Impact of decision-making in heat transfer courses on students' ability to solve authentic problems

1st Jiamin Zhang
Department of Physics
Auburn University
Auburn, United States
jiaminz@engr.ucr.edu

2nd Soheil Fatehiboroujeni
Department of Mechanical Engineering
Colorado State University
Fort Collins, United States
s.fatehiboroujeni@colostate.edu

3rd Matthew Ford
School of Engineering and Technology
University of Washington Tacoma
Tacoma, United States
mattford@uw.edu

4th Eric Burkholder
Department of Physics
Auburn University
Auburn, United States
ewb0026@auburn.edu

***Abstract*—This Research Full Paper presents a study of problem-solving skills in heat transfer courses. Although engineering programs stress the importance of teaching problem-solving skills, there are frequently reported gaps between the skills graduating engineers have and what employers want. Part of the reason for this gap is that many of the problems students solve, particularly in engineering science courses like thermodynamics and fluid mechanics, bear little resemblance to the authentic, unstructured problems they will be expected to solve as engineers. Previous work in problem-solving has been limited by a lack of a framework to describe how experts solve authentic problems, and a lack of assessments to measure authentic problem-solving. We have developed an assessment of problem-solving skills in the context of heat transfer to measure how well undergraduate engineering students are learning to solve authentic problems. We measured changes in students' problem-solving over the course of one chemical engineering heat transfer course and one biological engineering heat transfer course. Although students made improvements in some areas, the average student scores didn't change significantly for two-thirds of the questions on our assessment. Differences between the two courses can be explained by what decisions students practiced during the course. These results suggest that undergraduate students need more deliberate practice making the decisions that expert engineers do as they solve authentic problems. We hope to encourage other educators to use this assessment in their courses to measure how well they are preparing their students to solve real-world engineering problems.**

***Index Terms*—problem-solving, heat transfer, decision-making**

I. Introduction

Engineers are known as problem-solvers. In their work, they encounter ill-structured problems that require them to collect additional information, consider external constraints, and reflect on their solution process [1]–[6]. Although college graduates cite these problem-solving skills as the most important technical skills required of them in their everyday work [7], employers and researchers sometimes find that undergraduate students are not prepared for solving these kinds of problems when they graduate [8], [9].

One of the reasons for this skills gap is that the majority of problem-solving in traditional undergraduate engineering courses consists of solving textbook problems. Textbook problems require conceptual knowledge and can be quite challenging. However, textbook problems often already state assumptions and information needed and are designed to exercise a limited set of knowledge and skills. For example, Douglas *et al*. [10] examined end-of-chapter problems present in the two most commonly used statics textbook and found nearly all of the problems across these two textbooks were well-structured. Project-based learning, capstone design courses, and internship experiences do allow room for students to make problem-solving decisions, but these opportunities are often limited in undergraduate curricula.

If we want to teach undergraduate students to solve complex, real-world problems, we must be able to measure how well they are learning the necessary skills. However, real-world problem-solving skills are difficult to measure. Substantial work has been devoted to characterizing student and expert problem-solving in engineering [11]–[13] and physics [14]–[17], but there are almost no agreed-upon measures of problem solving [17].

Research on how experts solve authentic problems has revealed a framework that describes problem-solving as a set of 29 decisions-to-be-made, such as deciding what are the important underlying features of the problem, and what is the best solution [18]. How these decisions are made is highly context-dependent and requires deep disciplinary knowledge. Furthermore, our collaborators have developed a general template for assessing these decisions [19]. Using this framework and the general template, we have created an assessment of engineering problem solving in heat transfer [20].

In this work, we describe the use of our heat transfer assessment in two heat transfer courses, one in chemical engineering, and one in biological engineering. We sought

Howard Hughes Medical Institute

to answer the research question: how do students' problem-solving skills change as a result of one term of instruction in an engineering heat transfer course?

II. Methods

A. Assessment Design

We chose heat transfer as the context for the problem-solving assessment so that it would be useful to educators in a wide variety of engineering disciplines. When choosing the problem for the assessment, we wanted to ensure (1) the relevant physics includes the key concepts in heat transfer (e.g., convection and conduction), (2) any additional physics should be simple enough to explain in the assessment prompt, and (3) a professor who does research in heat transfer or someone from industry who regularly uses heat transfer to design systems should be "expert-enough" to solve the problem. The physical context we chose is the countercurrent heat exchange between arteries and veins in the human finger (Fig. 1 (a)). This countercurrent heat exchange mechanism reduces the degree to which the returning venous blood must be warmed by the core, at the expense of a lower average temperature in the extremities. The assessment starts by asking the participant how they would model the problem. Then, the participant is asked to answer three closed-response questions: the first question asks what assumptions from a given list of ten assumptions to make to simplify the problem. The second question asks what information is needed to solve the problem and asks participants to choose "definitely want", "might want" or "don't need" for 32 different pieces of information. The third question asks which variables the findings will be most sensitive to and participants can only choose five most important variables from a given list of nine variables. In the following text, we will refer to these questions as Assumptions, Information, and Sensitivity, respectively.

The solution space is too large for modeling the countercurrent heat exchange process in the finger to be included in open-response questions in the assessment. Thus, we used a "troubleshooting" design. We ask the participant to give feedback and note issues or missing features of two solutions proposed by their "colleague": an intentionally flawed solution (Fig. 1 (b)) and a revised solution (Fig. 1 (c)). Previous research has shown that such a task can still involve much the same expert problem-solving decisions and relevant content knowledge, but the solution space is more limited [17], [19]. Thus, the responses of both experts and students are straightforward to characterize and score. We then present the participant with experimental data and model predictions and ask them to give feedback on their colleague's validation plan. In the following text, we will refer to these groups of questions as Model 1, Model 2, and Experiment and Simulation, respectively. The assessment is delivered online via a Qualtrics survey and takes about one hour to complete. Details of the assessment can be found in our previous publication [20].

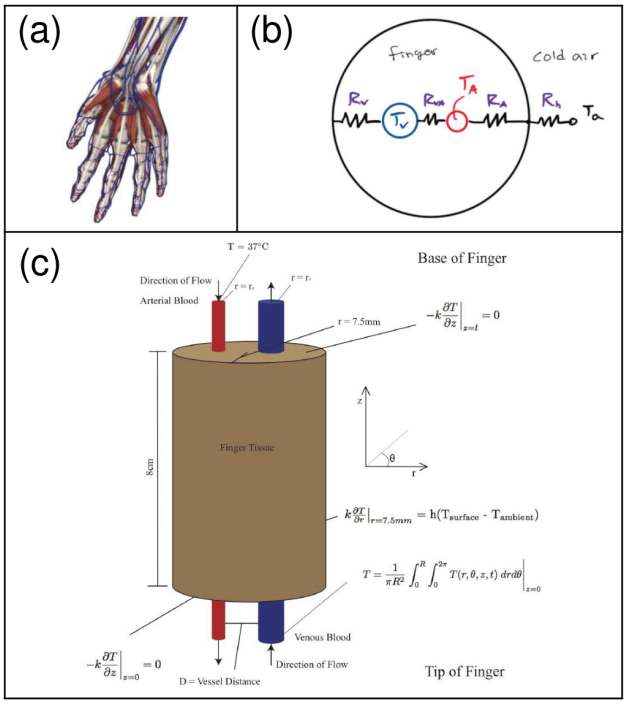

Fig. 1. Schematics in the heat transfer assessment (a) Blood vessels in a human hand (from Complete Anatomy App [21]), (b) Resistor network model (Model 1), (c) Finite element model (Model 2).

B. Assessment Validation and Scoring

We previously conducted a pilot study of this assessment to determine whether it reliably measures differences between novice and expert reasoning [20]. We recruited 12 experts (faculty members who do research in heat transfer, as well as professional engineers who design heat exchangers) and 12 undergraduate students who were taking or had recently taken heat transfer to participate in think-aloud interviews while taking the assessment.

We used data from the expert responses to create a scoring rubric for the assessment. Students were graded based on how well their responses agreed with the experts from the pilot study. For Assumptions and Sensitivity questions, possible answers to each item in the question are TRUE and FALSE, which translate to values of 0 and 1, respectively. For Information question, possible answers to each option in the question are Definitely Want, Might Want, and Don't Need, with values of 1, 0.5, and 0, respectively. Choices with stronger expert consensus (either for or against) were given more weight to the total score than choices on which experts disagreed. The score for each question was calculated using the following formula:

$$closed\text{-}response\ score = \sum_{items} (2 * response - 1) * (\%consensus - 50\%). \quad (1)$$

For example, if 75% of experts selected TRUE for an assumption, then a student would receive +0.25 points for selecting

TRUE for that assumption, or lose 0.25 points for selecting FALSE for that assumption. Equation 1 automatically discounts items which do not have strong expert consensus. The total score for a given question (Assumptions, Information, or Sensitivity) is the sum of the scores on the individual items.

To score open-response questions (Model 1, Model 2, and Experiment and Simulation), we developed a code book (Table I) corresponding to specific statements made by experts. We used a multi-round emergent coding scheme to develop stable codes, and then identified codes mentioned by three or more of the 12 experts as "expert consensus codes" (CC). Codes mentioned by fewer experts were identified as "expert non-consensus codes" (NC). Details of the coding process of expert responses and the development of the rubric are discussed in our previous work [20]. Statements made by students which contradicted expert consensus were identified as "extraneous codes" (EC). In a few instances, students asked for material properties when answering the question "are there important features missing from the model" for Model 1. Although the comment was not wrong, it was not mentioned by any of the experts and was of lesser importance than the other missing features mentioned by the experts (e.g., missing axial dependence for the temperature). Thus, this response was not labeled as extraneous and was simply ignored in the scoring.

TABLE I
RUBRIC FOR SCORING THE ASSESSMENT BASED ON RESPONSES FROM ELEVEN EXPERTS. THE RUBRIC ITEMS WERE MENTIONED BY AT LEAST THREE EXPERTS.

Topic Area	Expert Consensus Codes
Model 1	• Missing axial dependence • Problems with lumped flows or missing capillaries • Missing internal convective resistance • Unclear boundary conditions (T_a, T_v not specified)
Model 2	• Remove transient term in equation • Neglect circumferential conduction (in the theta direction) • Neglect axial conduction • Difficulty in estimating external convection coefficient • Problems with lumped flows or missing capillaries
Experiment and Simulation	Questions about experiment and data collection: • Asking about error bars • Asking about measurement probe • Asking about experimental control: environment • Asking about subject differences
	Questions about model prediction and validation: • Large temperature mismatch between model and experiment • Asking about error bars in simulation result • Asking about model parameter: perfusion rate

Students were scored according to how closely their codes matched experts, and penalized for incorrect or extraneous comments. The score for each group of open-response questions was calculated as follows:

$$open\text{-}response\ score = N_{CC} - 0.5N_{EC}. \quad (2)$$

For two out of the three closed-response questions, the average student scores were about 70% of the average scores of the experts. However, for the remaining closed-response question and two out of the three groups of open-response questions, the average student scores were less than 40% of the average scores of the experts. Building off of the previous study, in the current study, we sought to understand if students are improving on any of our measures of problem-solving during typical heat transfer courses.

C. Data Collection

We delivered the assessment as a pre- and post-test in two third-year heat transfer courses at a highly selective private research university in the northeastern USA. Course 1 is in the Chemical Engineering (ChE) department and Course 2 is in the Biological and Environmental Engineering (BEE) department.The assessment was assigned to students at the beginning and end of the term as part of homework assignments. Students were given credit for completing the assessment, not graded on the accuracy of their responses. The ChE course is a three-credit course that consists of three 50-minute lectures and one 2-hour discussion section per week for 15 weeks. Students did practice problems in the discussion sections and practice problems were mostly drawn from the textbook. Students completed the pre-assessment in week 1-2 and the post-assessment in week 14. In total, 25 out of the 41 students enrolled completed both the pre- and post-assessment and gave consent for using their data for research. The BEE course is a four-credit course that consists of three 50-minute lectures and one 50-minute discussion section, which is often used as additional lecture time, per week for 15 weeks. The BEE course covered the same topics about heat transfer as the ChE course but in a more biological context, which is the problem context of our assessment. Students completed the heat transfer pre-assessment in week 3 and the post-assessment in week 14. 12 out of the 25 students enrolled completed both the pre- and post-assessment and gave consent for using their data for research. Both courses are a requirement for their respective majors. 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. To have a meaningful pre-post comparison, we only included students who completed both the pre-test and the post-test.

III. RESULTS

A. Closed-Response Questions

For the three closed-response questions, the average scores of the students in the pre-test and post-test were all lower than that of the experts from the pilot-study (see Table II). For both courses, the average student scores in the post-test were higher than that in the pre-test. To check if the changes in the student scores are statistically significant, we calculated the paired Cohen's d for the effect size using the following

equation: $d = \bar{z}/s_z$, where $z = x_{post} - x_{pre}$ is the difference between pre-score and post-score for each student, $\bar{z}$ is the average of the difference, and s_z is the standard deviation of the difference. Effect sizes of 0.5 or greater are considered to be large, and effect sizes of 0.2-0.5 are medium [22]. Additionally, we used a paired t-test to compute the p-value to check whether the average difference of the post-score and pre-score is significantly enough from zero. Cohen's d and p-value are both included in Table II.

The question with the largest positive effect size was the Assumptions question for the BEE course, with Cohen's d = 0.47 and p = 0.13. Typically, a p-value less than 0.05 is considered statistically significant. Although the p-value for changes in the scores for the Assumptions question for the BEE course is larger than 0.05, given the small sample size ($n = 12$), the effect size of 0.47 is still worth mentioning. The question with the second largest positive effect size was the Information question for the ChE course, with Cohen's d = 0.42 and p = 0.05. Interestingly, the Sensitivity question for the BEE course had a large negative effect size, with Cohen's d = -0.84 and p = 0.01. For all of the other questions, the effect sizes were not large enough to be considered educationally significant.

TABLE II
DESCRIPTIVE STATISTICS FOR PRE-SCORES AND POST-SCORES FOR CLOSED-RESPONSE QUESTIONS

		Expert Pilot	ChE Pre	ChE Post	BEE Pre	BEE Post
Assumptions	**Average**	0.99	0.34	0.51	0.76	0.94
	Std	0.60	0.71	0.67	0.37	0.44
	Scaled Avg	1.00	0.35	0.52	0.77	0.96
	Cohen's d		0.16		0.47	
	p-value		0.44		0.13	
Information	**Average**	5.07	4.27	4.97	3.97	4.14
	Std	0.88	1.41	1.29	1.52	2.04
	Scaled Avg	1.00	0.84	0.98	0.78	0.82
	Cohen's d		0.42		0.11	
	p-value		0.05		0.72	
Sensitivity	**Average**	1.25	0.33	0.30	0.78	0.01
	Std	0.38	0.51	0.75	0.69	0.67
	Scaled Avg	1.00	0.27	0.24	0.62	0.01
	Cohen's d		-0.05		-0.84	
	p-value		0.82		0.01	

To compare the performance of students across these three questions that were on different scales, we rescaled the student pre-scores and post-scores using the expert average:

$$scaled\ score = \frac{score}{\text{mean}(expert\ scores)}. \quad (3)$$

Out of all the post-test scores, the scaled average for the Information question for the ChE course and the scaled average for the Assumptions question for the BEE course were the largest (0.98 and 0.96, respectively). These are also the questions that had the largest effect size between pre and post. This suggests for the ChE course, students gained more expertise in identifying what information is needed as a result of taking the course. Similarly, the result suggests for the BEE course, students became close to the average expert level for making assumptions as a result of taking the course. However, for both courses, the scaled post-average for sensitivity was very low: 0.24 for ChE and 0.01 for BEE, with nearly no change between pre and post for ChE and a large negative change between pre and post for BEE.

To illustrate the changes in the student performance between pre-test and post-test visually, we subtract the scaled pre-score from the scaled post-score and plot the changes in scores in Fig. 2. The box encompasses the 25th to the 75th percentiles of each score and the thick center line in the box plot indicates the median. Qualitatively, the relative location of the boxes convey similar information as Cohen's d, with the Sensitivity question for the BEE course having the largest negative change and the Information question for the ChE course and the Assumptions question for the BEE course having the largest positive change between pre-test and post-test.

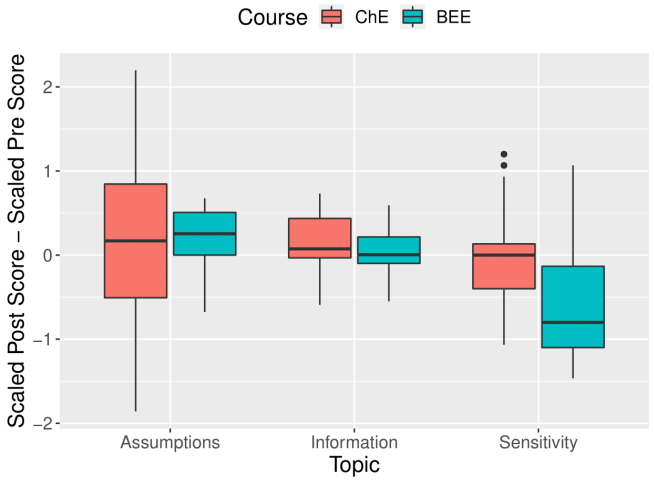

Fig. 2. Changes in scaled post-score compared to scaled pre-score for closed-response questions in both courses. The pre-scores and post-scores are scaled by dividing by the average expert score. For each student, the scaled pre-score is subtracted from the scaled post-score.

Each of the closed-response questions consists of a number of individual items and the score for that question is a sum of scores for each item, according to Eqn. 1. To understand the reason behind the questions with the largest positive and negative effect sizes, we calculated the change in the average student score for each item between pre and post and calculated the percent of students choosing the item in the pre-test and post-test. The Assumptions question for the BEE course had the largest effect size: 0.47. The Assumptions question asked the participant to choose appropriate assumptions from a list of 10 assumptions to simplify the problem. The assumption item with the largest positive change for the average score of students in the course was "neglect radiative heat transfer", with 50% of students choosing this assumption in pre-test and

67% of students choosing this assumption in post-test. The assumption item "assume thermal conductivity of blood and flesh are the same" had the second largest positive change for the average score of students in the course, with 67% of students choosing this assumption in pre-test and 25% of students choosing this assumption in post-test. Since the scores were calculated according to Eqn. 1 and the majority of experts didn't choose this assumption, a decrease in the percent of students choosing this assumption item resulted in an increase in the score. In our previous study, we found that experts are unlikely to make assumptions that may be quantitatively valid, but do not simplify the model in a useful way [20].

The Information question for the ChE course had an effect size of 0.42. The question asked participants to choose "definitely want", "might want" or "don't need" for 32 different pieces of information. The total score was calculated according to Eqn. 1. We will now compare the percentage of students who selected "definitely want" in pre-test and post-test for items that have the largest change in scores between pre and post. "Radius of finger" had the largest change in score between pre and post (32% students at pre-test and 88% at post-test), followed by "length of finger" (52% at pre-test and 72% at post-test), followed by "outside air temperature" (76% at pre-test and 92% students at post-test). None of the items had large negative changes in scores between pre and post on average.

The Sensitivity question for BEE had a negative effect size of -0.84. The question asked participants to choose five out of nine given variables that they thought the findings will be most sensitive to. Interestingly, none of the nine sensitivity items had a positive change in average score between pre and post. Three items had zero change on average and six items had negative change. "Viscosity of blood" was the sensitivity item with the largest negative change in the average score between pre and post, with 0% students choosing this item in pre and 33% students choosing this item in post. Since only a small percentage of experts chose this item (% consensus=16.7%), according to Eqn. 1, students were penalized when they choose this item. Thus, a large increase in the percent of students choosing this item resulted in a large decrease in the score from pre to post. The sensitivity item "depth of artery and vein within the finger" had the second largest negative change in the score, with 58% students choosing this item in pre and 25% students choosing this item in post. Since the majority of experts chose this item, a decrease in the % of students choosing this item resulted in a decrease in the average score. The sensitivity item "thermal diffusivity of blood" also had the second largest negative change in the score, with 42% students selecting this item in pre and 58% in post. Since none of the experts chose this item, the increase in percentage of students choosing the item resulted in a decrease in the average score.

B. Open-Response Questions

For the open-response questions, we calculated the scores for three groups of questions: Model 1, Model 2, and Experiment & Simulation. The statistics of the student pre-scores and post-scores are included in Table III. The scores of experts in the pilot-study are also included for comparison. For all three groups of questions, the student average pre-scores and post-scores were all lower than the expert average. By looking at the scaled score (student score divided by the average of expert score), we saw more clearly that both the scaled pre-scores and scaled post-scores for Model 1 were close to zero. The scaled scores for Model 2 were slightly higher, but with the highest average still less than 30% of the expert average. For the Experimental Questions, however, both the scaled pre-scores and post-scores were close to 50% of the expert average. We again calculated Cohen's d and p-value to find questions with large effect size between pre and post and determine whether the effect size is statistically significant. The groups of open-response questions with the largest effect size were Model 2 questions for the ChE course (Cohen's d = 0.30, p = 0.15) and Model 1 questions for the BEE course (Cohen's d = 0.27, p = 0.37). The analysis of the scaled score and effect size in combination suggest that before taking the course, students already had some but not enough decision-making abilities for interpreting experimental data and comparing experimental data with simulation results and taking the course only had a very small positive effect. Before taking the course, students had difficulty giving feedback to the two proposed solutions and taking the course only had a medium positive effect on the performance on Model 2 for students in the ChE course and Model 1 for students in the BEE course.

TABLE III
DESCRIPTIVE STATISTICS FOR PRE-SCORES AND POST-SCORES FOR OPEN-RESPONSE QUESTIONS

		Expert Pilot	ChE Pre	ChE Post	BEE Pre	BEE Post
Model 1 Questions	**Average**	1.67	-0.12	-0.04	0.00	0.17
	Std	0.87	0.75	0.54	0.52	0.54
	Scaled Avg	1.00	-0.07	-0.02	0.00	0.10
	Cohen's d		0.10		0.27	
	p-value		0.63		0.37	
Model 2 Questions	**Average**	1.73	0.06	0.22	0.33	0.50
	Std	1.56	0.46	0.56	0.75	1.22
	Scaled Avg	1.00	0.03	0.13	0.19	0.29
	Cohen's d		0.30		0.14	
	p-value		0.15		0.65	
Experiment and Simulation	**Average**	3.00	1.68	1.72	1.25	1.50
	Std	1.22	1.03	0.98	1.06	1.38
	Scaled Avg	1.00	0.56	0.57	0.42	0.50
	Cohen's d		0.03		0.19	
	p-value		0.90		0.52	

We plot the difference between the scaled post-scores and scaled pre-scores for the open-response questions in Fig. 3. The scores are scaled according to Eqn. 3. Due to the large standard deviation for the open-response scores, a box plot is not a very meaningful representation. Instead, we use a violin plot to represent the difference in the scaled scores for the two courses in Fig. 3. Although some students had a change in

scaled score of more than 0.5 points, the majority of students had little to no change.

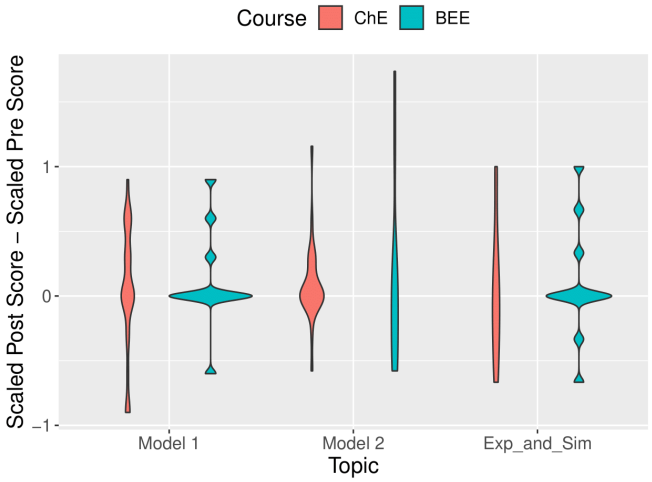

Fig. 3. Changes in scaled post-score compared to scaled pre-score for open-response questions in both courses. The pre-scores and post-scores are scaled by dividing by the average expert score. For each student, the scaled pre-score is subtracted from the scaled post-score.

To understand the reasons behind the questions with the largest effect sizes (Model 2 for ChE and Model 1 for BEE), we compared the difference in the number of expert consensus codes and extraneous codes mentioned by students in the pre-test and post-test. For Model 2 for ChE (Cohen's d=0.30), in the pre-test, out of the 25 students, only four students mentioned an expert consensus item: "problems with lumped flows or missing capillaries". In the post-test, a total of eight expert consensus codes were mentioned among seven students. Interestingly, only one student mentioned "problems with lumped flows or missing capillaries", all of the other expert consensus codes that were mentioned had to do with making simplifications: "remove transient term", "neglect circumferential conduction", and "neglect axial conduction". In both the pre-test and post-test, a total of five extraneous items were mentioned.

Model 1 for the BEE course had an effect size of 0.27. In the pre-test, out of the 12 students, only one student mentioned the expert consensus item "missing axial dependence", one student mentioned the expert consensus item "problems with lumped flows or missing capillaries", and four extraneous items were mentioned. In the post-test, three students mentioned three expert consensus items "missing axial dependence", "problems with lumped flows or missing capillaries", and "missing internal convective resistance", and two extraneous items were mentioned. The post-test had a 50% increase in the number of expert consensus items mentioned and a 60% decrease in the number of extraneous items mentioned. Because students got one point added for each expert consensus item and 0.5 points deducted for each extraneous item, the changes resulted in a moderate effect size (0.27). However, in terms of the individual improvements for students, with only one additional student mentioning an expert consensus item in the post-test compared to the pre-test, the change was not very significant.

IV. Discussion

A. Changes Reflecting Practice Making Decisions

When designing and conducting the current study, we sought to understand if students are improving on any of our measures of problem-solving during typical heat transfer courses. Broadly speaking, we see that students' problem-solving in certain areas improves over the heat transfer course but in other areas the improvement is very minimal. The improvements seem to be in areas where they are given practice in making certain expert decisions during the heat transfer course and the problem-solving gains are smaller when this practice is more limited.

For decisions that students had practice making in previous courses, students performed well on both the pre-test and post-test. For example, for the Sensitivity question in the BEE course, 83% of students selected "choice of boundary conditions" in both the pre-test and post-test. "Outside air temperature" was also selected by 83% of students in both the pre-test and post-test. Outside air temperature can also be considered a boundary condition. Boundary conditions are frequently stressed in required courses students took before heat transfer, such as differential equations and fluid mechanics. This continued emphasis in previous courses and in heat transfer results in the large percentage of students choosing choices related to boundary conditions in both the pre-test and post-test. For the open-response questions on experiment and simulation in both courses, the average student score is about 50% the average expert score in both the pre-test and the post-test. The questions about experiment and simulation probe the decisions "how believable is the information" and "how does information compare to predictions" from the 29 decisions in the work by Price *et al* [18]. Students have had practice making these decisions in previous courses, especially laboratory courses, which results in an acceptable level of mastery (50%). However, taking the heat transfer course doesn't improve making these decisions significantly. We hypothesize this is because students were not given additional opportunities to compare experimental data and model predictions in heat transfer courses.

For decisions that students had practice making in the heat transfer course, we observed larger effect size between pre and post. In particular, for Model 2 questions for the ChE course, the effect size is 0.30. As shown in Table I, three out of the five expert consensus codes for Model 2 are related to simplifications that the experts suggested: "remove transient term in equation", "neglect circumferential conduction", and "neglect axial conduction". These codes correspond to the decision "how to simplify the problem to make it easier to solve" from Price *et al*. [18] In the pre-test, none of the students mentioned these simplifications, whereas in the post-test seven out of 25 students mentioned at least one of these simplifications. We interviewed the graduate teaching assistant for the ChE course and found that students had practice

activities where they would cross out unnecessary terms in equations in the course.

Another important trend we noticed is that after taking the heat transfer course, students were able to recognize the importance of more variables but were still not good at choosing the most important variables from a list of variables. For example, blood viscosity is included as one of the 32 pieces of information in the Information question and is also included as one of the nine variables in the Sensitivity question. Blood viscosity is the item with the largest negative change in score between pre and post for the Sensitivity question in the BEE course, with 0% students selecting this variable in pre and 33% students selecting this variable in post. Since the majority of experts didn't choose this variable (only two experts out of twelve), students are penalized for selecting this variable according to Eqn. 1. In the Information question for the BEE course, 0% students chose "definitely want" for blood viscosity and 50% students chose "don't need" in the pre-test. In the post-test 58% students chose "definitely want" for blood viscosity and 8% chose "don't need". Since the expert consensus for selecting this item is 79%, students who chose "definitely want" were awarded and students who chose "don't need" were penalized, which results in a large increase in the score for this item between pre and post. Comparing the student responses for the same item in these two questions, we see that after taking the course, students become aware that blood viscosity is needed to model the convective heat transfer. However, when asked to select five most important variables from a list of nine variables, they fail to recognize that the viscosity of blood is not as important as some of the other variables, such as the flow rate of blood. We observe similar trends for the other variables in the Sensitivity question for the BEE course. All the items with an increase in the percentage of students choosing the item from pre to post are the ones with expert consensus less than or equal to 50%. All the items with a decrease in the percentage of students choosing the item from pre to post are the ones with expert consensus larger than 50%. This suggests that after taking the heat transfer course, students are more aware of more variables the findings could be sensitive to, but they don't have enough expertise to choose the most important variables from all the relevant variables. Overall, students are learning more content knowledge after taking the heat transfer course, but they need more practice distinguishing the most important variables or pieces of information.

B. Developing Expertise

In addition to studying if students improved on any of our measures of problem-solving during typical heat transfer courses, we are also interested in how students' problem-solving abilities at the end of the course compare to those of the experts. In Table II and Table III, we calculated scaled average scores by dividing the average student scores by the average expert scores. For some of the questions, the average student performance in the post-test is more than 80% of the expert average score: Assumptions question for the BEE course (scaled score = 0.96) and the Information question for the ChE course (scaled score = 0.98) and the BEE course (scaled score = 0.82). For some of the questions, the average student performance in the post-test is about 50% the expert average score: Assumptions question for the ChE course (scaled score = 0.52), Experiment and Simulation questions for the ChE course (scaled score = 0.57) and the BEE course (scaled score = 0.50). For all the other questions, the average student performance in the post-test is less than 30% the expert average score: Sensitivity question, Model 1 questions, and Model 2 questions. For the Assumptions question, the average student score in the post-test in the BEE course is much higher than the ChE course. We hypothesize this is because students in the BEE course had more practice making assumptions that are relevant to our assessment, which is also in the biological context. Both the ChE course and BEE course have high performance on the Information question in the post-test, which suggests students had practice deciding what information is needed while taking the heat transfer course.

It is reasonable to ask what is the desired mastery level we want the students to achieve at the end of the heat transfer course. Since the experts in the pilot-study are professors doing research in heat transfer or industry professionals who regularly use heat transfer in their work, we don't necessarily expect third year college students to achieve the same level of expertise as the experts at the end of the heat transfer course. Due to a lack of previous studies on assessing student and expert problem-solving in engineering, there isn't a clear consensus for what the desired level of understanding should be for third year college students. In the current study, the only questions with scaled student score larger than 60% of the average expert score and effect size larger than 0.2 are the Information question for the ChE course and the Assumptions question for the BEE course. The results in our study suggests that students need more opportunities to practice making a variety of problem-solving decisions in their course work.

C. Reliability of the Assessment

As we discussed in the Methods section, the pre-assessment and post-assessment were assigned as one of the homework questions on the first and last homework in the course and students received grades for completion only. To assess whether students put in enough effort in the assessment and if there is a change in the amount of effort between the pre-test and post-test, we calculated the median time spent on the pre-test and post-test for both courses: ChE pre 0.6 hr, ChE post 0.55 hr, BEE pre 1.4 hr, and BEE post 0.62 hr. Although for the BEE course, the time spent on the post is much shorter than that on the pre, the post time is similar to the time spent by students in the ChE course and is sufficient to finish the assessment with enough effort.

We also compared the average length of student responses (by word count) per group of questions (Model 1, Model 2, Experiment & Simulation) for the pre-test and post-test. For the pre-test and post-test in both courses, the average length of student responses in the questions for Model 2

(finite element model) is shorter than that for Model 1, even though the two models have similar questions. The length of responses per question for the Experiment & Simulation questions, which are given after the Model 2 questions are longer than responses for Model 2. This suggests students' efforts didn't diminish as they progressed in the assessment. Rather, the shorter answers in Model 2 are either a result of questions that look repetitive to those in Model 1 or students have less feedback to give for Model 2 because it's a more accurate model. We found that, overall, students provided less detail in this format than in the pilot-study [20]. In order to ask students to more thoroughly explain their answers, we need to revise the questions. For example, instead of asking "are there important features missing from this model", we will ask "what important features (if any) are missing from this model? Please list at most three missing features." In the current pre-post implementation, there seems to be slight decrease in effort between the pre-test and post-test as evident by the time spent for the BEE course and between a 10% to 35% decrease in the average length of responses for all groups of questions. To motivate students to put in more effort in the post-test, we can frame the post-test as an opportunity to review for the final exam.

D. Limitations and Future Work

Although the current heat transfer assessment was able to distinguish between problem-solving skills of experts and students and identify areas where students improved after taking the heat transfer course, several aspects of the assessment need to be improved. First, some questions generated similar responses from the participants and are thus redundant. Second, students interpreted some questions as yes/no questions instead of explaining their answers. Third, the resistor network approach (Model 1) is not consistently taught in heat transfer courses in different engineering majors. Heat transfer courses in mechanical engineering often put more emphasis on resistor network approach than heat transfer courses in chemical engineering and biological engineering do. In the new version of the assessment, we are making improvements in all the aforementioned areas and will use a 2D version of the finite element model to replace the resistor network model. Additionally, to have a more accurate measure of the amount of time students spend on the assessment, we plan to include timer questions in the Qualtrics survey to record the time spent on each question.

Another limitation of the current study is the small sample size: 25 students in the ChE course and 12 students in the BEE course. Additionally, the pre-post study did not include a heat transfer course in mechanical engineering. After doing pilot-testing with experts and students using the revised assessment, we plan to implement the assessment as a pre-test and post-test in more heat transfer courses in biological engineering, chemical engineering, and mechanical engineering. In future pre-post study, we will also recruit professors from more schools to have more variety in the student population.

We also plan to develop a closed-response version of all of the questions in the assessment to automate the scoring of the assessment. This will greatly reduce the time needed to analyze the student responses and enable a wider distribution of the assessment. Responses from initial pilot testing, think-aloud interviews with students, and expert responses will be used to replace the open-response questions with "choose-many" multiple-choice questions. This will ensure that the questions include reasonable distractors generated by students as well as answers chosen by a consensus of experts. We will compare performance of students in similar courses across and within institutions on the closed-response and open-response versions of the assessment to evaluate any information loss when moving to closed-response questions.

To give students more practice with making problem-solving decisions, we hypothesize it will be necessary to redesign heat transfer courses to focus on the problem-solving skills targeted by the assessment. This can be done using the principle of "deliberate practice", which is effortful practice of specific skills with timely and precise feedback from an instructor [23], as well as best practices for active learning as outlined by Jones, Madison, and Wieman [24]. The basic format of the in-class activities can follow a preparation for future learning model [25], where students are first presented with a scaffolded real-world problem and asked to make some of the problem-solving decisions before being told what the expert consensus is. The heat transfer assessment can then be used in a pre-post design in both traditional and newly designed heat transfer courses to assess the impact of deliberate practice as a research-based teaching method.

V. Conclusions

We administered an assessment of problem-solving as a pre- and post-test in two heat transfer courses, one in chemical engineering, and one in biological and environmental engineering. We find that in some areas, students' problem-solving skills improved after taking the heat transfer course. Students saw particularly large improvements making assumptions in the BEE course, asking for information needed in the ChE course, and making simplifications in the ChE course. This can be plausibly explained by the practice students received making these decisions in the course. Despite this progress, there are also areas where students showed little improvement. After taking the course, students become aware of the importance of more variables but are not quite able to distinguish the most important variables from a given list of variables. The results in our study suggest that students need more opportunities to practice making a variety of problem-solving decisions in their course work.

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. We hope to encourage other educators to use this assessment in their courses to measure how well they are preparing their students to solve real-word engineering problems.

REFERENCES

[1] ABET, "Criteria for accrediting engineering programs, 2019 – 2020," 2019.

[2] C. L. Dym, "Design, systems, and engineering education," *International Journal of Engineering Education*, vol. 20, no. 3, pp. 305–312, 2004.

[3] C. L. Dym, A. M. Agogino, O. Eris, D. D. Frey, and L. J. Leifer, "Engineering design thinking, teaching, and learning," *Journal of Engineering Education*, vol. 94, no. 1, pp. 103–120, 2005.

[4] D. Jonassen, J. Strobel, and C. B. Lee, "Everyday problem solving in engineering: Lessons for engineering educators," *Journal of Engineering Education*, vol. 95, no. 2, pp. 139–151, 2006.

[5] D. H. Jonassen, "Designing for decision making," *Educational Technology Research and Development*, vol. 60, no. 2, pp. 341–359, 2012.

[6] N. Shin, D. H. Jonassen, and S. McGee, "Predictors of well-structured and ill-structured problem solving in an astronomy simulation," *Journal of Research in Science Teaching*, vol. 40, no. 1, pp. 6–33, 2003.

[7] 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.

[8] Q. Symonds, "The global skills gap in the 21st century," 2018. [Online]. Available: https://www.qs.com/portfolio-items/the-global-skills-gap-in-the-21st-century/

[9] C. Grant and B. Dickson, "Personal skills in chemical engineering graduates: the development of skills within degree programmes to meet the needs of employers," *Education for Chemical Engineers*, vol. 1, no. 1, pp. 23–29, 2006.

[10] E. P. Douglas, D. J. Therriault, J. A. Magruder Waisome, E. Buten, and E. A. L. Bates, "Characterization of problem types in engineering textbooks," in *2022 ASEE Annual Conference Proceedings*, 2022.

[11] D. H. Jonassen, A. Johri, and B. Olds, "Engineers as problem solvers," in *Cambridge handbook of engineering education research*. Cambridge University Press New York, NY, 2014, pp. 103–118.

[12] C. J. Atman, R. S. Adams, M. E. Cardella, J. Turns, S. Mosborg, and J. Saleem, "Engineering design processes: A comparison of students and expert practitioners," *Journal of engineering education*, vol. 96, no. 4, pp. 359–379, 2007.

[13] D. R. Woods, "An evidence-based strategy for problem solving," *Journal of Engineering Education*, vol. 89, no. 4, pp. 443–459, 2000.

[14] J. Larkin, J. McDermott, D. P. Simon, and H. A. Simon, "Expert and novice performance in solving physics problems," *Science*, vol. 208, no. 4450, pp. 1335–1342, 1980.

[15] M. T. Chi, P. J. Feltovich, and R. Glaser, "Categorization and representation of physics problems by experts and novices," *Cognitive science*, vol. 5, no. 2, pp. 121–152, 1981.

[16] J. I. Heller and F. Reif, "Prescribing effective human problem-solving processes: Problem description in physics," *Cognition and instruction*, vol. 1, no. 2, pp. 177–216, 1984.

[17] W. K. Adams and C. E. Wieman, "Analyzing the many skills involved in solving complex physics problems," *American Journal of Physics*, vol. 83, no. 5, pp. 459–467, 2015.

[18] A. M. Price, C. J. Kim, E. W. Burkholder, A. V. Fritz, and C. E. Wieman, "A detailed characterization of the expert problem-solving process in science and engineering: Guidance for teaching and assessment," *CBE—Life Sciences Education*, vol. 20, no. 3, p. ar43, 2021.

[19] A. M. Price, E. W. Burkholder, S. Salehi, C. J. Kim, V. Isava, M. P. Flynn, and C. E. Wieman, "An accurate and practical method for measuring science and engineering problem-solving expertise," *Submitted*.

[20] J. Zhang, S. Fatehiboroujeni, M. J. Ford, and E. W. Burkholder, "Assessing authentic problem-solving in heat transfer," in *2022 ASEE Annual Conference Proceedings*, 2022.

[21] Complete Anatomy App. [Online]. Available: https://3d4medical.com/press-category/complete-anatomy

[22] J. R. Fraenkel, N. E. Wallen, and H. H. Hyun, "How to design and evaluate research in education," 2012.

[23] K. A. Ericsson, R. R. Hoffman, A. Kozbelt, and A. M. Williams, *The Cambridge handbook of expertise and expert performance*. Cambridge University Press, 2018.

[24] D. J. Jones, K. W. Madison, and C. E. Wieman, "Transforming a fourth year modern optics course using a deliberate practice framework," *Physical Review Special Topics-Physics Education Research*, vol. 11, no. 2, p. 020108, 2015.

[25] D. L. Schwartz and T. Martin, "Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction," *Cognition and instruction*, vol. 22, no. 2, pp. 129–184, 2004.