Testing an assessment of problem-solving in introductory chemical process design courses (WIP)

Dr. Eric Burkholder, Stanford University
Prof. carl e wieman

Testing an assessment of problem-solving in introductory chemical process design courses (WIP)

Introduction:

Problem-solving is consistently cited as one of the most important outcomes of an undergraduate education in engineering [1-3]. While it is generally held that scientists and engineers are trained to be good problem-solvers, there is very little research that confirms this belief. Indeed some work suggests that engineering graduates are ill-prepared to solve the complex problems they encounter in the workplace [4]. Substantial work has been devoted to characterizing student and expert problem-solving in physics [5-11] and engineering [12-14], but there are almost no agreed-upon measures of problem solving [8]. If we are to teach undergraduate students to solve complex, real-world problems we must be able to measure how well they are learning the necessary skills. In this work, we describe the testing of a new assessment to measure dimensions of problem-solving in undergraduate chemical engineering courses.

Much of the empirical work in problem-solving has focused on differences between experts and novices as they solve structured problems [e.g. 6]. While this has provided valuable insights, such as the fact that novices focus on the surface features of problems while experts focus on the concepts underpinning the problem, it provides a limited picture of problem-solving because the problems that scientists and engineers encounter in the workplace are not well-structured. These "ill-structured" problems may have conflicting goals, multiple solution methods, multiple forms of representation, and non-engineering success standards [12]. Indeed, Hong, Jonassen and McGee (2003) found that solving these ill-structured problems involved higher-order metacognitive skills when compared with solving well-structured problems [15].

Price et al. conducted an empirical study of expert problem-solving that frames the process of an expert solving an ill-structured ("authentic") problem in terms of the decisions that experts make [16]. They find a remarkably consistent set of approximately 30 decisions that experts make as they solve problems, such as deciding to decompose the problem into smaller pieces, deciding on an appropriate abstract representation of the problem (e.g. diagrams or equations), and deciding on the failure modes of a potential solution. These empirical findings are in line with theory that suggests decision-making represents the core processes in solving a variety of complex problems, such as design problems [17, 18]. Central to Price et al.'s empirical model of problem solving is an expert's *predictive framework*—similar to a mental model or schema in other problem-solving literature. The predictive framework is a mental representation of the problem's key features and the relationships between them, which allows the experts to make predictions and explain observations. A predictive framework has three key features: (1) it allows the expert to identify important problem elements and eliminate unimportant elements; (2) it allows the experts to explain relationships between these elements, which includes some degree of mechanistic reasoning; (3) the predictive framework is detailed enough that the expert is able to conduct thought experiments by manipulating important variables.

In light of the work of Price et al., we sought to develop an assessment of engineering problem-solving by posing a problem that would require the solver to make some of the same decisions

that an expert problem-solver makes. Textbook problems are not suited to this task because the expert decisions are often made for the solver—for example, assumptions are almost always given to the solver in physics problems, rather than allowing the solver to identify appropriate assumptions or simplifications. We found that a troubleshooting task, such as critiquing a flawed product design schematic, was well-suited for this assessment, as it requires the solver to make many of the expert decisions [19]. The general structure for the assessment, which may be applied in any science and engineering discipline, is depicted in Fig. 1. Here we create and test a chemical engineering problem-solving assessment based on this design.

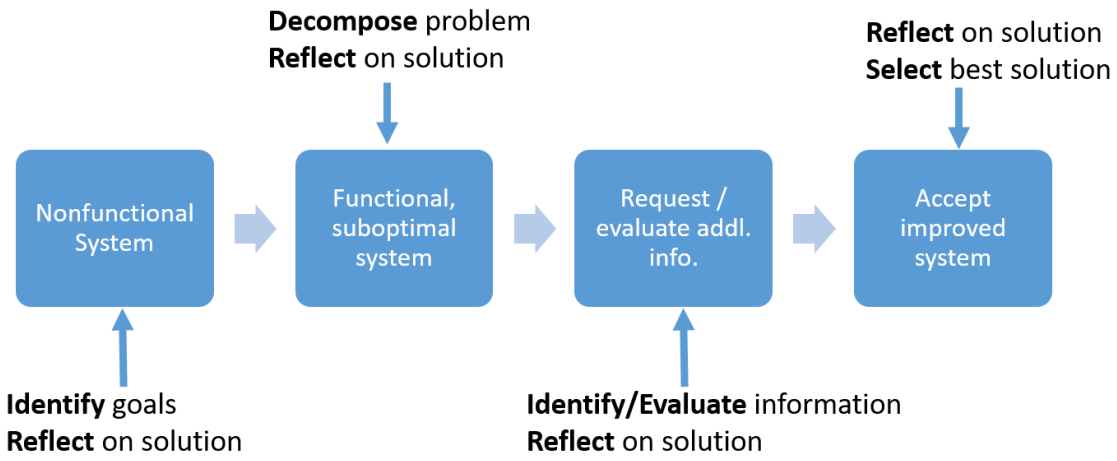

Figure 1: Outline of assessment design. In the first stage students are shown a nonfunctional system and asked to identify the criteria on which the system should be evaluated, requiring them to identify the goals of the problem and reflect on the solution. They are then shown a corrected system that is suboptimal and asked a series of increasingly detailed questions on how they will evaluate it. In the third stage they are asked what information they want to evaluate the system and how to use that information. Finally they are shown an optimized solution and must decide whether or not they will accept the proposed changes based on all the data they have gathered throughout the assessment.

The assessment consists of 10 free-response questions concerning the analysis of a chemical block flow diagram (Fig. 2) Participants are given some background information on the chemistry of the process, and then they are shown a block flow diagram that contains several errors (e.g. mass accumulation loops). They are asked what criteria they would use to evaluate the process, whether the process meets those criteria, and whether the process is physically feasible. If they indicate the process is not feasible, they are asked what modifications need to be made to make it feasible. Participants are then shown a corrected block flow diagram that represents a functional but suboptimal chemical process. They are asked what feedback they would give the designer of the process, whether the process was optimal with respect to mass and energy consumption (if not, what modifications need to be made), and what other information they would request about the process. Participants are then shown a list of 14 pieces of information relevant to the process, and asked to decide whether each piece is essential, secondary, or a minor detail. They are then asked what modifications they would make to the process based on the most essential pieces of information. Finally, students are shown a more optimal block flow diagram and asked whether they would accept the proposed changes and why. They are asked if they have any safety concerns about the process, and then asked to summarize all of the changes they would make to the original block flow diagram shown.

Throughout the assessment the questions progress from more generic questions (e.g. criteria for evaluation) to more specific questions (e.g. safety issues and optimal material consumption). The assessment takes approximately 60 minutes to complete, and is administered online in Qualtrics. Participants are not allowed to go back and modify responses to earlier questions when completing the assessment.

We previously conducted a pilot study of this assessment to determine whether it reliably measures differences between more and less expert-like reasoning [19]. To evaluate students on problem-solving, we compared their assessment responses with a responses collected from experts. In that study we found that students who had completed the capstone design course were more expert-like in their problem-solving than other students. Notably, we found no differences in problem-solving between students in their first, second, or third years of study after they had completed an introductory course in chemical engineering design. This suggests that the capstone design course was important for students to learn problem-solving, but it raises questions about how well these skills are developed in other core coursework.

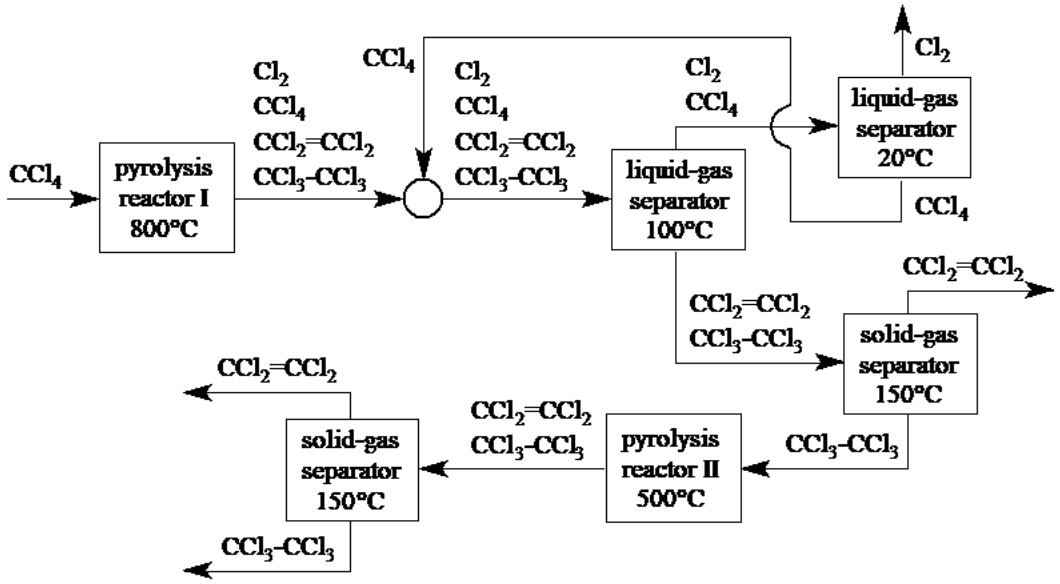

Figure 2: Flawed chemical block-flow diagram from the first part of the problem-solving assessment.

The previous study was quite limited in scope. In this study, we sought to test how well the assessment functions as a pre/post measure in introductory design and process analysis courses at two universities. Our research questions were: (1) do we see changes in students' problem-solving during the introductory chemical engineering design course? (2) Does the assessment provide reliable measures of specific decisions associated with expert problem-solving when administered in a low-stakes, pre/post format that is typical for inventories of conceptual understanding in biology and physics?

Methods:

Data consisted of student responses to the problem-solving assessment collected from introductory chemical process design courses at two large research universities. At University 1, the course covered the basics of mathematical and graphical analysis of chemical processes, dynamic scaling, and design choices related to process efficiency, product quality, economics, safety, and environmental considerations. The course consisted of one one-hour lecture and one 90-minute discussion section per week for 15 weeks. The course participants were freshmen or transfer students (typically juniors transferring from community colleges) intending to major in chemistry or chemical engineering. Students majoring in chemical engineering are required to take a second, more complete course in chemical process analysis following this introduction to design. Students completed the problem-solving pre-assessment in week 3 and the post-assessment in week 15 for extra credit points in the course. In total, 12 of the 62 students enrolled completed both the pre- and post-assessment. At University 2, the course was a sophomore-level course in chemical process design and analysis. It covered the same topics as the course at University 1, but in considerably more depth. The course consisted of three one-hour lectures and one two-hour discussion section per week for 15 weeks. The course participants were largely students intending to major in chemical engineering. Students completed the problem-solving pre-assessment in week 4 and the post-assessment in week 15 for participation credit in the course. Complete pre- and post-assessments were submitted by 32 of the 54 students enrolled in the course. We did not have any background information to determine whether the students who completed the survey had similar GPAs as other students in the course.

Dimension being evaluated	**Maximum Score**
(1) Important features (flaws/improvements noted in design)	11
(2) Criteria (number of expert criteria identified)	11
(3) Use of information (how ranking of information compares with experts)	14
Composite (Total=(1) + (2)/14 + (3)/11)	13

Table 1: Summary of scoring rubric for student responses.

The pre- and post-responses were coded along several dimensions that reflect how expert-like students' reasoning was. This included (1) how many important features of the design students noticed (flaws and improvements in the designs), (2) how students selected criteria and (3) used information to evaluate the design. Students were assigned a score on each metric (maximum scores given in Table 1) and then a composite score that was the sum of the number of important features they noticed, plus the percentage of criteria and information identified that experts also identified.

1. Responses were coded for which mistakes students noted in the original process, as well as what improvements students suggested or accepted to the process in their responses. There were four flaws in the original block flow diagram, and at least seven opportunities for improvement to the process, depending on how the students decided to use the information that was given to them. The maximum score was 11 points.

2. We also coded the responses to two questions for how closely student responses matched expert responses collected previously [19]. We first coded the criteria the students used to evaluate their process and how it matched the criteria listed by experts in the previous study. There were 11 expert criteria so students could score a maximum of 11 points.
3. We coded how students ranked each piece of information given to them and how that compared to expert responses. Students were given one point for each piece of information they ranked the same way the experts did for a maximum of 14 points.

We assigned a numerical "composite score" to each student based on the total number of flaws and improvements they noted in their responses, as well as how closely their information rankings and criteria listings matched the expert responses (i.e information score and criteria score were converted to fractions and added to the important features score). The flaws and improvements were weighted more heavily because identifying important features of the problem is one of the most essential elements of expert problem-solving. The maximum possible composite score was 13 points and the minimum was zero points. We also coded patterns of shortcomings in student solutions, but did not factor these into the composite score. Shortcomings included mixing up chemical species (e.g. what is a byproduct, what is the desired product, etc.), suggesting modifications to the process that would make it worse, and giving contradictory responses (e.g. suggesting that a compound be recycled while also suggesting the unit from which it comes should be removed).

Results:

The composite scores for students at both universities at pre-test and post-test are plotted in Fig. 3. The average pre-test composite score at University 1 was 3.6 points. There was a 0.23 point increase in scores from pre-test to post-test at University 1. The pre-scores at University 2 were 1.6 points higher than the pre-scores at University 1. At University 2 the scores actually decreased by 0.87 points from pre-test to post-test.

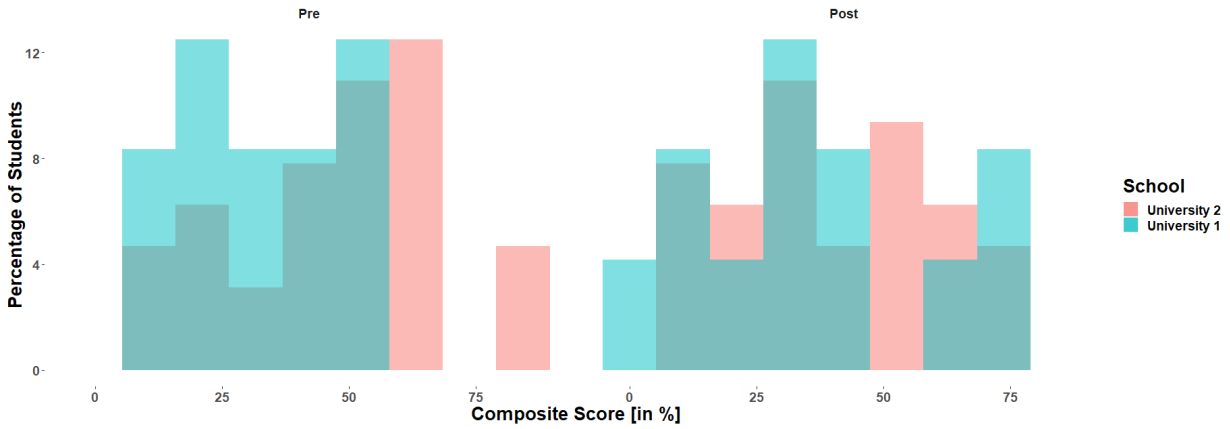

Figure 3: Histogram of composite scores for all students at both timepoints.

Criteria Score	Pre-test	Post-test
University 1	16%	15%
University 2	17%	18%

Table 2: Criteria score for students at both universities at pre- and post-test.

The criteria scores for all students at pre-test and post-test are summarized in Table 3. The experts listed 11 total criteria according to which they would evaluate this process, so the maximum score was 11 points. The average pre-test criteria score at University 1 was 1.8 points, meaning students listed approximately 2 of the 11 criteria that experts did at pre-test at that school. At post-test, students from university 1 listed 1.6 expert criteria on average. Students at University 2 listed 1.9 expert criteria on average at pre-test, and 2.0 criteria at post-test.

Important Information Score	Pre-test	Post-test
University 1	40%	31%
University 2	36%	39%

Table 3: Important information scores for universities 1 and 2 and pre- and post-test.

The information scores for all students at pre-test and post-test are summarized in Table 2. The average pre-test information score at University 1 was 5.6 points (40%), meaning that on average students categorized 5.6 pieces of information in the same way that experts categorized them (maximum score is 14 points). At post-test, students at University 1 only categorized 4.4 pieces of information the same way that experts did (31%). Students at University 2 had pre-test information scores of 5.1 points (36%), and post-test information scores of 5.5 points (39%).

We conducted a quantitative analysis of the students composite scores, information scores, and criteria scores using linear mixed effects models [20] to determine if there were differences between universities and between pre- and post-test. The mathematical model is:

$$\boldsymbol{Score} = \beta_0 + \beta_1 \boldsymbol{Time} + \beta_2 \boldsymbol{School} + \beta_3 \boldsymbol{Time} \times \boldsymbol{School} + \gamma_1 \boldsymbol{Z}$$

where **Time** is a binary variable that is 0 at pre-test and 1 at post-test, and **School** is a binary variable that is 0 for University 1 and 1 for University 2. The coefficient β_0 is thus the average pre-score at University 1, β_1 is the change in scores at University 1, $\beta_0 + \beta_2$ is the pre-score at University 2, and $\beta_1 + \beta_3$ is the change in scores at University 2. The unobserved random variations between students are modeled by the variable **Z**. As a measure of model quality we compute both the marginal and conditional R-squared, which tells you the proportion of variation in scores explained by the fixed effects (**Time, School, Time x School**) and the fixed effects plus the random effects (**Z**), respectively. The results for all three models are in Table 4.

	Composite Score	Information Score	Criteria Score
Intercept (β_0)	3.6 ± 0.62 ***	5.6 ± 0.50 ***	1.8 ± 0.27 ***
Time (β_1)	0.23 ± 0.73	-1.2 ± 0.64 †	-0.25 ± 0.32
School (β_2)	1.6 ± 0.73*	-0.46 ± 0.59	0.063 ± 0.31
Time X School (β_3)	-1.1 ± 0.86	0.82 ± 0.75	0.16 ± 0.37
Marginal R^2	0.072	0.037	0.011
Conditional R^2	0.35	0.22	0.31

*Table 4: Results of mixed-effects models of proximity, information and criteria scores as a function of institution and time the test was taken. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.10$.*

Most of the differences observed between pre- and post-test were not statistically significant (see Table 4). The one exception is that the decrease in important information scores at University 1 is marginally significant. There was one statistically significant difference between universities: the pre-scores at University 2 were significantly higher than the pre-scores at University 1. Overall, time and school explain only 1-7% of the variation in composite, information and criteria scores, whereas between student variations account for 20-30% of this variance. This suggests that most of the variations in scores are likely not due to changes over time, or which school the students attend.

Shortcoming	Pre-test	Post-test
Failure to Reflect	18	16
Superficial Feedback	5	0
Unknown/Incorrect Information or Assumptions	17	8

Table 4: Shortcomings in student responses at pre- and post-test.

The common shortcomings among students solutions not accounted for in the composite score, information score, or criteria score are listed in Table 4. The most common shortcoming was a failure to adequately reflect on the solution or problem-solving process, with 18 students doing this at pre-test and 16 at post-test. A common manifestation of this was for students to suggest modifications to the block flow diagram that would make the process worse. For example, some students suggested removing a solid-gas separator which would result in an unfavorable chemical equilibrium in the following reactor. Other students suggested removing the second reactor altogether, substantially limiting the amount of valuable product that was produced. We consider this a failure to reflect because, if the students had paused to consider the consequences of this modification, they would realize that they were making the process worse (they possessed the relevant content knowledge to make that decisions). The second-most common shortcoming was for students to not know or misuse important information or assumptions. For example, students would misread the table of physical properties and say that one of the solid-gas separators should be a liquid-gas separator, even though it is indicated in the information given to them that the compound to be separated sublimates. We saw 17 students not knowing or misusing information at pre-test, but this dropped substantially at post-test. At pre-test, five students gave superficial feedback regarding how the process flow diagram was drawn at pre-test; students did not focus on these superficial features at post-test.

Discussion:

We tested how well an assessment of problem-solving could be used as a measure of changes in students' problem solving skills during an introductory chemical process design course. Our research questions concerned whether there were measurable changes in problem-solving during the introductory course in chemical process design and analysis, and whether we could obtain reliable measures of problem-solving when administering this assessment as a low-stakes, pre/post assessment format typically used to administer concept inventories in other disciplines.

At university 1 we found a very small increase in composite scores (2.5%) from pre- to post-test, whereas at university 2 we found a 10% *decrease* in scores from pre to post-test, though neither of these changes were statistically significant. Students used a narrow range of criteria to evaluate the block flow diagram at both pre- and post-test. Students' most common criteria were the complexity of the process (the number of units used to accomplish a particular outcome), the yield of product, whether the separation temperatures were correct, and process economics. While these are all important criteria for evaluating a process, they represent a small subset of the criteria cited by experts. For example, experts cited process safety measures, failure modes, metallurgical constraints, and quantitative stream compositions and flow rates as important criteria, but students almost never mentioned these when evaluating the process.

There was more agreement between students and experts on what information given to them was important to the process, though the agreement was still less than 50%. For example, students thought that the price of electricity at the plant and the cost of high-density solid pumps were more important pieces of information than the experts did. Conversely, students thought the regulation of the reactants by the Montreal Protocol was very important, whereas experts typically considered this a minor consideration. There was no change in the criteria students used between pre- and post-test.

There are two explanations for the minimal change in students' reasoning from pre- to post-test. The first explanation is that students at neither university are learning to solve ill-structured problems the way that expert engineers do. Students could be learning the basic content knowledge associated with chemical process design (important design rules, mass and energy balances, cost analysis, etc.) which allows them to more easily read tables and more confidently understand what concepts the questions are targeting, but not be able to apply this knowledge to solving the problem. For example, the criteria listings and information rankings of students generally reflected a set of rules that are commonly taught in introductory design coursework. The criteria listed by experts, as well as how experts used various pieces of information, reflect their experience in designing real chemical processes where issues of safety and process failure are real concerns. Students may be implicitly or peripherally aware of these issues, but lack the experience and deep content knowledge needed to address these issues and apply this information.

The second explanation for the minimal change in scores and the decrease in information scores at University 1 is that the post-test responses are not valid because students were not taking the post-assessment seriously. It appears that the students at University 2 put less effort into their

post-test responses than they did the pre-test responses, as did some of the students at University 1. The responses were shorter and more vague, listing general principles they knew to be true (e.g. byproducts should be recycled, cost should be minimized), but lacking concrete applications to the problem at hand. Indeed, we found that students spent substantially less time on the post-test assessment (difference in median time = 18 minutes, $p = 0.05$), and that time spent was marginally correlated with composite score (Kendall's $\tau = 0.12$, $p = 0.09$). This suggests that participation credit or extra credit may not be sufficient to elicit reliable responses at post-test when students are often pre-occupied with final exams and other end-of-term responsibilities. The assessment requires substantially more time and thought than the concept inventories typically used in physics and biology. For example, the Force Concept Inventory [21] consists of 30 multiple choice questions about the relationship between force and motion, and takes about 15 minutes to complete. In contrast, this assessment has 10 free-response questions and takes about 60 minutes to complete if putting in a earnest effort. Ranking the various information given is particularly time-consuming, which could explain the decrease in information scores at University 1. This suggests that we may not have reliable post-test results, hindering our ability to measure changes in problem-solving with this assessment.

There is a statistically significant difference between composite pre-scores at the two institutions studied here. The pre-test responses were generally thoughtful and detailed, so we take them to be valid measures of students' reasoning and hypothesize that this difference may be due to population differences. The students from University 1 are largely freshman and transfer students who may not yet be committed to majoring in chemical engineering, whereas the students from university 2 are sophomores intending to declare a chemical engineering major. The students at University 2 may thus have been more invested in the assessment (at least at pre-test). Indeed, the participation rate at University 2 was three times that at University 1. Furthermore, many of the participants from University 1 were transfer students from community colleges, and thus have significantly different academic backgrounds from the traditional students at both universities.

Conclusions:

The results of this study suggest that care is needed when administering this assessment of problem-solving to measure changes over time. Namely, it is important that the assignment be perceived to have some value to the students beyond participation credit or extra credit. In another study, we were able to collect reliable post-test results by administering the assessment as part of the homework assigned in a senior design course. While there are differences between senior and introductory students, we hypothesize that administering this assessment as an actual assignment for the course would make the responses substantially more reliable.

In an upcoming study, we are testing this hypothesis to see if we can obtain more reliable results from introductory students. We are administering the assessment as an assignment early and late in the same course at University 1 to see if there are measurable pre/post differences in students' problem-solving. This will be used as a control group for studying an intervention designed to teach problem-solving. The intervention consists of a worksheet that students complete when they are doing a design exercise in the course. They are asked a number of questions that require

them to plan out their approach for solving the problem, and then reflect on their solution once they have reached it. Salehi and Wieman have shown that this leads to improved problem-solving that may even transfer to different contexts [22].

Another way to increase the reliability of the assessment is to make it shorter. We found that a number of questions provided essentially redundant information. Namely, we did not get substantially different answers when students were asked to give feedback on the corrected design in part 2 of the assessment, and when they were asked if it was optimal with respect to mass and energy balances. This suggests that we may be able to reduce three questions in Part 2 to a single question asking for feedback on the corrected design. Furthermore, we did not find much information in the final question where students were asked to summarize their plans for altering the original process. They would either be redundant with previous responses, or students would appear to put minimal effort into answering them.

Further refinement of this assessment will be an important advance in engineering education research, as it will provide a reliable way to measure problem-solving—an important, but not guaranteed outcome of an engineering education. It can be used to answer many different research questions, e.g. are there differences in outcomes between traditional capstone design courses and capstone courses that focus on experimental design? It can also be used as an assessment tool for various chemical engineering departments to decide whether their undergraduate programs are adequately preparing students for the workplace. Improving our ability to measure problem-solving is an important step in being able to improve the way we teach problem-solving to undergraduate students and prepare them for engineering careers.

References:

1. H. J. Passow, "Which ABET Competencies Do Engineering Graduates Find Most Important in their Work?," *Journal of Engineering Education*, vol 101, no. 1, pp. 95-118, 2012.
2. ABET Engineering Accreditation Commission. (2000). ABET criteria for accrediting engineering programs. Retrieved from https://www.abet.org/accreditation/accreditation-criteria/
3. National Research Council, Discipline Based Education Research. Washington D.C:National Academies Press, 2012.
4. D. H. Jonassen, J. Strobel, & C. B. Lee, "Everyday problem solving in engineering: Lessons for engineering educators." *Journal of Engineering Education*, vol 95, no. 2, pp. 1-14, 2006.
5. Larkin, J. McDermott, J., Simon, D. P., & Simon, H. A. (1980). Expert and novice performance in solving phyiscs problems. Science, 208(4450), 1335-1342.
6. M. T. H. Chi, P. J. Feltovich, and R. Glaser, Categorization and representation of physics problems by experts and novices, Cognitive Science 5, 121 (1981).
7. J. I. Heller and F. Reif, Prescribing effective human problem-solving processes: Problem description in physics, Cognition and Instruction 1, 177 (1984), https://doi.org/10.1207/s1532690xci01022.
8. W. K. Adams and C. E. Wieman, Analyzing the many skills involved in solving complex physics problems, American Journal of Physics 83, 459 (2015), https://doi.org/10.1119/1.4913923.
9. C. Wieman, Comparative cognitive task analyses of experimental science and instructional laboratory courses, The Physics Teacher 53, 349 (2015), https://doi.org/10.1119/1.4928349.
10. Hsu, L., Brewe, E., Foster, T.M., & Harper, K.A. (2004). Resource letter RPS-1: Research in problem solving. American Journal of Physics 72, 1147. http://dx.doi.org/10.1119/1.1763175
11. M. P. Čančula, G. Planinšič, and E. Etkina, Analyzing patterns in experts' approaches to solving experimental problems, American Journal of Physics 83, 366 (2015), https://doi.org/10.1119/1.4913528.

12. D. Jonassen, “Engineers as Problem Solvers,” in Cambridge Handbook of Engineering Education Research, A. Johri and B. Olds, Eds. Cambridge: Cambridge University Press, 2014, pp. 103-118.
13. C. Atman, et al., “Engineering Design Processes: A Comparison of Students and Expert Practitioners,” Journal of Engineering Education, vol. 96, issue 4, pp. 359-379, Oct 2007.
14. D. Woods, “An Evidence -Based Strategy for Problem Solving.” *Journal of Engineering Education*, vol. XX, iss. XX, pp. 443-459, 2000.
15. N. S. Hong, D. H. Jonassen, & S. McGee, “Predictors of well-structured and ill-structured problem solving in an astronomy simulation.” *Journal of Research in Science Teaching*, vol. 40, iss. 1, pp. 6-33, 2003.
16. A. M. Price, C. Kim, E. Burkholder, M. Flynn, A. Fritz, and C. E. Wieman, Identifying expert problem solving decisions (2019), unpublished.
17. D. H. Jonassen, “Designing for decision making.” *Educational Technology: Research and Development* , vol 60, pp 341-359, 2012.
18. B. Means, E. Salas, B. Crandall, * T. O. Jacobs, “Training decision makers for the real world,” in Decision making in action: Models and Methods, G. A. Klein, J. Orasanu, R. Calderwood & C. E. Zsambok, Eds. Ablex: Norwwod,NJ, 1993, pp 306-326.
19. E. Burkholder, A. Price, M. Flynn, and C. Wieman, Assessing problem-solving in science and engineering programs, *Phys. Educ. Res. Conf. Proceedings*, Provo, UT, 2019.
20. Stock, J. H., & Watson., M. W. (2015). *Introduction to Econometrics* (3rd ed.). Boston: Pearson.
21. D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, Phys. Teach. **30** (3), 141 (1992).
22. S. Salehi and C. E. Wieman (2020), A Problem-solving Framework: Characterizing Actions and Decisions Involved in Solving a Complex Problem, manuscript in preparation