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

New Orleans JAZZED about Engineering Education

Paper ID #15242

Performance Assessment in Elementary Engineering: Evaluating Student (RTP)

Dr. Cathy P. Lachapelle, Museum of Science

Cathy Lachapelle leads the EiE team responsible for assessment and evaluation of our curricula. This includes the design and field-testing of assessment instruments and research on how children use EiE materials. Cathy is particularly interested in how collaborative interaction and scaffolded experiences with disciplinary practices help children learn science, math, and engineering. Her work on other STEM education research projects includes the national Women's Experiences in College Engineering (WECE) study. Cathy received her S.B. in cognitive science from the Massachusetts Institute of Technology and her Ph.D. in educational psychology from Stanford University.

Dr. Christine M. Cunningham, Museum of Science

Dr. Christine Cunningham is an educational researcher who works to make engineering and science more relevant, accessible, and understandable, especially for underserved and underrepresented populations. A vice president at the Museum of Science, Boston since 2003, she founded and directs Engineering is Elementary^{TM}, a groundbreaking project that integrates engineering concepts into elementary curriculum and teacher professional development. As of September 2014, EiE has served 6.2 million children nationwide and 71,000 educators. Cunningham has previously served as director of engineering education research at the Tufts University Center for Engineering Educational Outreach, where her work focused on integrating engineering with science, technology, and math in professional development for K-12 teachers. She also directed the Women's Experiences in College Engineering (WECE) project, the first national, longitudinal, large-scale study of the factors that support young women pursuing engineering degrees. Cunningham is a Fellow of the American Society for Engineering Education and was awarded the 2014 International Society for Design and Development in Education Prize. She holds B.A. and M.A. degrees in biology from Yale and a Ph.D. in Science Education from Cornell University.

Performance Assessment in Elementary Engineering: Evaluating Student Work (RTP)

Abstract

With the new emphasis on engineering practices and engineering design (NGSS Lead States, 2013), teachers of and researchers studying K-12 engineering need to find ways to measure students' developing engineering skills. To efficiently measure student learning of engineering practices, there is need for a tool to capture student performances in a way that readily affords evaluation. The problem we pursue in this paper is how to accomplish this measurement. In this paper, we present a performance assessment instrument and coding rubric. We calculate inter-rater reliability for coders and present descriptive statistics for student scores to demonstrate the utility of the instrument for distinguishing a range of performances. We conduct an exploratory factor analysis to examine the internal structure of the unit and calculate internal consistency reliability. To build a case for validity for use of the assessment to measure student learning of engineering practices, we compare video of 30 students working on design challenges in their student groups, collected from 10 of the participating classrooms, to the same students' performance on the assessment. This also informs the use and limits of utility of the written performance assessment for measuring elementary students' engineering skills and understanding-in-use. Finally, we describe the time needed to score the assessments, and discuss its utility for larger-scale research studies.

Introduction

The Next Generation Science Standards[1] calls for all American students to learn engineering in addition to science in grades K-12. The NGSS places particular emphasis on students learning engineering practices and an understanding of engineering design. At the earliest grades, children learn about engineering as solving problems that people want solved. "Emphasis is on thinking through the needs or goals that need to be met, and which solutions best meet those needs and goals" [1 Appendix I]. Throughout elementary school, students' engineering work becomes more formal, as they define the criteria for success in solving a problem, define the constraints on design solutions, research a variety of possible solutions, and iterate towards an optimal solution. Engineering practices include defining problems, developing and using models, conducting investigations, analyzing data, designing solutions, arguing from evidence, and communicating information.

Teachers of K-12 students and researchers studying K-12 engineering need to find ways to measure students' developing engineering skills. To measure student learning of engineering practices efficiently, there is need for tools to capture student performances in a way that readily affords evaluation and scalability.

Purpose of the Study

We chose for this study to focus on the assessment of the engineering performance of elementary school students. The problem we pursue in this paper is how to accomplish the measurement of elementary students' skills with engineering practices, their understanding of engineering design, and their conception of engineering as a discipline. To address this need, we developed a performance assessment, which we have named the Elementary Engineering Performance Assessment (EEPA).

Our research questions are as follows:

- Can we evaluate individual elementary students' engineering skills with a quick pencil and paper design task?
- What are the characteristics of the EEPA when used with a sample of students from the target population?
- Can the instrument be efficiently and reliably coded using a rubric, so that researchers and teachers can make use of it?

Literature Review

Performance assessments are a form of contextual assessment where students engage in tasks within a context that affords the use of practices of interest to the assessor[2,3]. There are many advantages to performance assessment, including face validity, the emphasis on skills and the ability to deal with complexity, and relevance[2,3].

Performance assessment tasks should meet several criteria: they should elicit complex and observable performances, use a standard set of tasks, have high fidelity to "real life" performances, measure a variety of levels of performance, and afford improvement with practice[4].

In engineering in the K-12 setting in particular, there is need for assessment focusing on performance not content—extensive research into the issue of preparing high school students for college engineering found that what is most consistently called for is familiarity with the process of designing a product[5]. Elementary school engineering curricula also consistently focus on the process of design, engaging students in simple design tasks and familiarizing them with elements of the process[6–8]. The processes of design are also the primary focus of the NGSS[7].

An important challenge for the development of performance assessment is achieving reliability in determining the skill level of students (test-retest reliability)[2,4]. This is a frequent problem with performance assessments, because they are time-consuming to administer, which precludes administering a sampling of similar tasks; however, they also present opportunities for learning for both students and teachers, and as such may be worth an additional investment of time[3].

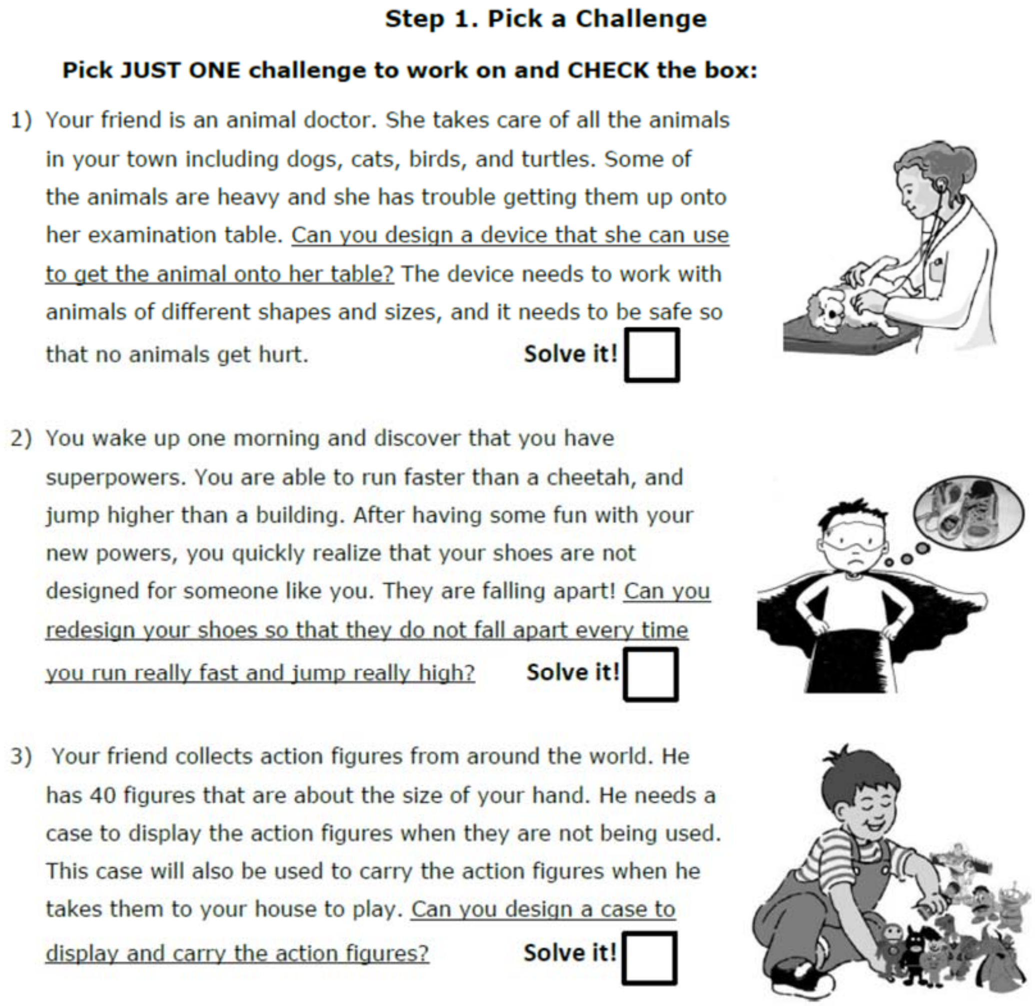
Method

Instrument Development

Our target population for assessment is elementary students. Therefore, we faced a number of constraints for instrument development. First, elementary school classes do not have consistent computer access, and requiring a computer assessment would have significantly impeded our data collection efforts, so we chose to implement the EEPA in a written format. Second, our target age range was 8 to 11 (grades 3 through 5), so we needed an assessment incorporating a third grade reading level, which could be completed within 30 minutes (a class period).

We chose to develop an instrument that would assess the learning objectives detailed in Table 1. Students are presented with three hypothetical design challenges to choose from. Each is presented in a single paragraph, to minimize reading time. Each describes a problem, a goal, and criteria for success (see Figure 1). Then, students are prompted to answer questions designed to elicit their understandings of the learning goals (see Table 1).

Figure 1: Challenge Scenarios



Step 1. Pick a Challenge

Pick JUST ONE challenge to work on and CHECK the box:

1) Your friend is an animal doctor. She takes care of all the animals in your town including dogs, cats, birds, and turtles. Some of the animals are heavy and she has trouble getting them up onto her examination table. Can you design a device that she can use to get the animal onto her table? The device needs to work with animals of different shapes and sizes, and it needs to be safe so that no animals get hurt. **Solve it!** □

2) You wake up one morning and discover that you have superpowers. You are able to run faster than a cheetah, and jump higher than a building. After having some fun with your new powers, you quickly realize that your shoes are not designed for someone like you. They are falling apart! Can you redesign your shoes so that they do not fall apart every time you run really fast and jump really high? **Solve it!** □

3) Your friend collects action figures from around the world. He has 40 figures that are about the size of your hand. He needs a case to display the action figures when they are not being used. This case will also be used to carry the action figures when he takes them to your house to play. Can you design a case to display and carry the action figures? **Solve it!** □

Table 1: Questions from the EEPA

Assessment Question	**Learning Objective**	**NGSS Connection**	**Coded Variable**
Imagine a Solution. THINK OF SOME IDEAS for how you might solve the problem in the challenge you chose.	Consider multiple solutions.	Practice: Constructing Explanations and Designing Solutions	Num_Ideas
Choose Your Best Design Idea. How will your BEST DESIGN IDEA solve the problem in the challenge?	Incorporate the given criteria within a design idea.	DCI: 3-5-ETS1.A	Meets_Criteria
	Make an argument for the proposed design idea that incorporates evidence.	Practice: Constructing Explanations and Designing Solutions	How_Will_Solve
Your Final Design. Take your BEST DESIGN IDEA and draw it here. Don't forget to label all of the parts and the materials that you would use!	Communicate a design idea using a combination of drawings and explanation.	Practice: Obtaining, Evaluating, and Communicating Information	Drawings
What is your design made out of? List the materials here.	Demonstrate understanding that matter exists as different substances with different properties, suited to different purposes.	DCI: K-2-PS1.A	Num_Materials
Tell Us More! What could you do next to make sure your design actually works?	Develop a plan to investigate whether the design idea would work.	Practice: Planning and Carrying Out Investigations	Do_Next
Do you think that the work you did for this activity is engineering? Why or why not?	Explain a variety of aspects of engineering.	Crosscutting Concept: Influence of Science, Engineering, and Technology on Society and the Natural World	Engineering?

Pilot Testing the Instrument

During development of the instrument, as part of our process of gathering evidence for validity, we conducted think-aloud protocols with 23 students in the target age range (7-11) who were participating in an engineering summer camp at a suburban school, to inform design and ensure that students were interpreting the instrument as intended. Fourteen of the students were male and 9 were female, and all students were white. Students were asked to read aloud and express their thinking aloud as they completed the assessments, while researchers watched, took notes, and prompted students to clarify the thinking that led to their responses.

Drawing on student and teacher feedback and student observations, we revised wording and revised the layout of the EEPA. We again tested using a think aloud protocol with another 12 students, ages 8 and 9, participating in an engineering summer program for academically at-risk students at an urban school. Seven of the students were male and five were female; 9 were Black and 3 were white.

Sample and Procedure for Coding Assessments

We collected assessments from grades 3-5 students in 274 classes from 129 schools. In each class, teachers had just finished implementing one or more engineering units. Schools were located in three states on the east coast, one in New England, one in the Mid-Atlantic region, and one in the South. Half the students were male and half female. Further demographics can be found in Table 2. Students completed the EEPA individually.

Table 2: Demographics of Participating Schools and Classes

Grade 3	**Grade 4**	**Grade 5**	**Urban**	**Suburban**	**Rural**	**Title 1**
28.3%	34.0%	37.6%	19.2%	42.9%	38.0%	68.8%*

*Title one status was not available for 6 classes from 2 private schools.

To characterize quality of performance on the EEPA, we developed a rubric (see Table 3) that focuses solely on aspects of the NGSS Engineering DCI and Practices that we planned to assess, as detailed above (Table 1). Three researchers each separately coded the same 130 assessments from 6 classrooms (a fully crossed design), then met to discuss the codes and come to consensus, in order to develop inter-rater reliability. Those 3 researchers then each independently coded another set of 161 assessments from 8 classrooms to check for inter-rater reliability (IRR). Inter-rater reliability was calculated using intra-class correlation (ICC), which is most appropriate for ordinal data [9] as in our rubric. We used a two-way absolute agreement type of model as appropriate for fully crossed design, and single-measures ICC, as we intend that the reliability of ratings made by 3 coders should be generalized to the full sample.

Table 3: Coding Rubric

Score:	**0**	**1**	**2**	**3**
Num_Ideas	None	1 Idea	2 ideas	3 or more ideas
Meets_Criteria	Blank	Meets none of the criteria stated in the problem	Meets 1 of the criteria stated in the problem	Meets 2 or more of the criteria stated in the problem
How_Will_Solve	Not Answered	No specific aspects of the design idea were explained or justified.	Contains one half of a linked explanation and justification: WHAT/HOW or WHY.	Explained and justified at least 1 specific WHAT/HOW and its WHY aspect of the design idea
Drawings	No Drawing	Drawing, 0 or 1 labels	Drawing with either labels or explanation.	Drawing with both labels AND explanation.
Num_Materials	None	Names or labels only objects (discrete materials)	1 or 2 materials that are NOT objects.	3 or more non-discrete materials are named or labeled.
Do_Next	Blank	Unrelated to design process. Does not show understanding that design may not work when put into use.	Names another step of the design process. Must show understanding that design may or may not work.	If they go beyond testing, or include specific ideas for how to test aspects of the design to see if they work.
Engineering?	Blank	No aspects of engineering.	One aspect of engineering.	Multiple aspects of engineering.

Once inter-rater reliability was established, we randomly selected 5 student assessments from each class to analyze, and divided that sample randomly by class into 3 parts. Each of the three coders independently coded one of the three parts. The final, coded sample was 1370 assessments from 274 classes. Coding took approximately 60 to 90 seconds per assessment. The demographic breakdown of this sample is presented in Table 4.

Table 4: Demographics of Students Sampled for Analysis

<table>
<tr><th rowspan="2">Gender:</th><th>Male</th><th colspan="2">Female</th><th>Missing</th><th>Total N</th></tr>
<tr><td>647</td><td colspan="2">650</td><td>3</td><td>1370</td></tr>
<tr><td colspan="6"></td></tr>
<tr><th rowspan="2">Racial/Ethnic Grouping:</th><th>Represented</th><th colspan="2">Underrepresented</th><th>Missing</th><th>Total N</th></tr>
<tr><td>809</td><td colspan="2">457</td><td>104</td><td>1370</td></tr>
<tr><td colspan="6"></td></tr>
<tr><th rowspan="2">Free or Reduced-Price Lunch:</th><th>Ineligible</th><th colspan="2">Eligible</th><th>Missing</th><th>Total N</th></tr>
<tr><td>386</td><td colspan="2">281</td><td>703</td><td>1370</td></tr>
<tr><td colspan="6"></td></tr>
<tr><th rowspan="2">From a Title 1 School:</th><th>Not Title 1</th><th colspan="2">Title 1</th><th>Missing</th><th>Total N</th></tr>
<tr><td>397</td><td colspan="2">942</td><td>31</td><td>1370</td></tr>
<tr><td colspan="6"></td></tr>
<tr><th rowspan="2">Grade in School:</th><th>Grade 3</th><th>Grade 4</th><th>Grade 5</th><th>Missing</th><th>Total N</th></tr>
<tr><td>388</td><td>467</td><td>515</td><td>0</td><td>1370</td></tr>
</table>

Exploratory Factor Analysis

We used the Principal Axis Factoring (PAF) method of Exploratory Factor Analysis (EFA) in SPSS 23[10] because we expected our coded data to be non-normal. The ratio of sample size (1370) to expected factors (<8) is quite high (170:1) so we expect that the sample size is sufficient for this procedure, even if extracted commonalities are low[11]. We used parallel analysis to determine the number of factors to retain. We used the Oblimin rotation with the PAF because we expect the resulting factors to be correlated to some extent. We then examined pattern matrices for item loadings, crossloadings, and internal consistency reliability (by calculating Cronbach’s alpha) to determine the best model fit and suitability of scales.

Comparison of Demographic Groups

We examined characteristics of the data we collected, both overall and across demographic groups. We tested for differences across demographic groups (gender, race, etc.) using the Independent-Samples Mann-Whitney U Test, after first checking the assumption that results were similarly distributed across groups using population pyramid graphs. We tested for differences across grades 3, 4, and 5 using the Kruskal-Wallis H test, with the assumption of similar distributions of scores assessed by visual inspection of boxplots.

Examining Validity Evidence

To compile evidence that the instrument is valid for its intended use—to measure the engineering process skills and knowledge of elementary school students—we used the rubric we developed to examine the videotaped interactions and journal inscriptions of 30 students from 8 engineering classes as they complete design challenges in their teams. Approximately 1 hour of engineering design work was examined from each of 11 student groups, with 3 or 4 students in each group. The sample includes 16 girls and 14 boys. Twenty students are White, 5 Black, and 5 Asian.

We compared the videotaped students' in-class performances to the work they did on the EEPA, to see if the skills we intend to measure are congruent. Because of time constraints, we limited our qualitative analysis to students' explanations and justifications during design planning, which corresponds to the "How Will Solve" section of the EEPA. We analyzed the transcript turn-by-turn, to identify student contributions where they explained, justified, or argued for their engineering design ideas, then scored these design element explanations according to the same rubric we used in analyzing the EEPA.

Results

Student feedback on the EEPA during pilot testing was that it was engaging and interesting to them, and student responses indicated that they were interpreting the prompts as intended. With this initial evidence of validity for use as an elementary school assessment of engineering design practices, we went on to collect and code a larger sample.

Inter-rater reliability for our 3 coders proved to be Good or Excellent for all variables (Table 5).

Table 5: ICC measures of inter-rater reliability

Coded Variable	ICC estimate	Qualitative Rating of Agreement
Num_Ideas	.804	Excellent
Meets_Criteria	.707	Good
How_Will_Solve	.669	Good
Drawings	.772	Excellent
Num_Materials	.826	Excellent
Do_Next	.848	Excellent
Engineering?	.810	Excellent

Exploratory Factor Analysis

Parallel analysis determined that one factor should be retained for the 7 coded items. We manually set the number of factors to retain in the analysis to 1. Exploratory Factor analysis confirmed one factor accounting for 19.75% of the variance. Though Cronbach's alpha was reasonable (.610) and Bartlett's Test of Sphericity ($p<0.001$) indicated high factorability, extracted communalities were all low (<.4). The lowest communality was for the variable Num_Ideas (.054), and reliability analysis also indicated that Cronbach's alpha would improve if this item were dropped from the scale. We repeated the analysis with this item dropped, leaving 6 items. Internal consistency reliability improved marginally (Cronbach's alpha=.617), as did the percentage of variance accounted for (22.19%) but communalities remained low (<.4). We concluded that all items should be analyzed separately.

Comparison of Demographic Groups

Nonparametric tests showed that girls consistently outperformed boys on the EEPA (see Table 6). On all but the “Number of Materials” section, girls did significantly better.

Table 6: Gender

Coded Variable	**Mean Rank** (Male)	**Mean Rank** (Female)	**Median** (Male)	**Median** (Female)	**Mann-Whitney U**	**p-value**
Num_Ideas	617.74	680.12	1	1	190049	<.001
Meets_Criteria	604.21	693.58	2	3	181299	<.001
How_Will_Solve	627.92	669.98	2	2	196637	.025
Drawings	595.25	702.58	2	2	175501	<.001
Num_Materials	634.33	663.60	2	2	200786	.122
Do_Next	615.41	682.44	3	3	188541	<.000
Engineering	625.52	672.37	2	2	195082	.015

N (Male)=647, N (Female)=650, N (Missing)=3

Students from racial and ethnic groups that are well-represented in engineering in the United States (White and Asian) performed consistently significantly better than students from underrepresented racial and ethnic groups (Black and Hispanic), on all parts of the EEPA (see Table 7).

Table 7: Race / Ethnicity

Coded Variable	**Mean Rank** (Represented)	**Mean Rank** (UnderRep)	**Median** (Repres.)	**Median** (UnderR)	**Mann-Whitney U**	**p-value**
Num_Ideas	655.02	595.4	1	1	167447	.001
Meets_Criteria	677.78	555.11	2	2	149034	<.001
How_Will_Solve	668.63	571.31	2	2	156436	<.001
Drawings	686.74	539.26	2	2	141789	<.001
Num_Materials	659.64	587.22	3	2	163709	<.001
Do_Next	665.07	577.61	3	2	159316	<.001
Engineering	679.4	552.24	2	2	147721	<.001

N (Underrepresented)=457, N (Represented)=809, N (Missing)=104

Students whose parents reported higher income performed consistently better on the EEPA than students whose parents reported lower income, as determined by student eligibility for the National School Lunch Program (NSLP--see Table 8). Higher-income students performed significantly better on all parts of the assessment except the “Number of Ideas” section. It is important to note that Race / Ethnicity is significantly correlated with eligibility for the National School Lunch Program, with a Pearson Correlation of .348, p-value<.001.

Table 8: Eligibility for National School Lunch Program

Coded Variable	**Mean Rank** (Ineligible)	**Mean Rank** (Eligible)	**Median** (Ineligible)	**Median** (Eligible)	**Mann-Whitney U**	**p-value**
Num_Ideas	338.18	328.26	1	1	52619	.446
Meets_Criteria	351.50	309.96	3	2	47477	.002
How_Will_Solve	356.50	303.10	3	2	45549	<.001
Drawings	353.37	307.40	2	2	46758	.001
Num_Materials	356.85	302.62	3	2	45414	<.001
Do_Next	353.03	307.85	3	3	46886	.001
Engineering	361.50	296.22	2	2	43617	<.001

N (Ineligible)=386, N (Eligible)=281, N (Missing)=703

It is also important to note the high rate of missing data for the NSLP variable (51.3%). Given this high rate, we also compared students from Title 1 schools (those with 40% or more students from low-income families) to those from schools not designated as Title 1 schools. Students from Title 1 schools performed worse than other students on all but 2 sections of the EEPA, the “Drawings” and “Do Next” sections (see Table 9).

Table 9: From a Title 1 School

Coded Variable	**Mean Rank** (Not Title 1)	**Mean Rank** (Title 1)	**Median** (Not Title 1.)	**Median** (Title 1)	**Mann-Whitney U**	**p-value**
Num_Ideas	745.23	638.3	1	1	157122	<.001
Meets_Criteria	705.35	655.10	2	2	172954	.017
How_Will_Solve	708.42	653.81	2	2	171735	.009
Drawings	692.15	660.67	2	2	178195	.132
Num_Materials	702.00	656.51	3	2	174284	.031
Do_Next	691.84	660.80	3	3	178316	.134
Engineering	706.38	654.67	2	2	172544	.016

N (Not Title 1)=397, N (Title 1)=942, N (Missing)=31

Finally, we found that the youngest (Grade 3) students performed consistently worse than other students—significantly worse on all but the “Do Next” portion of the EEPA—while the oldest (Grade 5) students in our sample performed consistently better (see Tables 10 and 11). All medians were significantly higher except for the variable coding the “Number of Materials” section of the assessment, which was near significance (p=.054).

Table 10: Grade 3

Coded Variable	**Mean Rank** (Grade 3)	**Mean Rank** (Grade 4-5)	**Median** (Grade 3)	**Median** (Grade 4-5)	**Mann-Whitney U**	**p-value**
Num_Ideas	654.80	697.63	1	1	178597	.031
Meets_Criteria	658.18	696.29	2	2	179908	.078
How_Will_Solve	642.18	702.62	2	2	173698	.005
Drawings	619.16	711.71	2	2	164767	<.001
Num_Materials	658.53	696.16	2	2	180044	.081
Do_Next	679.93	687.70	3	3	188348	.714
Engineering	641.58	702.85	2	2	173.468	.005

N (Grade 3)=388, N (Grade 4-5)=982, N (Missing)=0

Table 11: Grade 5

Coded Variable	**Mean Rank** (Grade 3-4)	**Mean Rank** (Grade 5)	**Median** (Grade 3-4)	**Median** (Grade 5)	**Mann-Whitney U**	**p-value**
Num_Ideas	658.65	730.08	1	1	243121	<.001
Meets_Criteria	648.26	747.32	2	3	252001	<.001
How_Will_Solve	649.19	745.78	2	3	251207	<.001
Drawings	650.59	743.46	2	2	250011	<.001
Num_Materials	670.96	709.63	2	3	232592	.054
Do_Next	651.19	742.47	3	3	249400	<.001
Engineering	645.05	752.65	2	2	254747	<.001

N (Grade 3-4)=855, N (Grade 5)=515, N (Missing)=0

Validity Evidence from Video Analysis

In our analysis of video of students' planning time, we found that most groups engaged in 3 distinct phases of planning activity:

1. Share-out of individually brainstormed ideas
2. Arguing out the details of a group plan
3. Inscribing the final plan in journals

During the share-out of individually brainstormed ideas, students took turns explaining their ideas to each other, as in the following example (names are pseudonyms).

> Dan: You- that's- My idea is a little triangle.
> Rayna: Can I share my ideas now? Can I share my ideas now?
> Dan: Yeah.
> Rayna: Okay. So, my idea one, is it's an aluminum foil outside with small windows so the sun can get in.

During the second phase of planning, students worked on coming to consensus about their group plan. They argued about what to do—especially the details of their design.

Dan: Question. Shouldn't we use the sturdy clear plastic?
Serena: Or we could use a piece of milk carton and then cover the top with– um-
Dan: No, but then light couldn't get through.
Serena: No. Cover the top with this.
Dan: You put something that's clear on top of something that's not clear you still can't see through it.
Serena: No, on top. The sun will get in through the top.
Dan: Oh, you mean the cling wrap, now I get it.

Finally, each group inscribed the plan they had all decided upon in their journals. This often involved further explanation and justification.

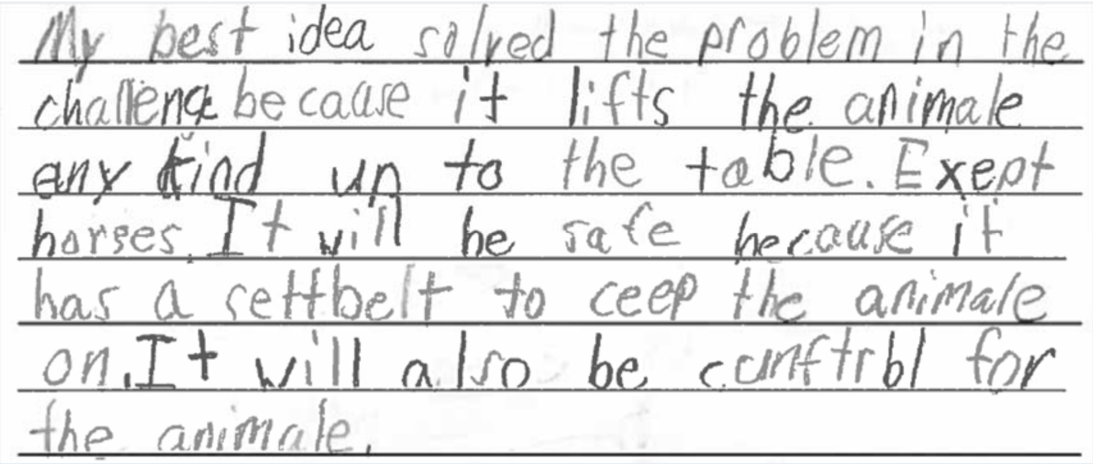
Dan: ((Points to Serena's drawing)) By the way, I think you need to (???) 'cause it's real confusing.
Rayna: ((Drawing in her journal)) Those are 25 cents, right?
Dan: ((To Serena)) No, it should look a little bit like this. ((Points to his journal))
Serena: Agh! What am I writing? All right, then we need-
…
Serena: 12 inches is 50 cents. So, we need a plastic bottle, and we need to measure that to see how much we'll need.
Dan: This is about a foot. This big, that big, I think it's enough.
Serena: Oh my gosh, yeah!

Analysis revealed that there were many differences between students' planning as a collaborative activity and their completion of the EEPA planning question. The motivation and audience for their explanations are quite different, with students working in groups needing to convince peers and come to group consensus, while work on the EEPA is for a grade or evaluation. Activity has a complexity that no assessment can match: coordination of joint activity, interactions, motivations, negotiation of ideas, and negotiation of standing.

However, we also saw important similarities between the planning activity and performance assessment of planning explanations. Just as on the EEPA, students explained and/or justified their ideas with differing degrees of skill during discussion. The form of the explanations and justifications was similar across tasks, and the same rubric could be used to code explanations and justifications across tasks.

Figure 2 below shows examples of explanations that both were coded as 3, the highest score: one from the engineering activity, the other from the EEPA. In speaking with her peers, Serena explains that in her design for a plant package, she wants to have "two straws holding up the plant." A little later, she justifies her idea to use straws: "because then the plant will be held up when we do the shake test." On the EEPA, similarly, the student's explanation and its justification are linked: "It will be safe because it has a settbelt to ceep the animale on (sic)."

Figure 2: Comparing Explanations across tasks

In Activity	Serena:	My idea is a milk carton at the bottom, plastic wrap over top, and **two straws holding up the plant.** On the inside, plastic wrap over the top of a foil bowl and the plant inside.
	…	
	Serena:	I like straws. **I think we should add the straws because then the plant will be held up when we do the shake test.**
On the EEPA	 My best idea solved the problem in the challenge because it lifts the animale any kind up to the table. Exept horses. It will be safe because it has a settbelt to ceep the animale on. It will also be cumftrbl for the animale.	

Discussion and Conclusion

The EEPA is a written performance assessment, intended to be used to measure some engineering skills and knowledge of elementary students, as outlined by the NGSS. Seven separate items can be coded or scored using a rubric by a teacher or researcher in one or two minutes, such that it can be used in classrooms and for research and evaluation studies with small to moderate populations; larger populations will require significant resources to complete the coding. Three researchers were able to reach inter-rater reliability after practicing on a handful of classes, suggesting that it can be reliably scored.

The coded responses do not form a strong factor or scale with internal consistency reliability and strongly loaded items, most likely because each item addresses a separate learning objective. Given the time needed for young students to complete the assessment is already high, the instrument should not have additional items.

However, we see from the nonparametric analysis comparing demographic groups that significant differences are readily apparent. Students from low-income families and schools, as well as students from racial and ethnic minority groups that are underrepresented in engineering (Black and Hispanic) do not perform as well as higher-income students, White students, and Asian students. Such achievement gaps are found on many assessments, including the National Assessment of Educational Progress (NAEP)[12], and so are unsurprising. The youngest (grade

3) students perform worse than older students, and the oldest (grade 5) students perform better than younger students, which again is unsurprising. These findings serve to demonstrate that the EEPA can distinguish between higher-achieving and lower-achieving students.

More surprising is the fact that girls perform much better on the assessment than boys. This may be due to the writing-heavy nature of the assessment, as there exists a persistent gender gap in the United States on writing assessments, with girls outperforming boys[13]. This suggests that scores should be read with caution: at least some component of the scores, with differences seen in the gender gap and in the gaps for the highly correlated groups of low-income students and underrepresented racial minorities, may be due to writing skills. It is possible that the EEPA is a test of writing to some extent—to what extent is unknown. Additional research is needed to gather evidence and distinguish the contributions of these possible explanations.

The similarity of students' engineering explanations on the EEPA to those they make to their peers, however, strengthens the argument that the EEPA is valid for use in measuring students' skills in "Constructing Explanations (in service of) Designing Solutions", a Practice called for by the NGSS[1]. In further work, we plan to make comparisons between students' other responses on the EEPA and their corresponding classroom engineering practices.

This preliminary work shows the EEPA to be a promising instrument for possible use in elementary engineering classrooms, research, and evaluation; however, additional work is needed to further establish whether the EEPA validly represents students' abilities in elementary engineering, to establish the relative roles of writing skills and engineering skills in determining scores, and to establish test-retest reliability.

References

[1] NGSS Lead States, 2013, Next Generation Science Standards: For states, By states, The National Academies Press, Washington, DC.

[2] Klassen, S., 2006, "Contextual assessment in science education: Background, issues, and policy," Science Education, **90**(5), pp. 820–851.

[3] Brown, J. D., 2004, "Performance assessment: Existing literature and directions for research," Second Language Studies, **22**(2), pp. 91–139.

[4] Shavelson, R. J., 2013, "On an approach to testing and modeling competence," Educational Psychologist, **48**(2), pp. 73–86.

[5] Abts, L. R., 2011, "Analysis of the barriers, constraints, and issues for dual credit and/or advanced placement pathway for introduction to engineering design," Proceedings of the ASEE Annual Conference & Exposition, Vancouver, BC.

[6] Kolodner, J. L., 2002, "Facilitating the learning of design practices: Lessons learned from an inquiry into science education," Journal of Industrial Teacher Education, **39**(3).

[7] Lachapelle, C. P., and Cunningham, C. M., 2014, "Engineering in elementary schools," Engineering in pre-college settings: Synthesizing research, policy, and practices, S. Purzer, J. Strobel, and M. Cardella, eds., Purdue University Press, Lafayette, IN, pp. 61–88.

[8] Cunningham, C. M., and Hester, K., 2007, "Engineering is Elementary: An engineering and technology curriculum for children," ASEE Annual Conference & Exposition, Honolulu, HI, p. 17.

[9] Hallgren, K. A., 2012, "Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial," Tutorials in Quantitative Methods for Psychology, **8**(1), pp. 23–34.

[10] IBM Corporation, 2012, IBM SPSS Statistics for Windows, IBM Corporation, Armonk, NY.

[11] MacCallum, R. C., Widaman, K. F., Zhang, S., and Hong, S., 1999, "Sample size in factor analysis," Psychological methods, **4**(1), p. 84.

[12] Reardon, S., 2011, “The widening achievement gap between the rich and the poor: New evidence and possible explanations,” Whither Opportunity?: Rising Inequality, Schools, and Children’s Life Chances, G.J. Duncan, and R.J. Murnane, eds., pp. 91–116.

[13] Klecker, B. M., 2006, “The gender gap in NAEP fourth-, eighth-, and twelfth-grade reading scores across years,” Reading Improvement, **43**(1), p. 50.