R. “Validating Content of a Sustainable Design Rubric Using Established Frameworks.” in *Proceedings of the 2017 ASEE Annual Conference and Exposition, June, 2017, Columbus, OH.*
[13] Watson, M.K., Barrella, E., Cowan, C.M., & Anderson, C.M. “Validating a Sustainable Design Rubric by Surveying Engineering Educators.” in *Proceedings of the 2018 ASEE Annual Conference and Exposition, June, 2018, Salt Lake City, UT.*

Appendix A. Sustainable Design Rubric for Individual Ratings

Criterion	**Earned Points (_/3)** *Instructions: Indicate points earned, considering three dimensions (see next table).*	**Evidence supporting your ratings** *Instructions: List examples of how your project satisfies each criterion. Provide a page number and brief description for examples that are documented in your project report. You may also describe decisions that are not well documented in your report. Provide citations (e.g., a specific code or best practice) when needed to support your examples.*
Environmental Category		
A1. Minimizes the use of non-replenishable raw materials; requires minimal energy input or uses renewable energy sources		
A2. Minimizes quantity of consumable waste (e.g., water, materials) output; manages quantity and quality (benign, usefulness) of waste		
A3. Protects or enhances natural ecosystems (water, air, soils, flora, fauna, etc.)		
Social Category		
B1. Identifies and engages stakeholders in the design process		
B2. Addresses needs of diverse stakeholders, acknowledging culture and other differences among individuals and groups		
B3. Protects human health and physical safety of users and society		

B4. Promotes human well-being and enhances quality of life for users and society		
Economic Category		
C1. Evaluates economic impacts of environmental design criterion		
C2. Evaluates economic impacts of a social design criterion		
C3. Considers affordability for users and/or demonstrates cost competitiveness or cost reduction for client/sponsor		
C4. Evaluates economic costs and benefits to inform decisions		
Trade-off Category (consider project holistically)		
T. Final design impacted by trade-offs among environmental, social, and economic criteria and reflects balance of dimensions		
Bonus Category (consider project holistically)		
X1. Uses and/or creates innovation(s) in its specific field to achieve sustainability		
X2. Worked with experts from other disciplines to enhance process or final design		

Appendix B. List of Capstone Design Project Topics

1. Eco car competition
2. Wind harvesting along highways
3. Collegiate Wind Competition (2 teams)
4. Drones for agricultural monitoring
5. Drones for climate effects monitoring in developing countries
6. Water treatment for mine drainage
7. Greenway design
8. Surgical instruments
9. Green buildings and historic properties
10. Electric vehicle educational demo
11. Leg prosthetics
12. Coffee roasting process
13. Biopharmaceutical process
14. Surgical robots

Appendix C

Reflection Questions: (10 - 15 minutes)
After completing your individual ratings, consider questions 1-3 and then after discussing consensus scores with your teammates, respond to the remaining questions. Reflection should be completed individually but you may compare with teammates or classmates after you submit the assignment.

Open-ended:

1) Which criteria were easiest to score for your project? Why?
2) Which criteria were most difficult to score for your project? Why?
3) How much time did it take you to individually complete your scores?
4) How much time did it take you to arrive at consensus scores?
5) What did you learn from discussing scores with your teammates and trying to reach agreement? Have you identified any areas for improvement or future work?

Likert:

a. The rubric templates, with score and evidence columns, were easy to use.
b. The rubric had an appropriate amount of criteria to measure sustainability.
c. In general, I was able to understand the meaning of each criterion.
d. The scoring process, individual scoring and then team consensus, was easy to understand.
e. Creating consensus scores contributed to my overall understanding of the project's sustainability.
f. Using the rubric and results of the scoring process will help me improve my capstone project.