A Content Analysis of How Engineering is Assessed in Published Curricula

Dr. Kerrie Anna Douglas, Purdue University, West Lafayette (College of Engineering)

Dr. Douglas is an Assistant Professor in the Purdue School of Engineering Education. Her research is focused on methods of assessment and evaluation unique to engineering learning contexts.

Prof. Tamara J. Moore, Purdue University, West Lafayette (College of Engineering)

Tamara J. Moore, Ph.D., is an Associate Professor in the School of Engineering Education and Director of STEM Integration in the INSPIRE Institute at Purdue University. Dr. Moore's research is centered on the integration of STEM concepts in K-12 and postsecondary classrooms in order to help students make connections among the STEM disciplines and achieve deep understanding. Her work focuses on defining STEM integration and investigating its power for student learning. Tamara Moore received an NSF Early CAREER award in 2010 and a Presidential Early Career Award for Scientists and Engineers (PECASE) in 2012.

Hillary Elizabeth Merzdorf, Purdue University
Tingxuan Li, Purdue University
Miss Amanda C. Johnston, Purdue University, West Lafayette (College of Engineering)

From Standards to Classrooms: A Content Analysis of How Engineering is Assessed in Published Curriculum

Two of the major shifts brought about by *Next Generation Science Standards*[1] are an increased emphasis in students' capabilities to perform higher-level reasoning skills and integrate content understanding into science practices. At the same time, NGSS has made engineering integration into science education a priority, and it is an exciting time of reform as schools are exploring curriculum resources and teachers are being trained in engineering design. When engineering is a part of science instruction, there must also be corresponding measurement of student learning, yet many teachers who are new to engineering are also unfamiliar with the process of assessing design practices. In addition, teachers must grapple with how to assess higher order skills, including how students use science to make design decisions. The practices, crosscutting concepts, and core ideas of NGSS represent general patterns of thinking and understanding that students may exhibit at each grade level. Assessment must be able to capture learning on each of the three dimensions to be informative, and it should support classroom learning of science and engineering in line with framework recommendations.[2] Because the goal of NGSS assessment is to provide evidence of higher level learning, it is imperative that teachers are provided with the means to properly monitor student learning of both content and engineering practices.[3]

Currently, there are few engineering-related assessments for elementary and middle-school. A large-scale engineering assessment was implemented by the National Center for Education Statistics[4] to measure 8th grade students' technology and engineering literacy using the National Assessment of Educational Progress (NAEP) Technology and Engineering Literacy (TEL) assessment. It is a computer-based assessment where the competency being measured is the students' ability to apply technology and engineering skills to real-life settings. Also, many researchers have developed two types of assessment tools: *cognitive* and *non-cognitive*. The cognitive assessment tools aim to assess students' thinking skills such as problem-solving. In engineering education, Doppelt[5] aimed to assess students' problem-based learning, while Denson, Buelin, Lammi, and D'Amico[6] developed a web-based tool as creativity assessment to measure the innovation of students' design products. Kelly, Capobianco, and Kaluf[7] used think-aloud protocols to assess student cognition during the design process, and found that they emphasize brainstorming more than other aspects such as testing or refinement of design solutions. Non-cognitive assessment tools aim to assess students' "soft skills" such as interests, perception, or attitudes. These skills are important in learning and instruction, because the research has found them to be correlated to students' learning outcomes, such as self-reported learning gains or the scores on the standardized tests.[8] Douglas and Strobel[9] developed a STEM goal-specific hope scale to identify students' ability regarding their current effort in STEM subjects with future hope, thus laying the foundation for motivation and achievement. Capobianco, Ji, and French[10] developed engineering identity development scale to examine elementary school students' identity development in engineering. By looking into the difference of scores produced from the instrument before and after the unit, the statistical significance suggested that students improved their ratings of academic identity, career identity and engineering aspirations.

However, there is a gap in the literature concerning assessments for teachers to assess

students' learning of both science content and engineering practices in the classroom. Furthermore, while much NGSS reform has focused on pedagogy and curriculum, there has been less resources readily available for assessments aligned to NGSS. One place teachers and schools can look for example assessments are in the integrated STEM curricula units commercially available. By examining current STEM assessments with two frameworks, this study aims to answer the following research questions: (1) What aspects of engineering are being assessed in common engineering or integrated STEM curricular units? (2) What level of cognitive demand is being referenced by these assessments? (3) What level of cognitive demand is assessed for each aspect of engineering design? Using a purposeful sampling strategy, the authors reviewed nine engineering curricula units published by 3 different publishing companies. To address the research questions, assessment tasks were coded based on the Task Analysis Guide in Science (TAGS) framework, and on the engineering process of design (POD) and engineering and technology literacy.

Theoretical Background

Task Analysis Guide in Science (TAGS)

Task Analysis Guide in Science (TAGS)[3] is a framework for analyzing the level of learning for assessment tasks developed as part of science learning. Simply put, a task or an item in the assessment can be characterized into different levels of learning by using this framework. On the vertical dimension, it has three categories (a) *scientific practice*, (b) *science content*, and (c) *integration of content and practices*. Scientific practices encourage science classrooms to mimic a scientific community. Scientific practices require students to go beyond memorized understanding of the content to application of the content in genuine scientific practice. The categories for scientific practices required by the NGSS include asking questions, developing and utilizing models, brainstorming and investigating, presenting data, applying mathematics, forming the interpretations, connecting the interpretations from the evidence, and presenting the results. Science content is that knowledge of scientific explanations. It also includes the basic facts such as formulas, terminology, or a set of procedures related to a scientific principle. *Integration of content and practices* requires students to connect the authentic science practices and meaningful disciplinary core ideas. Students complete the tasks about scientific practices within the core disciplinary knowledge. A task asks students to propose a model and show the relationship as an explanation of a real-world phenomenon.

In addition to the three categories mentioned above, the TAGS framework also contains the cognitive demands at the vertical dimension: Memorized Practices (MP), Memorized Content (MC), Scripted Practices (SP), Scripted Content (SC), Scripted Integration (SI), Guided Practice (GP), Guided Content (GC), Guided Integration (GI), and Doing Science (DS). For the purposes of this paper, we replace the Doing Science dimension with Doing Engineering (DE), and use engineering in place of science as appropriate in our descriptions of the dimensions. A MP task requires students to reproduce descriptions of science /engineering practices. A MC task requires students to memorize a collection of definitions as whole. A SP task requires students following a set of procedures. A SC task requires students to use steps related to a specific principle. A GP task requires students to create explanations about a specific science/engineering practice. A SI task requires students to follow basic procedures within both content and practice. A GC task requires students to have high cognitive processes such as producing ideas. A GI task contains more written text and requires students to have high-level thinking. A DE task is very

open-ended and requires students to develop a solution with the combination of practice and content.

The TAGS framework has many similarities with the revised Bloom's taxonomy[7]. The revised Bloom's taxonomy is for characterizing educational objectives. It also has two dimensions, knowledge and cognitive processes. On the knowledge dimension, four categories are used: factual knowledge, conceptual knowledge, procedural knowledge, and metacognitive knowledge. On the cognitive processes dimension, six categories are used: remember, understand, apply, analyze, evaluate, and create. These categories are in a hierarchical order. The revised Bloom's taxonomy has been applied to many subjects such as English or mathematics classrooms. The advantage of the TAGS framework is that science/engineering content and practice can be reflected together as *integration*. In contrast, the revised Bloom's taxonomy does not have this advantage, because the integrative nature of science and engineering content and practice is missing. Therefore, we chose TAGS in this research.

Process of Design (POD), Engineering Literacy, and Technology Literacy

The Process of Design (POD) is a framework derived from the key indicators identified by Moore, Glancy, Tank, Kersten, Smith, & Stohlmann[12] within their *Framework for Quality K-12 Engineering Education.* It is a research-based, rigorously evaluated framework which maps to the common design processes presented in literature and is intended to guide engineering-based inquiry. By providing a definitive set of concepts that are essential to engineering education, it allows us to examine the ways these concepts are reinforced by assessment in integrated STEM curricula.

The rank of each indicator in the framework indicates its relevance for equipping students with fundamental engineering knowledge and skills. According to this structure, the most important material that engineering education should include is the process of design (POD), which the framework divides into three steps. *Problem* and *Background* (POD-PB) stages teach students to scope an engineering problem, identify criteria and constraints to guide solution brainstorming, and collect relevant information from a variety of sources. Engineering students will apply this information to *Plan* and *Implement* (POD-PI) a solution and create a prototype, and draw conclusions and make decisions about the fit of the solution based on the prototype's performance in *Test* and *Evaluate* (POD-TE). The remaining indicators in the framework are practices necessary to engineering, but are outside of POD and may also relate to other disciplines. Students apply science, engineering, and mathematics (SEM) by learning from problems that stress the interdisciplinary nature of these subjects. *Engineering Thinking* (EThink) is a mindset that students strive for by problem-solving, critically examining challenges, managing uncertainty, and using metacognition during the design process and other relevant engineering activities. Instruction in engineering will also help students develop *Conceptions of Engineers and Engineering* (CEE) as they understand the many fields of work within engineering and engineers roles in society. Becoming adept with the *Tools, Techniques, and Processes* (ETools) for successfully accomplishing tasks is a goal of engineering education outside of the design process itself. When studying design problems, students should be mindful of the surrounding *Issues, Solutions, and Impact* (ISI) and the global systems they affect, while adopting the *Ethical Responsibility* (Ethics) of following engineering regulations and standards. Finally, *Teamwork* (Team) and *Communication* (Comm-Engr) are essential to authentic K-12 engineering education, where students are prepared to collaborate and interact with fellow

engineers, clients, and colleagues.

The primary coding variable used in this study was process of design (POD), and we considered the indicators to be six distinct categories instead of three. Additionally, we included *Communicate* as a seventh step within the design process, in which students communicate design solutions to clients. The secondary coding variable was Engineering Literacy, and it consisted of the remaining framework indicators outside of POD. Technology Literacy was the third coding variable for test items that were meant to assess students' knowledge of particular technology in the curricula without connecting it in any way to engineering. This variable included ideas such as, but not limited to, vocabulary words about technology, learning about how a technology works, and learning about how technology is used in the real world. We recognize that many of the indicators from the *Framework for Quality K-12 Engineering Education* would fall both within common definitions of engineering literacy and technology literacy, but for this study we defined it as above. If the curriculum had presented a design problem before assessment, we coded items as testing students on one of the seven design steps. If the design problem was not yet introduced, the assessment items were coded to the appropriate engineering literacy indicator. Assessment items were coded to technology literacy if they tested students only on their knowledge of technology related to the curriculum. An overview of our codes and their definitions are presented in Table 1.

Table 1. Definitions of TAGS and POD coding terms.

Task Analysis Guide in Science (TAGS)	
Memorized Practice	Reproducing descriptions of scientific/engineering practices
Memorized Content	Memorizing a collection of scientific/engineering definitions
Scripted Practice	Following a standard set of procedures
Scripted Content	Using steps related to a standard principle
Scripted Integration	Following basic procedures within both content and practice
Guided Practice	Creating explanations about a scientific/engineering practice
Guided Content	Using higher cognitive processes, such as producing ideas
Guided Integration	Using higher level thinking within both content and practice
Doing Engineering	Developing a solution combining content and practice
Process of Design (POD)	
Problem	Scoping an engineering design problem
Background	Collecting relevant information for solution
Plan	Formulating and selecting solution ideas
Implement	Creating a prototype of solution
Test	Performing experiments with prototype

Evaluate	Making decisions about the fit of solution
Communication	Consolidating solution information for client

Methodology

We used content analysis[13] in this research. The targeting data unit we analyzed was each item embedded in the curriculum unit. In particular, we used a top down/deductive method based on the existing frameworks— the *Task Analysis Guide in Science* (TAGS) and *Process of Design* (POD) from the *Framework for Quality K-12 Engineering Education*. To explore what aspects of engineering design and the level of learning expectation are commonly assessed in integrated STEM elementary curricula, we purposefully chose nine curricula units designed for grades 3-5 from three publishers: ETA hand2mind; Invention, Innovation, and Inquiry (I^3), and Engineering is Elementary (EiE). In total, we located 1079 assessment items as part of worksheets, end-of unit quizzes, or post-tests. Unit rubrics were coded as well as assessments, due to their role in measuring progress and guiding learning.

In this section, we provide the computational details of the inter-rater reliability (IRR). In this research, we used four coders to rate the items on the assessment embedded in the curriculum. Two coders rated the items based on TAGS framework. The other two coders rated the items based on POD framework. The coded results are the data on a nominal scale. Therefore, the Krippendorff's alpha reliability coefficient is used to compute the IRR[13]. The IRR originated from the classical test theory (CTT).[10] Equation 1 shows the observed score X is the sum of true score T and the measurement error E. Therefore, the variance of the observed scores can be decomposed to two parts, that is, $Var(X) = Var(T) + Var(E)$. The IRR reliability coefficient in equation 2 can therefore determine that the amount of variance in the observed scores is explained by the variance of the true scores after the measurement error variance is removed.

$$X = T + E \qquad \text{Equation 1}$$

$$IRR = \frac{Var(T)}{Var(X)} = \frac{Var(X) - Var(E)}{Var(X)} = \frac{Var(T)}{Var(T) - Var(E)} \qquad \text{Equation 2}$$

In our study, the IRR for the TAGS is 0.67, which indicates that the 67% of the variance in the observed scores is due to true score variance without accounting the measurement error between coders; 33% of the variance is due to the differences between coders. The IRR for POD is 0.80. According to Hallgren,[15] the value between 0.61 and 0.80 indicates substantial agreement between coders.

Results

We use six sections to describe our findings. The first section reports the findings related to the level of cognitive demand assessed in all curriculum unit tasks. The second section reports the results related to what aspects of design are assessed in all curriculum units. The third section combines the level of cognitive demand with each step in the process of design to understand the assessment characteristics on both dimensions. The fourth section describes the cognitive demand of assessment items for each publisher. The fifth section reports the aspects of design being assessed by each publisher. The sixth section compares each publisher on the combined cognitive demand and aspects of design.

Level of Cognitive Demand

To see what level of cognitive demand is being referenced by these assessments, we used the frequency analysis based on the data collected from the TAGS framework. Examples of assessment items within the categories are given in Table 2. They illustrate differences among memorized, scripted, and guided codes for items measuring engineering practice. Memorized practice assessment asks students to provide information related to engineering practice. Scripted assessment requires students to perform according to a set of instructions. At a higher level, guided assessment is more open-ended but includes prompts, while assessment of doing engineering is unstructured by expecting students to incorporate engineering practice and content into one response.

Table 2. Examples of levels of cognitive demand.

TAGS - Practice	**Example**
Memorized Practice	Fill in the blank: "One part of our model that did not work well was____"
Scripted Practice	"Sketch two ideas of the vehicle on the grid below"
Guided Practice	Open ended with framing: "Observe the materials that can be used to make your system. Think about how each of these materials could contribute to its structure"
Doing Engineering	"How would you change your design based on test results?"

The frequency plot in Figure 1 explains the pattern among all the categories. As shown in Figure 1, Memorized Practice is the most frequent type of assessment task, with 510 tasks across the nine units. These tasks require a basic understanding of practices, where students are expected to provide a definitional answer or explanation of an engineering practice.[3] The next most frequently assessed type of tasks was Memorized Content, which is similar to the revised Bloom's taxonomy[11] of remembering. In contrast, the tasks requiring a higher level of thinking, labeled as Guided Science or Guided Practice, only accounted for 35 of the 1079 tasks, or approximately 4% among all tasks. They are similar to the revised Bloom's taxonomy of analyzing and organizing. In this context, Guided refers to scaffolding;[3] students are expected ''to grasp a particular concept or achieve a particular level of understanding.''[16] Additionally, Scripted Content and Scripted Practice together accounted for 126 tasks. These types of tasks provide a certain amount of instruction to students, which are similar to a "cook-book procedure" in science classrooms.[3] Students only need to follow the pre-written procedures in order to complete the tasks. In another word, students do not need to understand the underlying scientific principles in order to successfully complete the items. Doing Engineering accounted for eight tasks. These kinds of tasks provide little guidance to students. It is similar to the revised Bloom's

taxonomy[11] of creating. The typical item that is in this category includes drawing a design from scratch.

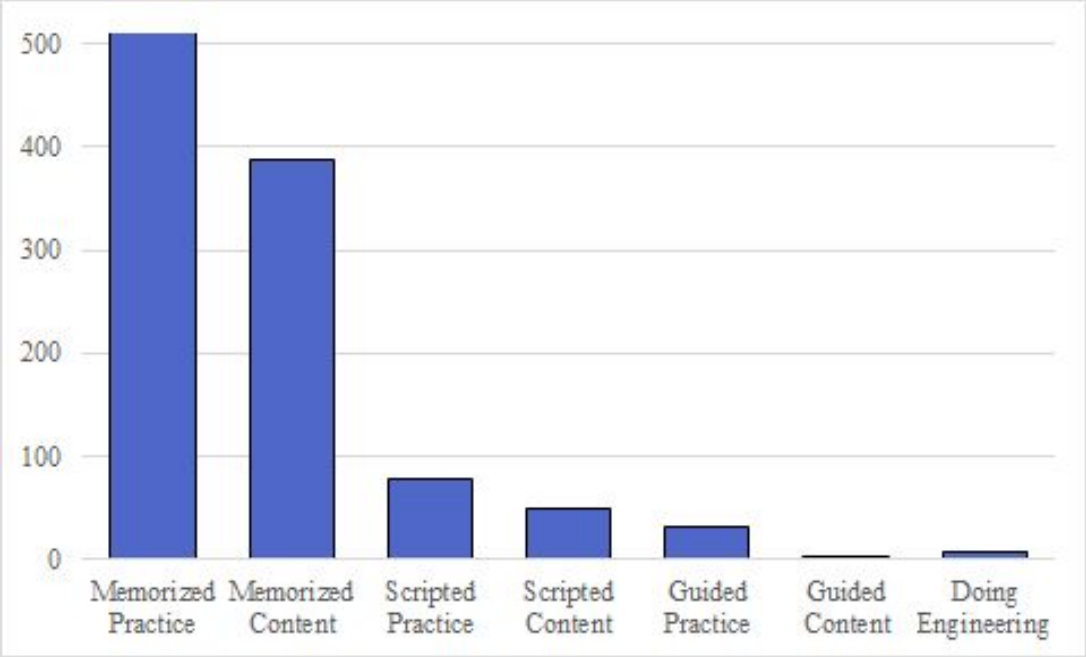

Figure 1. Frequency Analysis on the TAGS Framework.

POD, Engineering Literacy, and Technology Literacy

To learn what aspects of engineering design received the strongest focus in curricula, we analyzed the assessment items according to the engineering education framework developed by Moore et al.[12] In Figure 2, of all items coded, 440 tested the process of design (POD), 329 evaluated Engineering and Technology Literacy, and 305 were Outside of Engineering.

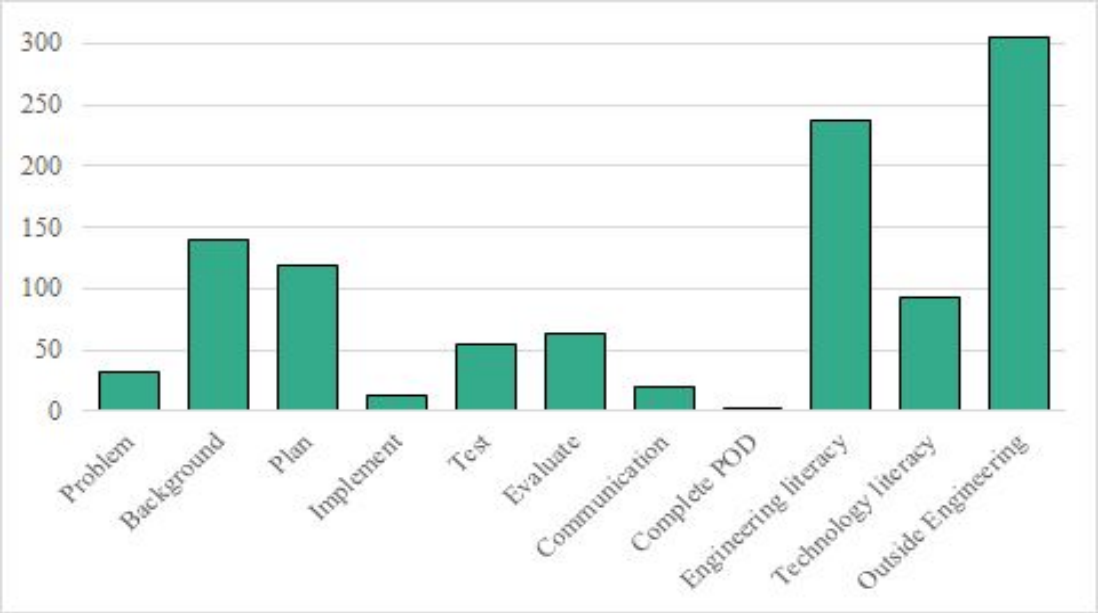

Figure 2. Frequency Analysis on the POD Framework.

Within POD, assessment items were most often from *Background*, in which students were primarily tested about solution materials, background information for the solution, or both. *Plan* was the second most commonly assessed step, with providing information about the plan and communicating ideas being the strongest focus of these items. The third step with a large number of assessment items was Evaluate*,* with students making design decisions, brainstorming changes for redesign, supplying evidence for these choices, and deciding if the design met criteria and constraints. Test, Problem, and Communicate all contained a small proportion of items. Finally, 12 items assessed the Implement step, and one item completely assessed all steps of POD. Within Engineering Literacy and Technology Literacy, assessment items most often tested students' conceptualizations of engineering and engineers (CEE), their ability to use science, engineering, and mathematics (SEM), and their use of engineering thinking (EThink). If assessment items tested learning Outside Engineering, they had either been used by the curricula for either data collection during an activity or for vocabulary checks related to the lesson content.

Cognitive Demand of POD, Engineering Literacy, and Technology Literacy

To investigate integrated STEM assessment on the dimensions of content focus and cognitive demand, we combined frequencies from both frameworks to determine the levels of thinking required for each step of POD, and for engineering and technology literacy. Figure 5 shows the proportion of items at each level of cognitive demand, within the POD and engineering or technology literacy categories. Aside from one item representing complete POD which tested Memorized Content, the most homogenous steps were Implement and Test. Memorized Practice was the demand of nearly all POD categories, while Guided Content was only reached in Background assessment items. Scripted Practice was used most often during Implement, Evaluate, and Plan, but higher-level cognitive demands such as Doing Engineering and Guided Practice received very little attention from the assessments in any category.

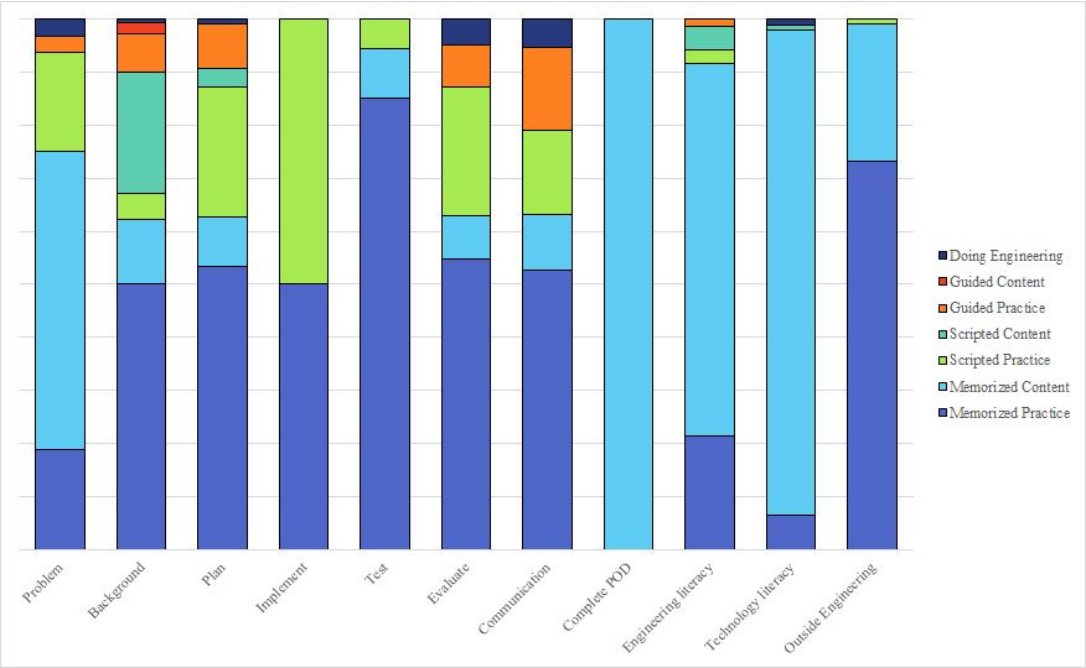

Figure 3. Cognitive Demand of POD Assessment Items.

Level of Cognitive Demand by Publisher

To determine the level of cognitive demand of assessments from each publisher, we reported the number of items within each coding category of the TAGS framework. As Figure 4 shows, ETA hand2mind curricula used slightly more Scripted Practice items than Memorized Content items, but also contained the most Memorized Content items. I3 contributed relatively few assessments to the analysis, but the majority of their items targeted the Memorized Content level of demand. While assessment by EiE relied mostly on Memorized items, they also published the only curricula to assess the three highest levels of cognitive demand, Guided Practice, Guided Content, and Doing Engineering.

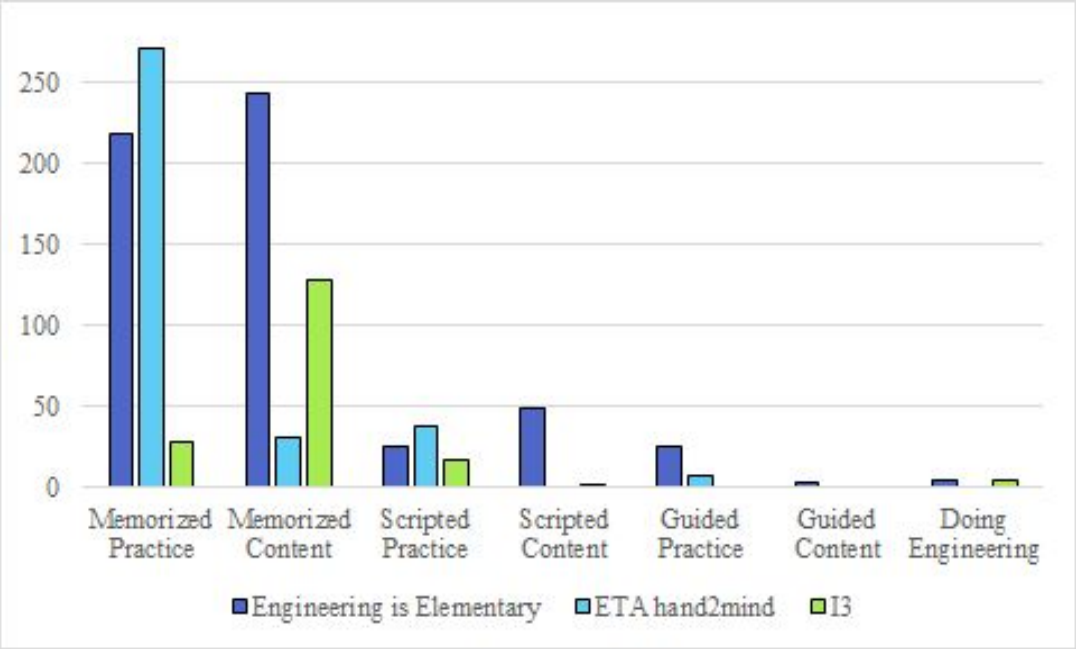

Figure 4. Frequency Analysis of the TAGS Framework by Publisher

POD, Engineering Literacy, and Technology Literacy by Publisher

To analyze the assessment characteristics from the three curricula per publisher, we examined the number of items for the process of design, engineering and technology literacy, and outside engineering. Similar to the overall results, Figure 5 shows that the individual publishers focused primarily on Memorized Practice and Content, and all three publishers had few items asking for higher levels of thinking. The breadth of assessment items measured against the POD framework varied by publisher, with EiE focusing primarily on Background and engineering literacy items, I3 primarily on plan and technology literacy. The majority of ETA hand2mind items did not fall into the POD framework, however, those that did fall are more evenly across the range of POD.

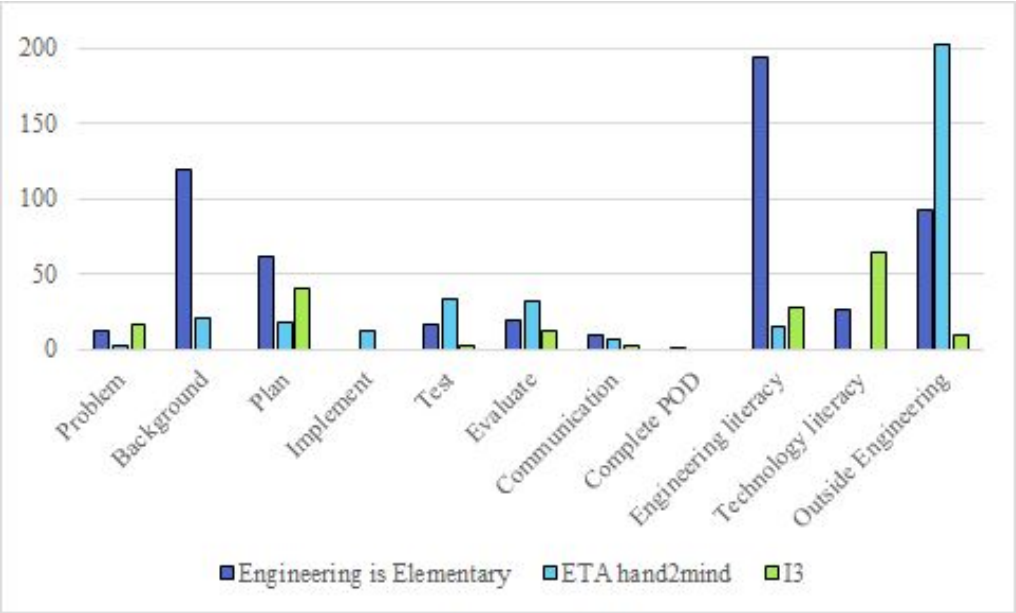

Figure 5. Frequency Analysis on the POD Framework by Publisher.

Cognitive Demand of POD by Publisher

After using the TAGS and POD frameworks to simultaneously describe the assessment items, we examined the distribution of these items across publishers. Figure 6 shows the number of items assessing each POD step and the curricula containing them, along with the cognitive levels targeted by each step. Background is the most frequently assessed POD category, and EiE implements Background assessments at every level of cognitive demand, but most often at Memorized Practice and Scripted Content. All steps of POD include items that target Scripted Practice, with hand2mind containing these items most often. I^3 assessed the Evaluate step as Doing Science more often than EiE, while hand2mind assessed Evaluate as Memorized or Scripted Practice. EiE and I^3 both assessed Test as Memorized Content or Practice, and hand2mind contained the most Test items overall at Memorized or Scripted Practice. Overall, the publishers focused on Background and Plan, and were more evenly distributed for the remaining categories. Even though EiE items contained the widest range of cognitive levels, there is little similarity among publishers within POD steps concerning the cognitive levels assessed.

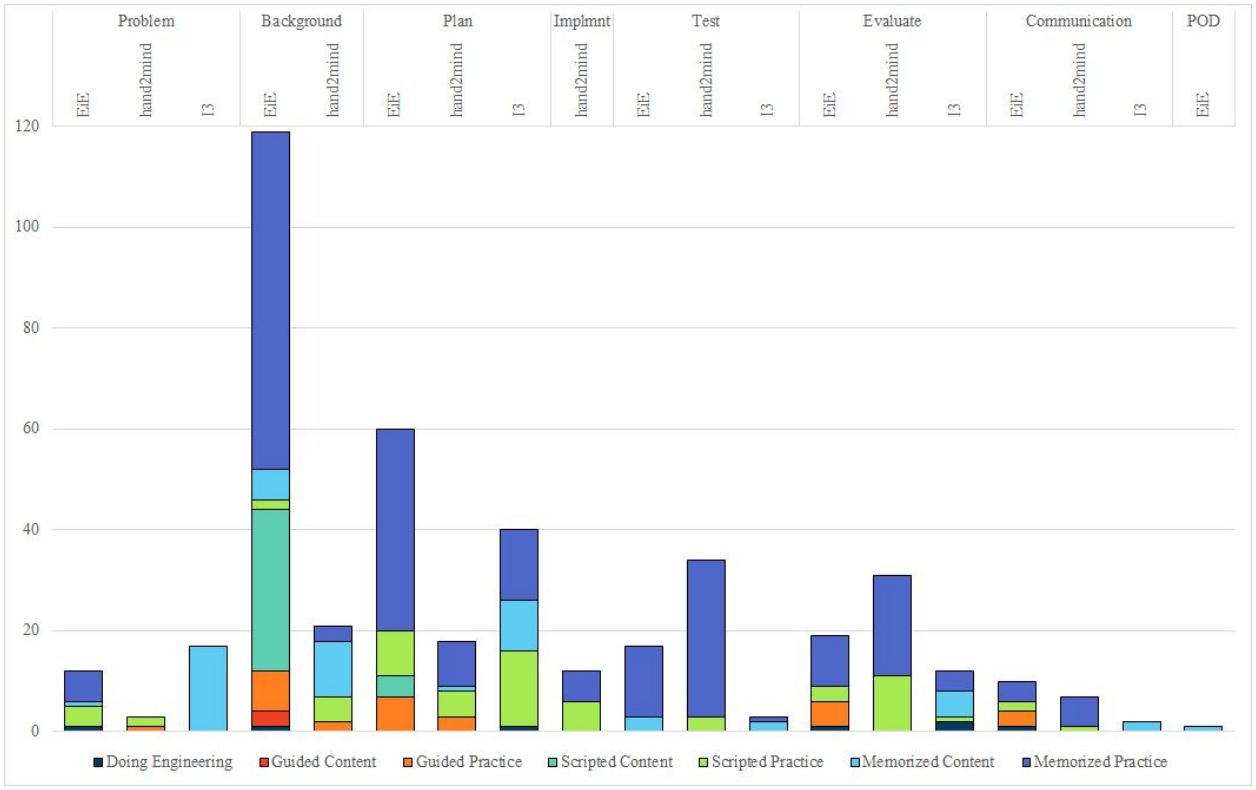

Figure 6. Cognitive Demand of POD Assessment Items by Publisher.

Cognitive Demand of Engineering and Technology Literacy by Publisher

For assessments of Engineering and Technology Literacy, as well as Outside Engineering, the publishers EiE and I^3 contained items from all three categories, at the levels of Memorized Practice and Content. The majority of assessments outside of POD were in hand2mind curricula, testing students Outside Engineering at the Memorized Practice level. EiE utilized the second highest number of assessments outside of POD in Engineering Literacy, and these items were primarily Memorized Content with a mix of Memorized Practice, Scripted Practice, Scripted Content, and Guided Practice. In comparison, the Engineering Literacy items for hand2mind focused on Memorized Practice and Content and Guided Practice, while I^3 assessed Memorized Content only. Of the two publishers that assessed Technology Literacy, Memorized Content was almost their entire emphasis.

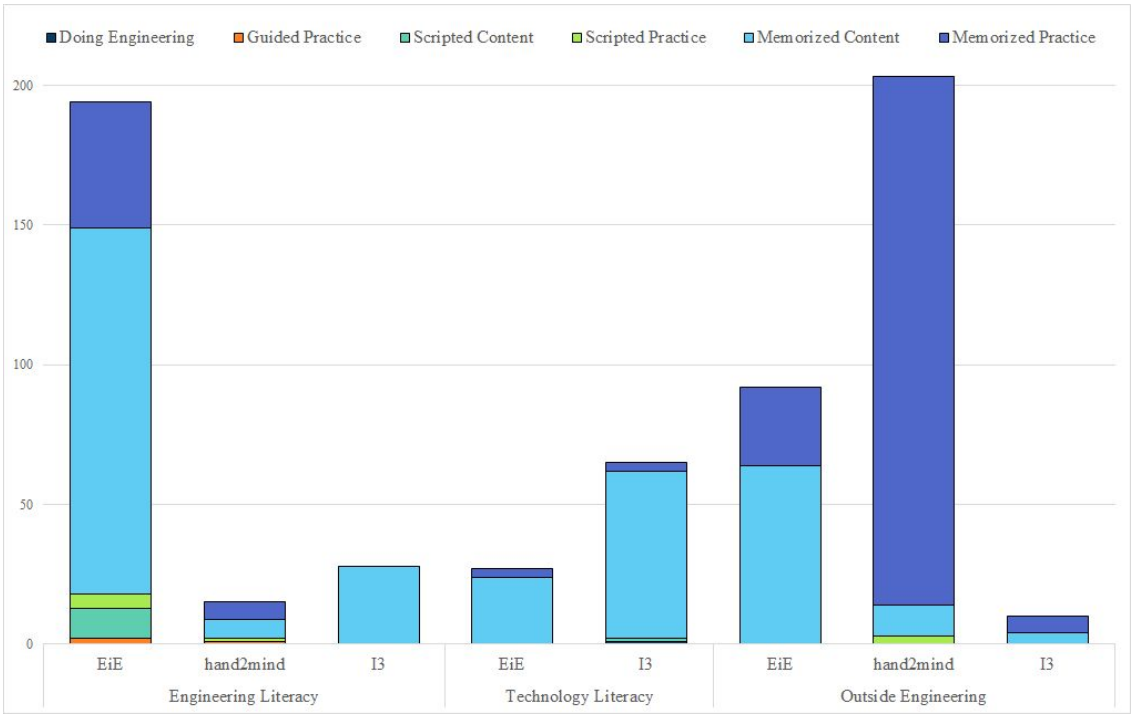

Figure 7. Cognitive Demand of Engineering and Technology Literacy Items by Publisher.

Conclusion

The majority of engineering assessment items from these units were dedicated to the process of design and engineering literacy. However, the early stages of POD, such as *Background* and *Plan*, had more items than later stages involving building, testing, and evaluating a prototype or model. In general, the proportion of assessment items dedicated to POD was very low compared to Engineering and Technology Literacy and Outside Engineering. In some cases, the curricula provide teachers resources to assess students' engineering design process during activities, but not outside of the actual activities in a summative manner. Many units dedicated a greater proportion of lessons to teaching preparatory science, technology, engineering, and mathematics content, causing the design challenge itself to be introduced relatively late. Based on our coding results on the items separate from POD, within-learning assessment is generally more focused on vocabulary related to the science content and data collection, compared to Engineering or Technology Literacy topics. Assessments were integrated into these units as guided activities more so than actual assessment of what students understood. For example, the curricula provide some opportunity for reflection and making inference, but overall, the emphasis of the worksheets was for recording observations and performing calculations. While these types of formative assessments are useful for projects, they do not adequately measure students' abilities to make engineering decisions from a depth of content understanding or their ability to scope an engineering design problem.

The findings based on the level of cognitive demand also supports the lack of assessment of students' abilities to make engineering designs and problem scope. Lower cognitive demand categories such as Memorized Practices or Memorized Content are the foundations for students

to develop high-order thinking. However, by studying these curricula, we noticed that students can be "hands on but not minds on", in the process of solving the tasks. In particular, Guided Content and Guided Practice tasks are lacking in the assessments. This implies that in reality students have limited opportunities to reflect or make inferences, given these items. In order to solve problems in the context of applying engineering design in the science classroom, the students will need to have the opportunities to use the guided information to solve the problems, rather than being asked to simply record the observations or perform calculations.

The intersection of TAGS and POD clearly demonstrated a lack of higher-level cognitive demands in several important areas. For example, Evaluate requires students to use critical thinking to examine their design after testing, but assessments only reached Memorized or Scripted Practice. Engineering Literacy and Technology Literacy items promoted deeper student thinking, by being Guided instead of Scripted or Memorized. However, the most typical cognitive demand for every step of POD was indisputably Memorized Practice, meaning that students are being tested on their ability to recall information about scientific practices. Integrated STEM curricula are meant to teach engineering design and literacy as ways of thinking, not facts to memorize or scripts to follow. We suggest that the corresponding measurement is in need as the part of curriculum when engineering is part of science instruction.

After examining the types of assessment within each of the nine curricula, we concluded that publishers who included a greater number of assessment items often placed them at the beginning of the design process or in Engineering Literacy. ETA hand2mind assessments were mainly Memorized Practice, and I^3 assessments were largely Memorized Content, but most Engineering is Elementary items assessed both Memorized Practice and Content. The majority of ETA hand2mind's assessments were Outside Engineering, at the level of Memorized Practice. Engineering is Elementary's items were concentrated on engineering design Background at all cognitive levels, and Engineering Literacy at the level of Scripted Practice. I^3 contributed few assessments to this study, and their items were frequently in at Memorized and Scripted levels for all POD steps, and at the Memorized levels in Engineering and Technology Literacy and Outside Engineering. From analyzing the combined results for each curriculum publisher, we conclude that both the frequency and the quality of assessment should be considered. Limiting classroom assessment to a high number of low-level tests or very few cognitively demanding assessments will not support effective, long-term engineering learning and instruction.

The results from this study imply that the evidence of higher levels of learning is mostly missing in engineering assessment. In addition, the tasks embedded in the curricular do not provide teachers opportunities to assess how students use science to make design decisions. Assessments that helps students reach these higher levels are necessary, if we hope to obtain a complete understanding of how students learn from integrated STEM curricula.

Implications

For NGSS reform to be successful, assessment systems must be developed for the classroom.[2] While schools and teachers can readily find published integrated STEM curriculum, teachers cannot rely on the curriculum to provide high-quality assessment tasks aligned to the expectations of NGSS that allow teachers to be able to see student learning gains. It is imperative that teachers and students have access to high-quality assessment both to support their development of deeper levels of understanding and skills, and also to have classroom experience with being tested with expectations beyond rote memorization. Standardized test companies will

align their testing programs to the NGSS, and students need to have had plenty of opportunities (with and without scaffolding) to demonstrate their content knowledge and practices prior to taking the tests. One potential in-road to preparing students is through developing high quality assessments as part of integrated STEM curriculum. More research is needed to understand how to design such assessments in a manner that allows teachers to fluidly assess students learning in engineering as they implement engineering-based STEM integration curricula.

References

1. Committee on a Conceptual Framework for New K-12 Science Education Standards National Research Council (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academies Press.
2. Pellegrino, J. W., Wilson, M. R., Koenig, J. A., & Beatty, A. S. (2014). *Developing Assessments for the Next Generation Science Standards*. Washington, DC: National Academies Press.
3. Tekkumru-Kisa, M., Stein, M. K., & Schunn, C. (2015). A framework of analyzing cognitive demand and content - practices integration: Task analysis guide in science. *Journal of Research in Science Teaching*, *52*(5), 659-685. doi: 10.1002/tea.21208
4. National Center for Education Statistics. (2014). *The Technology and Engineering Literacy (TEL) assessment* Washington, DC: U.S. Department of Education. Retrieved from https://nces.ed.gov/nationsreportcard/tel/
5. Doppelt, Y. (2005). Assessment of project-based learning in a Mechatronics context. *Journal of Technology Education*, *16*(2), 7–24. doi: 10.21061/jte.v16i2.a.1
6. Denson, C. D., Buelin, J. K., Lammi, M. D., & D'Amico, S. (2015). Developing Instrumentation for Assessing Creativity in Engineering Design. *Journal of Technology Education*, *27*(1), 23-40. doi: 10.21061/jte.v27i1.a.2
7. Kelley, T. R., Capobianco, B. M., & Kaluf, K. J. (2015). Concurrent think-aloud protocols to assess elementary design students. *International Journal of Technology and Design Education*, 25(4), 521-540. doi: 10.1007/s10798-014-9291-y
8. Garcia, E. (2014). The need to address noncognitive skills in the education policy agenda (Briefing Paper No. 386). Washington, DC: Economic Policy Institute.
9. Douglas, K. A., & Strobel, J. (2015). Hopes and Goals Survey for use in STEM elementary education. *International Journal of Technology and Design Education*, 25(2), 245-259. doi: 10.1007/s10798-014-9277-9
10. Capobianco, B. M., Ji, H. Y., & French, B. F. (2015). Effects of engineering design-based science on elementary school science students' engineering identity development across gender and grade. *Research in Science Education*, 45(2), 275-292. doi: 10.1007/s11165-014-9422-1

11.Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice, 41*(4), 212–218. doi: 10.1207/s15430421tip4104_2
12. Moore, T. J. , Glancy, A. W., Tank, K. M. , Kersten, J. A. , Smith, K. A. , & Stohlmann, M. S. (2014). A framework for quality K-12 engineering education: Research and development. *Journal of Precollege Engineering Education Research, 4*(1), 1–13. doi: 10.7771/2157-9288.1069
13. Krippendorff, K. (1980). Content analysis: An introduction to its methodology. Beverly Hills, CA: Sage.
14. Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, *3*(1), 1-18. doi: 10.1016/0022-2496(66)90002-2
15. Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23-34. doi: 10.20982/tqmp.08.1.p023
16. Maybin, J., Mercer, N., & Stierer, B. (1992). 'Scaffolding' learning in the classroom. In K. Norman (Ed.), *Thinking voices: The work of the National Oracy Project* (pp. 186-195). London, UK: Hodder & Stoughton.