

Assessing problem-solving in science and engineering programs

Eric Burkholder and Argenta Price

Department of Physics, Stanford University, 382 Via Pueblo, Stanford, CA, USA 94305

Michael Flynn

Department of Mechanical Engineering, Stanford University, 440 Escondido Mall, Stanford, CA, USA 94305

Carl Wieman

*Department of Physics, Stanford University, 382 Via Pueblo, Stanford, CA, USA 94305 and
Graduate School of Education, Stanford University, 485 Lasuen Mall, Stanford, CA, USA 94305*

Science and engineering (S & E) programs frequently claim that they teach undergraduate students how to be good problem solvers. However, there has been little research to-date that demonstrates this, in no small part due to the fact that measuring problem-solving is quite difficult. Our recent work characterized experts' problem-solving as carrying out a series of several dozen decisions made in solving novel problems; these decisions are remarkably consistent across S & E disciplines. Based on this, we developed a template for an assessment that measures students' problem-solving skills by posing questions that require them to make a subset of these expert decisions in suitable contexts. Preliminary results show that the assessment captures a wide range of problem-solving skills among students, and exposes key weaknesses in problem-solving. Most importantly, the data show that students' predictive frameworks—their mental models of a system's key features and the relationships between them—are less robust than experts', limiting students' ability to make predictions and explain observations while problem-solving. We provide detailed results from a pilot-test of this assessment in the context of chemical engineering design, which applies many concepts from physics. These results point to a general deficiency in undergraduate S & E programs: students are not being given the opportunity to practice expert decision-making, and thus do not develop robust predictive frameworks by the end of their undergraduate programs.

I. INTRODUCTION

Among the desired outcomes of an undergraduate education in science and engineering (S & E) is the ability to solve problems [1]. Indeed, recent graduates cite problem-solving as one of the most important skills needed to be successful in their careers [2]. It is widely held that students in many S & E disciplines are trained to be good problem-solvers, but very little research exists to show that this is the case [3]. There have been some studies that aim to characterize problem-solving [4–10], but there have been few attempts to measure student problem-solving *performance* [11]. Furthermore, there is still some ambiguity as to what expert problem-solving even looks like.

Previous work in expert problem-solving—which has been studied mainly in the context of physics—has found that experts categorize textbook problems according to their deep conceptual structure, whereas novices categorize them by surface features [5]. While this provides insight into expert-novice differences in knowledge organization, it does not tell us anything about how experts solve novel problems; these “exercises” are not truly problems for the expert (we elaborate on this distinction in the next section). More recently, expert thinking and problem-solving in physics has been characterized using cognitive task analyses [11–13].

In a qualitative study based on the critical decision method of cognitive task analysis [14], we identified a set of several dozen decisions (see Fig. 1 for some examples) that experts make as they solve novel problems in their work [15]. These decisions are consistent across physics and other sciences, as well as multiple engineering fields. Underlying most of these decisions are the experts’ “predictive frameworks,” mental models of a system’s key features and the relationships between them that can be used to make predictions and explain observations, and can adapt to new data and information. An example of such a framework from condensed matter physics is conservation of probability in colloidal suspensions: key features include various processes that give rise to species fluxes in a suspension such as Brownian motion, chemical reaction, and advection with bulk fluid flow. These features are connected by a conservation equation for the particle probability distribution. One can use this equation to make predictions about how complex fluids flow under various conditions.

This theoretical framework of problem-solving as decision-making informs the question addressed in this paper: how do we measure problem-solving? We propose that we can measure problem-solving by having students complete an authentic task—one that requires them to make some of the same decisions as an expert in their discipline would make while doing an authentic problem. Here, we describe in detail the development and pilot-testing of a template for such tasks that applies generally in physics and other S & E disciplines.

II. METHODS

To assess how well students are thinking like experts requires an authentic problem—one that forces them to make the same decisions as experts. Most textbook “problems” rob the students of the opportunity for authentic decision-making, and are more accurately called exercises. Particularly noticeable in textbook exercises are the lack of opportunities for students to reflect on their solutions and problem-solving process or strategy. Experts, in contrast, reflect quite frequently, often in the context of “troubleshooting”—trying to figure out why a system of interest is not functioning as expected, and/or how an existing system could be improved. The system of interest varies by scientific discipline: in physics it might be an optical trap, in medicine it is often a sick human being, and in engineering it might be a design for a product or process. The decision to troubleshoot, however, is nearly universal, making it a reasonable context for assessing a person’s ability to make expert decisions.

We designed a four-part template to assess troubleshooting in a generic way; its structure is depicted in Fig. 1. Each part of the template consists of multiple questions that require students to make a specific subset of expert decisions, which are listed in Fig. 1.

1. Students are first shown a representation of an existing system; they are not told whether the system is functioning properly, and are given incomplete information. They are then asked to decide what criteria they would use to evaluate the system’s function, and decide whether the test system they have been given meets those criteria—if it is functioning. If it is not, they are asked to decide what modifications are necessary to make it function.
2. Students are then shown a representation of the same system that has been modified such that it meets some minimum standard of functionality. They are asked what further feedback they would have about the function of the system, and then asked to decide if the system meets certain discipline-specific criteria for optimal functioning. They are then asked how they would modify the system such that its function is improved.
3. In the third part, students are asked what additional information they would request to evaluate the system. Students are then given a list of relevant information, and asked to rank the importance of each piece of information given to them. They are finally asked how they would use the most important information to modify the system of interest.
4. They are presented with a representation of a modified system whose function is likely to be superior to the system seen in Part 2. These modifications are presented as suggestions from a colleague, and the students are asked whether they would accept these suggestions. Finally, they are asked to summarize any and all changes they would make to the original system.

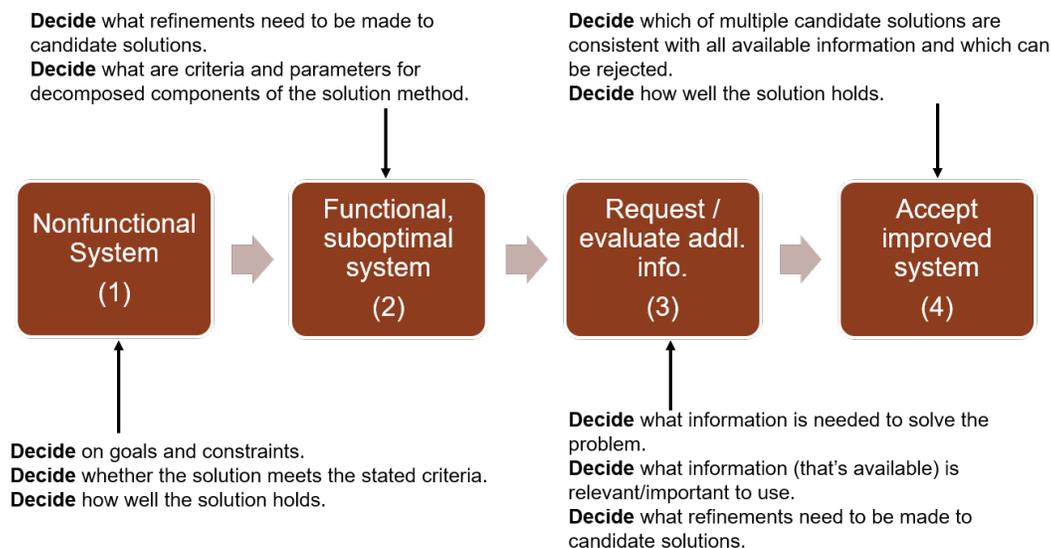


FIG. 1. Illustration of the assessment structure used to evaluate students' ability to make expert decisions in problem solving. Expert decisions are given for each step.

We argue this template can be applied to relevant problems in many S & E disciplines, and we have already done so in mechanical engineering and chemical engineering—we discuss the latter in this paper. In a pilot study, we adapted a chemical process analysis exercise from an introductory chemical engineering textbook to fit this template [16]. The context of the problem is that the student is an engineer who has assigned a summer intern to design a chemical process that produces tetrachloroethylene via pyrolysis of carbon tetrachloride (the system); the student is presented with their intern's preliminary process flow diagram (the representation). They are given a table containing selected physical properties of each chemical species, which they may reference throughout the troubleshooting process. One typical goal of these tasks is to design the process in the most cost-effective way possible. While these are not problems typically seen by physicists, they rely upon content knowledge and applications familiar to most physicists: conservation of mass (species) and energy, thermodynamics, and basic chemistry. The predictive frameworks tested by this assessment are efficient use of energy and material, and economic viability of chemical processes.

The assessment was administered in a think-aloud interview format with respondents filling in answers via Qualtrics. Respondents were given 60 minutes to complete the 10 questions. Respondents were undergraduate students studying chemical engineering and expert engineers that teach senior level design courses at two highly selective U.S. universities (#1 and #2). The students' level of experience in the domain of chemical process design is described in Table I.

All assessment responses were coded using an emergent scheme that focused on two areas:

1. Expert-student differences in criteria used to evaluate

TABLE I. Respondent experience in chemical process design.

Design Experience	Univ. #1	Univ. #2
Intro. Design Course		
Years 1-3	19	8
Senior Design Course		
Year 4	2	12
Industry Job or Internship	2	6

the design (Part 1 in Fig. 1), and what information was considered important (Part 3 in Fig 1).

2. Overall quality of (student) decision making based on the mistakes they noted in the original process flow diagram and the improvements they suggested or accepted to the modified process flow diagram. We also coded all 10 questions for "shortcomings" in their responses. Shortcomings fell into two categories: lack of content knowledge and deficiencies in predictive frameworks.

We assigned a numerical score to the student solutions that was the sum of the number of mistakes correctly noticed in the original design and (good) improvements [17] suggested to the functional design, minus the number of shortcomings coded in their responses. We take this score to be a proxy for the overall quality of student decision making. Differences in score by design experience were analyzed using the non-parametric Mann-Whitney U-test. The effect size reported for between-group differences is the Hodges-Lehmann estimator—the difference in pseudo-medians of two distributions [18].

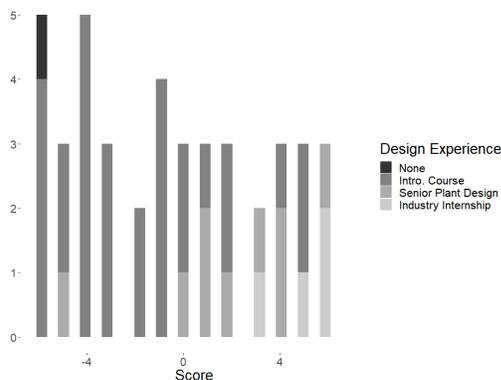


FIG. 2. Histogram of student scores on the assessment. The score was computed as the number of errors and improvements the students noted minus the number of shortcomings in their responses. Shortcomings indicated either a lack of content knowledge or a deficiency in the student’s predictive framework.

III. RESULTS

The results of the pilot study show weaknesses in students’ problem solving that we expect will be applicable in physics and across S& E disciplines as well. There were clear expert-student differences in the sophistication of their predictive frameworks, and their ability to use those frameworks to connect information to decisions about the solution. In particular, experts’ predictive frameworks had different key features than the students’ frameworks. Students focused on technical details such as stoichiometry and stream composition, or fundamental physical balances (e.g. making sure mass is conserved). Meanwhile, experts expressed concerns about process safety and operating conditions, which required predictive frameworks that included connections between the fundamental physics/chemistry and outcomes in the manufacturing process:

- “Exotherms and endotherms - Defines degrees of risk Potential runaways - potential for fires, explosions, loss of containment”
- “Very corrosive environment. Metallurgy constrains [sic] vis a vis corrosivity and temperature.”

These responses represent more sophisticated predictive frameworks that focus on more important features of the design [19].

Interestingly, students and experts were relatively consistent in which pieces of provided information they thought were important. The main difference here was the ability to use this information to modify their solution as guided by their predictive framework. Students frequently offered no concrete suggestions to account for the important information in their improvements to the process flow diagram. Our work adds to other work that previously noted expert-student differences in knowledge structure [5], by showing that a key difference is in how students and experts use this knowledge.

A histogram showing the distribution of student scores by

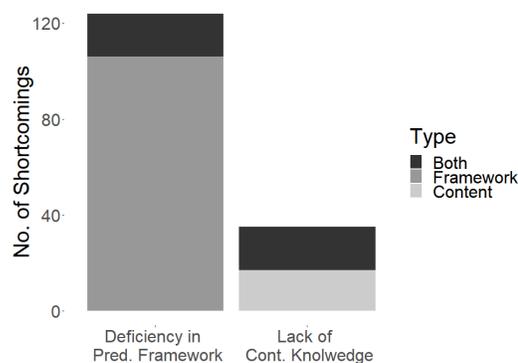


FIG. 3. Shortcomings in student responses coded by type: lack of content knowledge or a deficiency in the student’s predictive framework. Shortcomings that fall into both categories are added to both bars.

prior design experience is shown in Fig. 2. Overall, students who had taken a senior-level course in chemical plant design outperformed students who had only taken the introductory course in process design ($HLE = 4, p = 0.03$). Students who had completed an industry internship where they did some design work in addition to the senior design course outperformed the introductory-level students by a larger margin ($HLE = 7, p = 0.02$). The difference between seniors who did or did not have industry experience was not statistically significant ($HLE = 3, p = 0.12$); this is due to the small sample size (4). Notably, there was no significant difference in performance by year in college (1-3) for students who had only taken the introductory design course ($p > 0.5$ for all pair-wise differences).

Because the scores showed no difference between students in their first, second, or third year of college, we analyzed the shortcomings in students’ responses to determine if this could be explained by missing content knowledge students should learn in the fourth-year course. We coded the shortcomings into two categories:

- Lack of content knowledge, e.g. missing a sense of scale for large volume chemical processes
- Deficiencies in predictive frameworks (elaborated on below).

Some shortcomings reflected an inability to incorporate relevant conceptual knowledge into the problem solution, and were coded in both categories.

Only 17 of the 140 total shortcomings were reflective of missing content knowledge. The missing items are typically taught in the introductory design course, but it is possible that some students may have forgotten this content or not learned it until the senior design course.

Another 18 shortcomings were characterized as both lack of content knowledge and deficiency in the predictive framework. These were typically instances demonstrated some superficial knowledge but lacked sufficient detail and struggled to apply it:

“I still think some processes could be rearranged so the temperature ranges are not that vast, but I am not sure at the moment how to do that.”

“I think some of the processes seem to lack energy conservation.”

If the predictive framework lacked causal relationships between key features, the student was unable to make predictions or explain observations.

The remaining 105 shortcomings were coded as deficiencies in student predictive frameworks. Such deficiencies are evident when students are able to correctly invoke a particular concept or identify an important feature, but are unable to apply the knowledge appropriately in the problem context. One class of such deficiencies was inconsistency in answers—not using the identified criteria to evaluate the solution, or giving inconsistent suggestions to improve the system. For example, many students suggested removing one of the reactors to reduce cost (at the expense of yield), then later suggested recycling a stream of intermediate to the same reactor they wanted to remove. The second class of predictive framework shortcomings reflected insufficient connections in their predictive framework between formal knowledge (e.g. chlorine gas is corrosive) and the impact that would have in the context of the problem. For example, many students did not accept the proposed improvements to the process in Part 4 because it introduces chlorine gas, which is corrosive. However, there was already chlorine in the process, and the total amount was unlikely change as a result of this modification.

There is a floor effect for students who have completed only the introductory design course (see Fig. 2). Students in the senior design course had more developed predictive frameworks (fewer shortcomings based on predictive framework deficiencies), possibly because the senior course provided them with more opportunities to practice making decisions and using their predictive frameworks.

IV. CONCLUSIONS & FUTURE WORK

In all, the results indicate that, while students are learning appropriate content knowledge in their discipline, they are likely not being given the opportunity to practice making expert decisions in their coursework. Thus, they do not develop robust predictive frameworks that allow them to *use* their content knowledge to make meaningful decisions—which they certainly need to do as practicing scientists and engineers.

The next step in analysis of the current data includes coding student responses for the quality of individual decisions made—the scores presented in Fig. 2 are only a proxy for evidence of the students’ general decision-making skills. More finely resolved coding will allow us to see whether there are particular decisions students have trouble making, and if so, where educational emphasis should be placed.

Note that Figs. 2 and 3 only contain analysis of *student* responses. In the pilot-testing, we found that the context of

the exercise was crucial in determining experts’ responses. First, the assessment originally said that an engineer had produced the initial design, causing the experts (who managed engineers in their careers) to assume a certain level of competence, i.e. that basic physical laws were respected. Second, proper analysis of chemical processes requires far more data than respondents were given. Experts refused to answer many of the questions without more technical information; this was a marker of their expertise, but not captured by the scoring system we implemented. We opted not to provide the students with this information, as it would make the task too difficult.

In the next phase of this project, we will use our assessment to test the efficacy of an instructional intervention in the introductory design course at university #2. The intervention will be in the spirit of Salehi [20] and Holmes et al. [21]: we will have students practice making expert decisions—with particular emphasis on reflection and planning—as they complete design exercises throughout the semester. Through this, we expect that students will develop more sophisticated predictive frameworks. This will be in the form of a worksheet accompanying their design exercises, and students will be graded in part on the quality of their decision-making processes.

We believe that our troubleshooting assessment template can be easily adapted to multiple disciplines. For example, we can adapt a problem suitable for physics students: determining whether a garden gate is tall enough to prevent a small dog from jumping over it. Following the structure from Fig 1: (1 - nonfunctional system) Students are given the height of the gate, distance between the house and the gate, and an approximate dog size. After evaluating whether the dog can jump the gate, they are asked what modifications would be required for the dog to be kept behind the gate. (2 - functional sub-optimal system) Students are shown a new scenario in which the dog is not able to jump over the gate (perhaps the dog is smaller). They determine how they might modify the gate so that it is robust to a wider range of dog and yard sizes. (3 - request/evaluate additional info) Students are given some key pieces of information such as the dog’s top speed. They are asked how this information would be relevant to evaluating the effectiveness of a gate. (4 - reflect on improved solution) They are presented with a gate that is unable to be jumped by most dogs and asked whether they would prefer this gate or not. We plan to test our template across a range of disciplines. We expect, and have preliminary data showing, that the failure to develop and effectively use predictive frameworks in making decisions is not unique to chemical engineering, and assessments in other disciplines would yield similar results.

ACKNOWLEDGMENTS

This project was funded by the Howard Hughes Medical Institute, Grant No. GT10478. The authors thank Candice Kim for her thoughts on the manuscript.

-
- [1] P. A. Koen and P. Kohli, Abet 2000: What are the most important criteria to the supervisors of new engineering undergraduates?, in *1998 Annual Conference (ASEE Conferences, Seattle, Washington, 1998)* <https://peer.asee.org/6903>.
- [2] H. J. Passow, Which abet competencies do engineering graduates find most important in their work?, *Journal of Engineering Education* **101**, 95 (2012), <https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2168-9830.2012.tb00043.x>.
- [3] N. R. Council, *Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering*, edited by S. R. Singer, N. R. Nielsen, and H. A. Schweingruber (The National Academies Press, Washington, DC, 2012).
- [4] J. H. Larkin and F. Reif, Understanding and teaching problem-solving in physics, *European Journal of Science Education* **1**, 191 (1979), <https://doi.org/10.1080/0140528790010208>.
- [5] 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), <https://onlinelibrary.wiley.com/doi/pdf/10.1207/s15516709cog0502>.
- [6] 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/s1532690xc0102>.
- [7] P. Heller, R. Keith, and S. Anderson, Teaching problem solving through cooperative grouping. part 1: Group versus individual problem solving, *American Journal of Physics* **60**, 627 (1992), <https://doi.org/10.1119/1.17117>.
- [8] D. R. Woods, An evidence-based strategy for problem solving, *Journal of Engineering Education* **89**, 443, <https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2168-9830.2000.tb00551.x>.
- [9] M. Prince and B. Hoyt, Helping students make the transition from novice to expert problem-solvers, in *32nd Annual Frontiers in Education* (2002).
- [10] T. Yildirim, M. Besterfield-Sacre, and L. Shuman, Scale development for engineering modeling self efficacy, in *2010 Annual Conference & Exposition*.
- [11] 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>.
- [12] 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>.
- [13] 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>.
- [14] R. R. Hoffman, B. Crandall, and N. Shadbolt, Use of the critical decision method to elicit expert knowledge: A case study in the methodology of cognitive task analysis, *Human Factors* **40**, 254 (1998), <https://doi.org/10.1518/001872098779480442>.
- [15] A. M. Price, C. Kim, E. Burkholder, M. Flynn, A. Fritz, and C. E. Wieman, Identifying expert problem solving decisions (2019), unpublished.
- [16] T. M. Duncan and J. A. Reimer, *Chemical Engineering Design and Analysis* (Cambridge University Press, 2019).
- [17] A "good" improvement was defined as one that would make the process more efficient with respect to energy and material consumption, or more economically viable.
- [18] E. Lehmann, *Nonparametrics: statistical methods based on ranks* (Springer, 1996).
- [19] K. A. Ericsson, R. R. Hoffman, A. Kozbelt, and A. M. Williams, *The Cambridge handbook of expertise and expert performance* (Cambridge University Press, 2018).
- [20] S. Salehi, *Improving problem-solving through reflection*, Ph.D. thesis, Stanford University (2018).
- [21] N. G. Holmes, B. Keep, and C. E. Wieman, Developing scientific decision-making and practices through structured student agency (2019), submitted.