ASEE's 123rd Annual • Conference & Exposition • New Orleans, LA • June 26-29, 2016

New Orleans JAZZED about Engineering Education

Paper ID #14781

Rubric-Based Energy Literacy Assessment of Student Posters: Effects of Extended Calibration and Addition of Raters

Quinn Langfitt, Washington State University

Quinn is a PhD candidate in the Department of Civil and Environmental Engineering at Washington State University. His research is mostly focused on sustainability, including work on life cycle assessment and energy literacy assessment.

Dr. Liv Haselbach P.E., Washington State University

Dr. Liv Haselbach is the author of the McGraw-Hill GreenSource book, The Engineering Guide to LEED-New Construction, Sustainable Construction for Engineers. She has authored numerous papers on sustainability related to topics such as low impact development and carbon sequestration, and is active in the sustainability education community. Dr. Haselbach is a licensed professional engineer and a LEED AP (BD+C). Prior to her academic career she founded an engineering consulting company in the New York – Connecticut area. Her degrees include a BS in Civil and Environmental Engineering from Cornell, an MS in Chemical Engineering from UC Berkeley, and a PhD in Environmental Engineering from the University of Connecticut. She is currently an Associate Professor in Civil and Environmental Engineering at Washington State University, an Associate Director of the USDOT Tier 1 UTC: Center for Environmentally Sustainable Transportation in Cold Climates (CESTiCC) and a Fulbright-ALCOA Distinguished Chair in the Environmental Sciences and Engineering.

Rubric-Based Energy Literacy Assessment of Student Posters: Effects of Extended Calibration and Addition of Raters

Abstract

Energy literacy encompasses knowledge of energy principles in technical, social, and economic realms, as well as the ability to critically apply that knowledge to solve problems and form opinions. Collective advancement of energy literacy among the general population is thought to be instrumental in implementing sustainable energy solutions in the near future. As efforts to improve energy literacy have advanced, so has the need to assess the outcomes of those efforts. This paper describes advancements in a recently developed approach of examining energy literacy in student projects through application of a rubric, and the results of a case study using the methodology on the Imagine Tomorrow high school energy competition. Changes made to the approach include a more detailed rater calibration session and a significant increase in the number of raters over a previous cases study which used the same rubric. Similar to the previous study, results show that raters exhibit moderate to substantial agreement when interrater reliability is measured by Kendall's coefficient of concordance. As a component of this paper, group-wise comparisons of raters (pairs, triplets, and quadruplets) are examined to see if conclusions might have been different with different subsets of raters, both in terms of agreement statistics and in terms of energy literacy characteristics exhibited by various discrete groupings of students. No subset of raters would have resulted in significantly different conclusions in terms of scoring trends, though reliability statistics would be slightly altered. With respect to the competition energy literacy characteristics, it was found that posters created by students participating in more techno-centric challenges, with competition experience, or when mentored by returning advisors scored slightly higher than others. The energy literacy observed was unaffected by gender of the students or the teaching subject of advisors. Continual assessment, and improvement of assessment instruments, is vital as project-based learning continues to be a focal point for teaching about energy, and as organizers plan how to best shape future events to improve energy literacy of our current and future decision-makers.

Introduction

Energy literacy encompasses knowledge of energy principles in technical, social, and economic realms, as well as the ability to critically apply that knowledge to solve problems and form opinions. Collective advancement of energy literacy among the general population is thought to be instrumental in implementing sustainable energy solutions in the near future. Langfitt et al.[1,2] have previously developed a rubric-based scoring system for rating energy literacy displayed in the deliverables of a competition or course. The previous research showed some significant trends between energy literacy levels and factors about the participating students, mentors, and competition challenge entered (entry categories included behavior, biofuels, design, and technology). Rater agreement had been moderate as analyzed at an overview level in those works[1,2] and more thoroughly in a later paper[3]. Gotch et al.[3] outlined potential improvements to the rubric application process in order to increase interrater reliability, which included a calibration session of increased immersion and addition of more raters. The research presented in this paper was conducted by following these recommendations and interrater reliability was again tested to analyze if these changes caused improvements over previous levels of reliability.

Energy Literacy

Humans harness energy for many essential tasks such as transporting goods and people, heating and lighting homes, and growing crops. The forms of energy we use frequently have impacts on human health and the environment, may be in limited supply in the foreseeable future, can threaten national security, and affect people economically[4]. Various sources of energy vary in the degree of these impacts (including tradeoffs in many cases), making choices about what energy sources to use and support through policy important. Additionally, developing and utilizing energy efficient technology and making behavioral changes to reduce overall energy consumption can have a positive effect. These choices are best made by an informed, or *energy literate*, community[5-7]. Indeed, energy knowledge has been shown to affect decisions related to energy use[8,9] and support for various energy sources[10,11].

The original sole definition of "literacy" was being able to read and write. In time it also became associated with being highly educated[12] and now many dictionaries include competence or knowledge about a subject as one definition[13-15]. However, in the scholarly literature on literacy, many regard the term to be a more complex, multi-dimensional concept, where possessing specific knowledge is only one portion of the greater meaning[16]. Literacy is often defined as also recognizing the need to ask questions and the ability to find, process, and communicate information. Due to the potential breadth of the term there is significant inconsistency in definitions of literacy even within the same field[12,17,18] indicating that a term like "energy literacy" may have different conceptualizations by different organizations and authors. Therefore, defining energy literacy in any study assessing it is important.

DeWaters et al.[19] devoted significant effort to defining energy literacy. Using a detailed literature review they examined various other literacy fields including cultural literacy, environmental literacy, and technological literacy, with the latter two informing much of their final conceptualization of energy literacy. They found that literacy is usually defined as more than possessing certain knowledge; it also usually includes analyzing and assessing information for use in solving problems and communication, and applying learned skills and knowledge to benefit society. Specific tenants identified included understanding basic energy concepts, energy's impacts on society and the environment, the need for conservation, and personal energy use. Finally, it also included demonstration of these through one's beliefs and personal choices. The US Department of Energy (DOE)[20] has also defined energy literacy stating that, "energy literacy is an understanding of the nature and role of energy in the universe and in our lives…[and] is also the ability to apply this understanding to answer questions and solve problems." In this paper and rubric approach, energy literacy is defined similarly as the ability to recognize energy issues, understand basic energy concepts, find relevant energy information, and use that information to develop appropriate solutions that consider relevant stakeholders. However, behavioral effects are not part of the definition or assessed criteria because this rubric scoring approach is only intended to assess the problem-solving aspects of energy literacy.

Assessment of Energy Literacy

Assessment of energy literacy can be accomplished in a number of ways. The most common approach is through tests or questionnaires, which have generally found low levels of energy literacy in both children[7,21-23] and adults[24-27]. In this setting, respondents are typically asked to respond to multiple choice questions about basic energy principles, sources of energy currently used, energy consumption of various sectors, and more. Some of these also include energy knowledge self-assessment questions for a comparison of professed and actual understanding [24,26,27]. Another method for energy literacy assessment is analyzing behavioral data. For example, Brounen et al.[28] defined energy literacy through awareness of personal energy bill data and attempted to correlate that to behaviors. Finally, energy literacy can be measured by evaluating written works for indicators of energy literacy, such as through a scoring rubric[1,2,29]. All three approaches have various advantages and drawbacks and are applicable in different settings and within different definitions of energy literacy. The rubric-based approach, which is used in this paper, is most applicable when assessment of applied skills is desired, such as with a project or report on energy. This type of rubric may also be a more feasible option when access is not available directly to subjects and/or time does not permit other types of energy literacy testing.

With reference to the previously discussed definitions of energy literacy, this rubric-based approach may also be better aligned with the latter half of the Department of Energy[20] definition than a testing approach. That is, it more explicitly evaluates applied and investigative forms of literacy than many questionnaire-type assessments. This is because the subject is given the opportunity to demonstrate knowledge, prove that they recognize a need to find additional information, find and use that information, form arguments, and communicate results. The energy literacy rubric in this paper does address many of these points and, therefore, lends validity to the rubric approach as based on the Department of Energy[20] definition. This type of energy literacy applied to a deliverable and examined with a rubric should be differentiated from energy literacy knowledge examined through most of the testing procedures previously identified since few of these allowed students to directly demonstrate that they have the capability to find, interpret, use, and communicate energy information from external sources.

Imagine Tomorrow Competition

Imagine Tomorrow is a “problem-solving competition” for energy issues. These issues cover a wide range of topics including development of new alternative energy technologies, eco-design of buildings and parks, proposals and evaluations of behavioral campaigns to save energy, and implementations of biofuel use, just to name a few. The annual competition attracts high school students from Washington, Oregon, Idaho, and Montana, and is held at Washington State University in Pullman, WA. Students compete in teams of 3-5, enter themselves into one of four challenges based on the topic of their project (challenges include Biofuels, Behavior, Design, and Technology), and are guided by a mentor who is usually a teacher from the students’ school. Final deliverables are an electronically submitted abstract of each team’s work and a poster which students present to judges from academia, industry, and the community at the competition.

Objectives

Previous applications of this rubric-based assessment have appeared successful in understanding the efficacy of the Imagine Tomorrow program in promoting energy literacy amongst various demographic and other groups. Use of the rubric appeared appropriate in analyzing energy literacy based on similarities in scoring trends between raters on evaluations of both abstracts and posters in previous Imagine Tomorrow competitions. These have included similarities such as returning students outperforming new students, gender neutrality, students in middle grades outperforming the lowest grades, and teams competing in more technically-based challenges outperforming those in less technically-based[1,2,29]. Still, rater reliability as evaluated in those studies and by Gotch et al.[3] might be improved. This was especially true with respect to project abstracts, which are short and often written prior to completing the project, making them much less informative than the posters. Therefore, it was concluded that the approach is better suited to evaluating posters or other more detailed works[2,3], and this paper only assesses posters. The objective of this paper is to examine how changes to the rubric application process may have affected interrater reliability on posters, whether or not scoring trends would be altered by the number and subsets of raters, and to make a preliminary recommendation on the number of raters that should be used for future assessments. Specifically, the changes made included (1) a poster-based calibration session of considerably more detail and length than previous studies and (2) use of more raters on the posters (from two raters to five raters).

Methodology

Rubric

The energy literacy rubric used in this assessment (Table 1) was developed by Langfitt and Haselbach[29] and previously examined by Langfitt et al.[2]. A major intent of this subsequent research was to examine how changes in rater factors, including the number of raters and rater preparation, influenced interrater reliability. Therefore, the rubric is identical to the previous version.

Table 1: Energy Literacy Rubric[29]

	Points			
Topic	*0*	*1*	*3*	*5*
Issue	Not addressed	Identify the issue	Frame the issue	Professionally frame the issue
Solution	Not addressed	Identify solution to the issue	Discuss a solution	Develop appropriate solution
Impacts	Not addressed	Identify broader Impacts	Discuss broader impacts	Examine broader impacts
Stakeholders	Not addressed	Identify stakeholders	Consider stakeholder perspectives	Understand and address stakeholder perspectives
Technical Concepts	Not addressed	Identify technical concepts	Discuss technical concepts	Examine technical concepts as they relate to the project
Outside Information	Not addressed	Identify basic info from outside sources or that this information exists	Discuss information from outside sources	Examine information as it relates to the project

The rubric for energy literacy used in this assessment was initially developed[29] based on a rubric used for evaluating student projects in civil and environmental engineering at Washington State University and from a rubric on sustainability writing[30], in the style of a holistic rubric. Subsequent alterations included separating criteria into six dimensions (*Issue, Solution, Impacts,* *Stakeholders, Technical Concepts, and Outside Information)* and explicitly relating those dimensions to sub-principles of energy literacy from the DOE[20]. Content validity comes from both the basis of sustainability criteria from Timmerman et al.[30], and from the relation to the principles of energy literacy. For more detailed information on rubric development and validity, please refer to Langfitt and Haselbach[29].

In this research, raters were required to award posters a score of 0, 1, 3, or 5 in each rubric dimension based on the criteria listed. Overall scores for each poster were then developed from the dimension scores using the scheme described in Table 2.

Table 2: Integer and Word Score Conversion Scheme

Sum of Dimension Scores	*Integer Number Score*	*Word Score*
0	0	Absent
1-6	1	Emerging
7-12	2	Developing
13-18	3	Competent
19-24	4	Effective
25-30	5	Mastering

Raters

This study utilized five raters, who are individually referred to as Raters 1 through 5. All five raters were students in civil and environmental engineering at Washington State University. Two raters were PhD students, two were upper level undergraduates, and one was a Masters' student. Two raters had prior experience applying the rubric to posters. All five raters scored every poster based on the same set of artifacts.

Reliability

In rubric assessments, reliability refers to the degree to which multiple raters agree. In much of the literature, reliability is used as a general term to describe both consensus and consistency measures[31-33]. Consensus is when raters agree on the scores that projects should receive, including the magnitude of those scores. Consistency is less restrictive and only requires raters to agree to on the trend in scores, or more specifically, the ranking of works[32]. Therefore, consistency allows raters to have different benchmarks for translating performance into scores, but still captures whether raters agree on which works deserve higher scores than others. This type of reliability is the focus of this study. (It should be noted that in the literature, reliability is usually used as a general term to describe both or either consensus or consistency measures, but sometimes "reliability" is used exclusively for consistency and "agreement" exclusively for consensus.)

Consistency can be measured in a number of ways. Common methods for generating a single number reliability score are Pearson's r, Spearman's rho, and Chronbach's alpha[32]. Previous versions of this rubric approach have utilized Spearman's rho[34] and a variation on it for more than two raters - Kendall's coefficient of concordance (KCC)[35]. This study again utilized Spearman's rho (for rater pairs) and KCC (for groups of 3 or more raters) in order to make reliability scores comparable to previous studies.

Both approaches rely on evaluating the degree of agreement between raters on poster rankings. Rankings are ordinal numbers determined by ordering the total scores from lowest to highest, starting at "1" for the lowest, for each individual rater. In the case of ties for any scores from the same rater, all these scores are assigned the same averaged rank in the ordering (averaged ranks can include fractional values). Specifically, Spearman's rho is calculated by[36]

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{1}$$

where,
d_i is the difference in rank between the two raters for poster i,
n is the total number of posters evaluated, and
ρ is Spearman's rho.

Kendall's coefficient of concordance is calculated by[36]

$$KCC = \frac{12\sum_{i=1}^{n}(r_i - \bar{r})^2}{m^2(n^3-n) - m\sum_{j=1}^{k}(t_j^3 - t_j)} \tag{2}$$

where,
r_i is the sum of all raters' rankings of poster i,
$\bar{r}$ is the mean of all r_i,
m is the number of raters,
n is the total number of posters evaluated,
k is the number of tied ranking series,
t_j is the count of posters within tied ranking series j, and
KCC is the Kendall's coefficient of concordance.

For both measures, values closer to 1 are higher correlation and values closer to 0 are lower correlation. One interpretation of intermediate values in this type of measure[37], is to subdivide the scale into slight correlation (0-0.2), fair correlation (0.2-0.4), moderate correlation (0.4-0.6), substantial correlation (0.6-0.8), and nearly perfect correlation (0.8-1). As a note of caution, these are simply general descriptors and should not be considered valid for every application.

A major intent of this research was to examine how the number of raters selected could have changed any outcomes in terms of reliability of the rubric application and the scoring trends identified. That is, would the results of the assessment have varied if any smaller subset of the raters had been used instead of all five? Therefore, the Spearman's rho or KCC was calculated and reported for every combination of two, three, four, and five raters. Additionally, another major objective was to assess whether or not an extended calibration session increased interrater reliability. Because the previous study[2] only used two raters, this was done on the basis of comparing the two rater Spearman's rho scores with that obtained previously. Pairs of raters are directly identified in the results section, however, triplets and quadruplets are identified by lettered groups to make presentation more streamlined. Table 3 shows the raters in each lettered group, by placing an “x” in the row of each rater included in that column’s group. For example, Group B consisted of Rater 1, Rater 2, and Rater 4.

Table 3: Lettered Groupings of Raters

	Triplet Groupings										*Quadruplet Groupings*				
Rater	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*	*M*	*N*	*O*
Rater 1	x	x	x	x	x	x					x	x	x	x	
Rater 2	x	x	x				x	x	x		x	x	x		x
Rater 3	x			x	x		x	x		x	x	x		x	x
Rater 4		x		x		x	x		x	x	x		x	x	x
Rater 5			x		x	x		x	x	x		x	x	x	x

Calibration

Calibration of raters is a technique that can increase interrater reliability[38]. Raters should iteratively score and discuss scoring differences to identify and resolve inconsistencies[39]. Specifically, the main activity in rubric calibration usually consists of all raters coming together to score a small sample of works and discuss those ratings to resolve significant differences.

A calibration session was carried out prior to rating the posters. All five raters were present for this activity which lasted about two and a half hours. The basic structure and procedures used approximately followed recommendations from the VALUE rubrics manual[38] and from a frame of reference training outline[40]. The general framework was based mostly on Graham et al.[40], as were most of the details on rater variability and biases. Rhodes and Finley[38] provided useful suggestions to facilitate discussion on both the rubric and on the sample works scores. The calibration session contained the following steps:

1: Background knowledge
2: Introduction to the rubric
3: Introduction to rater variability and biases
4: Overview of the rating process
5: Example rating
6: Rating practice

Calibration began by attaining background knowledge helpful to applying the rubric. The session leader briefly introduced the nature of the research, its importance, and the intended use of the ratings. Next, the group read through the energy literacy principles[20] aloud to ensure everyone understood the concept bases of energy literacy. This document was distributed well before the session to allow raters to preview it. The session leader interjected with important points of discussion such as clarification of some sub-principles and comments on how students have addressed some of these sub-principles in past projects. Only minor discussion occurred during this portion of the calibration, likely because all of the raters have an engineering background, however, a more diverse group of raters might require a more lengthy discussion.

Following this reading, the focus turned towards the rubric. All raters were issued a copy of the energy literacy rubric and a copy of a matrix which mapped the rubric dimensions to energy literacy guide sub-principles[29]. It was explained that the mapping document shows one interpretation of how the energy literacy principles may be used to guide rating with the rubric. However, the raters were also instructed that this was meant as a general guide and their own judgment should be used to categorize energy ideas outside those explicitly covered in the guide. Then the rubric itself was covered in more detail. Raters were instructed to score each rubric dimension with a 0, 1, 3, or 5 according to the rubric criteria. These criteria were introduced by the session leader through the use of a simple example (a fictional scenario where a group proposes installing solar panels on the roof of their school) to convey roughly what was meant by various terms such as "identify", "discuss", and "examine."

Biases and rater errors can impact scoring and lead to greater diversion in rater reliability. Consequently, common rater biases and errors[40] were discussed in an attempt to help the raters avoid them. Raters were instructed to avoid biases from personal connections (e.g. a project is focused on an issue that the rater is passionate about, so it receives a higher score) and from personal beliefs (e.g. the rater disagrees with the premise of the project, so it receives a low score). Rater errors that the raters were instructed to avoid included leniency (always giving the benefit of the doubt), halo (allowing ratings on one dimension to influence that on another), compensation (awarding a higher score on one dimension due to a near miss on another), central tendency (all ratings right around the middle), initial impression (judging by the first few things seen, rather than the poster as a whole), and gut feeling (using subjective criteria rather than the explicit criteria of the rubric). Additionally, raters were instructed not to consider factors such as appearance, language proficiency, and organization in the scores.

To better align raters, the next point of discussion was the general process that raters were encouraged to use. The entire poster should be read, rather than simply skimmed, to ensure that the rater could cohesively understand the poster. Rubric dimension scores could be developed while reading or after, however, if developed while reading, raters were to re-evaluate scores given in the context of the entire poster upon completion. Raters were instructed to keep the rubric at hand and consider the criteria specifically when judging each poster. Points should be awarded based on what is explicitly contained in the poster, rather than assuming what other knowledge they might possess. Finally, raters were instructed to give more weight to inquiry than to technical accuracy. That is, the focus is on increasing familiarity with energy literacy by the students examining energy concepts rather than deep and exact technical knowledge. A simple example of this would be a high score for the *technical* dimension for a group proposing a perpetual motion machine (considered impossible under the current theories of physics), but who still identify, discuss, and examine important concepts of energy and motion.

Transitioning into rating, the session leader covered five posters from the 2014 Imagine Tomorrow competition, by showing the posters to the raters and explaining what scores might be given and why. These posters were chosen to reflect a range of quality. The session leader had rated two sets of posters in the past and was one of the core developers of the methodology, lending some credibility to his interpretation. Still, this was only intended as a starting point to give the new raters a general feel for using the rubric; changes in interpretation remained open for the final activity – independent rating and subsequent group discussion of sample posters.

Finally, scoring activities took place in order to practice applying the rubric to sample works. These sample works came from the previous year's competition and were chosen by the session leader to represent a range of the quality in energy literacy characteristics. This took place in two rounds which followed the same procedure. First three posters were scored independently by the raters and discussed by the group, then three additional posters were rated followed by discussion. In the discussions the scores were compared for each poster and rubric dimension individually, with raters briefly providing a verbal reasoning for giving the scores. For scores that were close (i.e. only one level apart), typically no further discussion took place, particularly if it was realized that differences were mostly due to overall higher or lower scores given by certain raters. (This is because the goal was consistency, not consensus, and these characteristics do not signal poor consistency.) Larger discrepancies were discussed in more detail because it was thought that these were due to more substantive issues in information classification and rubric application, rather than in interpretation of the extent of meeting a criterion (e.g. identify vs examine). There were relatively few of these larger differences. Two examples of issues discussed and agreed upon included (1) whether to award *technical* points for technical information not directly related to energy (decision: no) and (2) whether to award *solution* points for solution-type information coming directly from outside sources (decision: points for *information,* but only 1 point for *solution* unless further developed and/or discussed by the students). Both rating exercises appeared to produce relatively consistent results and the raters were dismissed. Further ratings would have taken place if more significant deviations had occurred.

Data Collection

The energy literacy rubric was applied to all 113 posters from the 2015 Imagine Tomorrow competition in Pullman, WA. Posters were rated based on photographs taken on the day of the competition. These photographs did not include any additional materials such as brochures or prototypes, except in the cases where the main deliverable was clearly of this nature. To mitigate any possible bias these pictures also did not include any students. Data about the teams, advisors, schools, and entry challenges were obtained in a spreadsheet from the Imagine Tomorrow event organizers and subsequently cross referenced with the poster scores. Trends in energy literacy were analyzed for the following variables (classifications within that variable in parentheses):

- *Gender (Male, Female)*
- *Repeat Student Participants (New Student, Repeat Student)*
- *Repeat Advisor Participation (New Advisor, Repeat Advisor)*
- *Challenge (Behavior, Biofuels, Design, Technology)*
- *Student Grade Level (9, 10, 11, 12)*
- *Project Setting (Class, Extracurricular)*
- *Advisor Teaching Subject (STEM, Non-STEM)**

*STEM is an acronym for Science, Technology, Engineering, and Math.

Results

Reliability

Consistency evaluation was carried out using Kendall’s coefficient of concordance on integer number scores. The result for all five raters was 0.606. This indicates moderate to substantial agreement, but is lower than the value of Spearman’s rho for agreement between the two raters evaluating posters in the previous assessment[2] (0.818). The direct comparison from year to year between these same two raters also revealed a lower agreement (0.699 versus 0.818) despite the longer calibration, suggesting that the calibration may not have had the desired effect of raising interrater agreement. However, it is noteworthy that raters 1 and 4 are the same raters who took part in the previous assessment and these raters showed the highest interrater reliability of all possible pairings. Reliability measured using Spearman’s rho for all pairs of raters and KCC for all triplets and quadruplets of raters are in Tables 4 through 6. For rater pairs, Table 4 is organized in matrix form, where the Spearman’s rho of any pair of raters is the value in the table for the first rater (row) and the second rater (column). Tables 5 and 6 show the Kendall’s coefficients of concordance by the groupings of three and four raters previously defined.

Table 4: Pairwise Comparisons of Raters for Spearman's Rho

		Second Rater			
		2	*3*	*4*	*5*
First Rater	*1*	0.501	0.399	0.699	0.558
	2		0.366	0.599	0.554
	3			0.465	0.311
	4				0.627

Table 5: Triplet Groupings of Raters Evaluated With KCC

Group	A	B	C	D	E	F	G	H	I	J
KCC	0.550	0.698	0.666	0.551	0.551	0.718	0.582	0.547	0.699	0.576

Note: Refer to Table 3 for raters included in each lettered group.

Table 6: Quadruplet Groupings of Raters Evaluated With KCC

Group	K	L	M	N	O
KCC	0.566	0.530	0.659	0.570	0.556

Note: Refer to Table 3 for raters included in each lettered group.

Based on average overall reliability scores in Table 7, the three rater case is very similar to the five rater case. Standard deviations of the KCC from groups of three or more raters are sufficiently small that it is unlikely that varying the raters within each group size would show significantly more agreement. This is further supported by examining the reliability scores in Table 5, where KCC ranged from 0.55 to 0.72.

Table 7: Summary of Reliability Scores by Rater Group Size

Rater Group Size	*Average KCC*	*Standard Deviation of KCC*
2	0.487*	0.109*
3	0.614	0.072
4	0.576	0.049
5	0.606	N/A

*Note that these are Spearman's rho

Scoring Trends

Scoring trends related to competition, student, and advisor variables were analyzed in detail based on all five raters' scores by Langfitt and Haselbach[41]. Variables which seemed to have a noticeable impact on energy literacy included challenge, project setting, repeat advisor, and grade level of students. Variables which seemed to have little or no impact on energy literacy displayed included gender, advisor teaching subject, and repeat student participation. The following discussion of scoring trends examines whether or not trends may have been interpreted differently if a smaller subset of the five raters were used instead of the full group.

Variables with Noticeable Impact on Energy Literacy

Table 8 shows the ranking by size of rater group of each classification within a variable that had a noticeable trend. For example, in challenge, under 2 raters, every pairing of raters gave the lowest average score to behavior projects, the highest average scores to biofuels projects, and different pairs of raters varied in their ordering of design and technology between the rankings of 2 and 3. Based on this table, it appears that having fewer raters would not significantly impact the trends identified, suggesting that fewer raters could be used. In light of the previous analysis of reliability by rater group size and the resolution of discrepancy in student grade level ranking, three raters appears to be a sufficient number. As an illustrative example, the average score by rater grouping is shown in the following for the challenge variable with pairs, triplets, quadruplets, and all five raters in Figures 1 through 3.

Table 8: Average Score Rankings within Four Variables under Each Rater Size Grouping

Number of raters	***Challenge***				***Student grade level***				***Setting***		***Repeat advisor***	
	Beh	**Bio**	**Des**	**Tech**	**9**	**10**	**11**	**12**	**Class**	**Extra**	**Repeat**	**New**
2	4	1	2/3	2/3	3/4	2	3/4	1	1	2	1	2
3	4	1	2/3	2/3	4	2	3	1	1	2	1	2
4	4	1	2/3	2/3	4	2	3	1	1	2	1	2
5	4	1	2	3	4	2	3	1	1	2	1	2

Note: Beh=Behavior, Bio=Biofuels, Des=Design, Tech=Technology, Extra=Extracurricular.

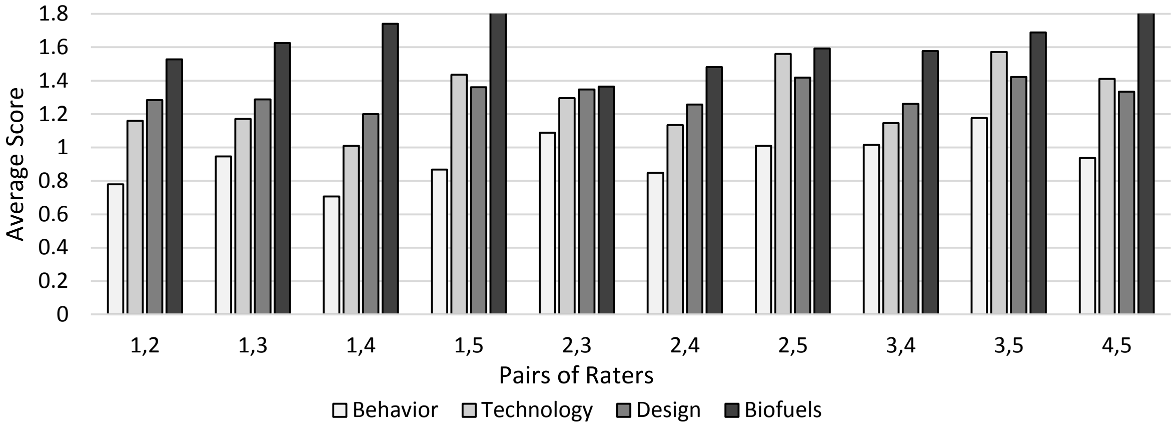

Figure 1: Average Score by Challenge for Every Pair of Raters

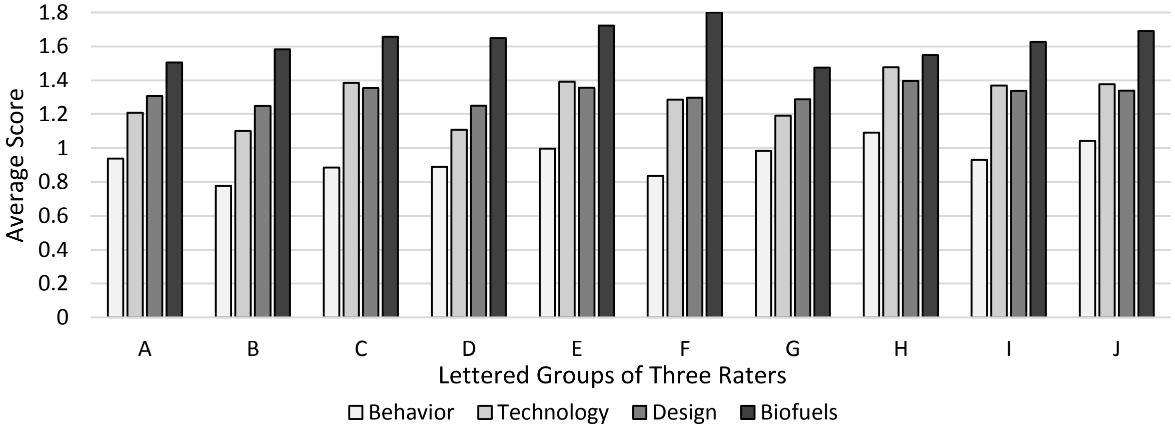

Figure 2: Average Score by Challenge for Every Triplet of Raters

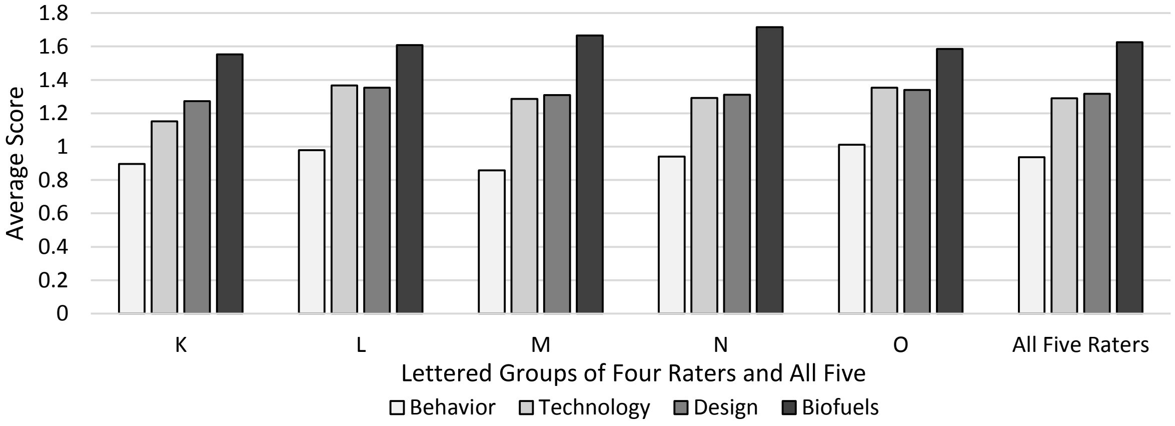

Figure 3: Average Score by Challenge for Every Quadruplet of Raters and All Five Raters

Variables with Little to No Impact on Energy Literacy

Variables with minor to no impact were those with only slight differences between classifications (gender, advisor teaching subject, and repeat student participation). This was evident within all the subsets of raters. Sometimes the rankings were switched from rater group to rater group, but with very little difference in the actual scores. Two highlights of this is that energy literacy appeared fairly equivalent for either gender participating in the competition and between STEM and non-STEM advisor teaching subjects. The latter suggests that all advisors can contribute positively to this knowledge effort.

Conclusion

Five raters were used to assess all of the posters in the Imagine Tomorrow competition for energy literacy characteristics. However, a smaller subset of raters would likely have resulted in similar reliability as measured by Kendall's coefficient of concordance and shown similar variable-linked trends. Specifically, the use of three raters appears to be sufficient as every grouping of three raters produced reasonably reliable scoring and the same scoring trends as the entire set of five raters. Pair groupings showed more non-concordance, however, would probably be useful if no more raters could be used. These conclusions were based purely on the reliability and trending observed in this study which used only civil and environmental engineering students as the raters. Therefore, it is possible that a single rater in other evaluations could have very different interpretations of the rubric and alter the efficacy of a three rater evaluation.

A higher number of raters in this evaluation would have allowed for a larger set of two and three rater groupings, leading to more generalizable results. Additionally, carrying out the analyses with people of more varied backgrounds (all raters were engineering students) would more fully accommodate the potential effects of a wider array of people. Similarly, there were two returning raters and three new raters. It is possible that the returning raters may have had biases and predispositions associated with how they interpreted and applied the rubric in previous assessments. However, it would be expected that if this rubric were used for ongoing assessment elsewhere, some raters would be new to the rubric and others have experience with it. Therefore, it is paramount when any new interpretations of the rubric become the standard at a calibration session, that returning raters are instructed to follow the new interpretation.

The somewhat lower reliability scores by pairs of raters in this study (when compared to the previous study[2]) indicate the expected improvement of holding a more substantial calibration session may not have been successful in raising scoring agreement. However, do note that the first pair of raters in the earlier case study may have been especially aligned in their interpretation of the rubric

Finally, it is noted that the trends in energy literacy were very consistent based on the variables of the competition in most of the rater groupings, especially with 3 or more raters. These trends continue to support some important conclusions from the earlier studies that the competition is gender neutral with respect to energy literacy and advisors from any field can be effective mentors. As expected, more technological focused submittals appear to display more energy literacy and repeat advisors may help in promoting more knowledge.

Acknowledgements

The authors would like to thank the Center for Environmentally Sustainable Transportation in Cold Climates (CESTiCC), the Ecoworks Foundation, Bank of America, Boeing, and the many other sponsors of the Imagine Tomorrow Competition for their generous support. The authors gratefully acknowledge the Northwest Advanced Renewables Alliance supported by the Agriculture and Food Research Initiative Competitive Grant no. 2011-68005-30416 from the USDA National Institute of Food and Agriculture, for funding part of this work. Appreciation is also for other members of the Imagine Tomorrow Competition steering committee. Additionally, the authors thank Brandon Werner, Mina Yekkalar, Benjamin Nantasai, and Trace Sendele for their contributions.

Bibliography

1. Langfitt, Q., Haselbach, L., and Hougham, R. J. (2015). "Artifact-Based Energy Literacy Assessment Utilizing Rubric Scoring." *Journal of Professional Issues in Engineering Education and Practice*, 141(2), C5014002.
2. Langfitt, Q., Haselbach, L., and Hougham, R.J. (2015). "Refinement of an Energy Literacy Rubric for Artifact Assessment and Application to the Imagine Tomorrow High School Energy Competition." *Journal of Sustainability Education*, 8.
3. Gotch, C. M., Langfitt, Q., French, B. F., and Haselbach, L. (2015). "Determining Reliability Scores from an Energy Literacy Rubric." *Proceedings of 122nd ASEE Annual Conference & Exposition*, Seattle, WA.
4. Asif, M., and Muneer, T. (2007). "Energy supply, its demand and security issues for developed and emerging economies." *Renewable and Sustainable Energy Reviews*, 11(7), 1388–1413.
5. Turcotte, A., Moore, M. C., and Winter, J. (2012). *Energy Literacy in Canada*. School of Public Policy, University of Calgary.
6. US Department of Energy (DOE). (2011). "Strategic Plan." *DOE/CF-0067*.
7. DeWaters, J. E., and Powers, S. E. (2011). "Energy literacy of secondary students in New York State (USA): A measure of knowledge, affect, and behavior." *Energy Policy*, 39(3), 1699–1710.
8. Schwartz, T., Denef, S., Stevens, G., Ramirez, L., and Wulf, V. (2013). "Cultivating energy literacy: results from a longitudinal living lab study of a home energy management system." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 1193–1202.
9. Abrahamse, W., Steg, L., Vlek, C., and Rothengatter, T. (2005). "A review of intervention studies aimed at household energy conservation." *Journal of Environmental Psychology*, 25(3), 273–291.
10. Bang, H.-K., Ellinger, A. E., Hadjimarcou, J., and Traichal, P. A. (2000). "Consumer Concern, Knowledge, Belief, and Attitude toward Renewable Energy: An Application of the Reasoned Action Theory." *Psychology & Marketing*, 17(6), 449–468.
11. Hobman, E. V., and Ashworth, P. (2013). "Public support for energy sources and related technologies: The impact of simple information provision." *Energy Policy*, 63, 862–869.
12. Bybee, R. W. (1997). *Achieving scientific literacy: from purposes to practices*. Heinemann, Portsmouth, NH.
13. Merriam-Webster. (2015). "Literacy." *Merriam-Webster Online Dictionary*. <www.merriam-webster.com/dictionary/literacy> Accessed 25 May 2015.
14. Dictionary.com. (2015). "Literacy." <dictionary.reference.com/browse/literacy?s=t> Accessed 25 May 2015.
15. Oxford Dictionary (2015). "Literacy." <www.oxforddictionaries.com/us/definition/american_english/literacy> Accessed 25 May 2015.
16. Frisch, A-L., Camerini, L., Diviani, N., and Schulz, P. J. (2012). "Defining and measuring health literacy: how can we profit from other literacy domains?" *Health Promotion International*, 27(1), 117–126.
17. Berkman, N. D., Davis, T. C., and McCormack, L. (2010). "Health Literacy: What Is It?" *Journal of Health Communication*, 15(S2), 9–19.

18. Gormally, C., Brickman, P., and Lutz, M. (2012). “Developing a Test of Scientific Literacy Skills (TOSLS): Measuring Undergraduates’ Evaluation of Scientific Information and Arguments.” *Cell Biology Education*, 11(4), 364–377.
19. DeWaters, J., Powers, S., and Graham, M. (2007). “Developing an Energy Literacy Scale.” *Proceedings of the 114th Annual ASEE Conference & Exposition*, Honolulu, HI.
20. US Department of Energy (DOE). (2014). “Energy Literacy: Essential Principles and Fundamental Concepts for Energy Education Version 3.0.” *DOE/EE-1123*.
21. Barrow, L. H., and Morrisey, J. T. (1989). “Energy literacy of ninth grade students: a comparison between maine and new brunswick.” *Journal of Environmental Education*, 20(2), 22–25.
22. Gambro, J. S., and Switzky, H. N. (1999). “Variables Associated With American High School Students’ Knowledge of Environmental Issues Related to Energy and Pollution.” *Journal of Environmental Education*, 30(2).
23. Bodzin, A. (2012). “Investigating Urban Eighth-Grade Students’ Knowledge of Energy Resources.” *International Journal of Science Education*, 34(8), 1255–1275.
24. National Environmental Education & Training Foundation (NEETF). (2002). “Americans’ Low ‘Energy IQ’: A Risk to Our Energy Future.”
25. Bittle, S., Rochkind, J., and Ott, A. (2009). “The Energy Learning Curve.”
26. Southwell, B., Murphy, J., DeWaters, J. E., and LeBaron, P. (2012). “Americans’ Perceived and Actual Understanding of Energy.” *No. RR-0018-1208*, RTI Press, Research Triangle Park, NC.
27. Cotton, D., Winter, J., Miller, W., and Muneer, R. (2015). “Informal learning on campus: a comparative study of students’ energy literacy in UK universities.” *Education for Sustainable Development Pedagogy: Criticality, Creativity, and Collaboration*, 15.
28. Brounen, D., Kok, N., and Quigley, J. M. (2013). “Energy literacy, awareness, and conservation behavior of residential households.” *Energy Economics*, 38, 42–50.
29. Langfitt, Q., and Haselbach, L. (2014). “Imagine Tomorrow High School Energy Competition 2014: Energy Literacy and Biofuels Literacy Assessment of Abstracts and Posters.” *Report to the Imagine Tomorrow Steering Committee, Washington State University*.
30. Timmerman, B. E. C., Strickland, D. C., Johnson, R. L., and Payne, J. R. (2011). “Development of a ‘universal’ rubric for assessing undergraduates’ scientific reasoning skills using scientific writing.” *Assessment & Evaluation in Higher Education*, 36(5), 509–547.
31. Moskal, B. M., and Leydens, J. A. (2000). “Scoring Rubric Development: Validity and Reliability.” 7(10).
32. Stemler, S. (2004). “A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Stemler, Steven E.” *Practical Assessment, Research & Evaluation*, 9(4).
33. Oakleaf, M. (2009). “Using rubrics to assess information literacy: An examination of methodology and interrater reliability.” *Journal of the American Society for Information Science and Technology*, 60(5), 969–983.
34. Spearman, C. (1904). “The Proof and Measurement of Association between Two Things.” *The American Journal of Psychology*, 15(1), 72–101.
35. Kendall, M. G., and Smith, B. B. (1939). “The Problem of m Rankings.” *The Annals of Mathematical Statistics*, 10(3), 275–287.*36.*
36. Gwet, K.L. (2012). *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters.* Advanced Analytics, LLC, Gaithersburg, MD.
37. Landis, J.R., and Koch, G. G. (1977). “The measurement of observer agreement for categorical data.” *Biometrics*, 33(1), 159–174.
38. Rhodes, T. L., and Finley, A. P. (2013). “Rubric calibration.” *Using the VALUE rubrics for improvement of learning and authentic assessment*, Association of American Colleges and Universities, Washington, DC, 23–25.
39. Maki, P. (2010). *Assessing for learning: building a sustainable commitment across the institution*. Stylus Pub, Sterling, VA.
40. Graham, M., Milanowski, A., and Miller, J. (2012). “Measuring and Promoting Inter-Rater Agreement of Teacher and Principal Performance Ratings.” *Center for Educator Compensation Reform.*
41. Langfitt, Q., and Haselbach, L. (2015). “Imagine Tomorrow High School Energy Competition 2015 Energy Literacy Assessment of Posters.” *Report to the Imagine Tomorrow Steering Committee, Washington State University*.