Teaching and Assessing Decision-Making Across the Physics Curriculum

by

Michael E. Robbins

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
May 10, 2025

Keywords: Problem Solving, Decision-Making, Graduate Physics, Quantum Mechanics, Introductory Physics Labs

Copyright 2025 by Michael E. Robbins

Approved by

Eric W. Burkholder, Chair, Assistant Professor, Department of Physics
Jordan Harshman, C. Harry Knowles Associate Professor, Department of Chemistry and Biochemistry
Guillaume M. Laurent, Professor, Department of Physics
Marcelo A. Kuroda, Thomas and Jean Walter Associate Professor, Department of Physics
Karen S. McNeal, Molette Endowed Professor, Department of Geosciences

Abstract

Problem solving is prevalent in all levels of physics and receives considerable attention in physics education research. This dissertation focuses on investigations of problem-solving in understudied areas of the physics curriculum: courses for non-STEM majors and graduate-level courses. The first study describes the design, implementation, and evaluation of laboratory activities for non-science majors which scaffolded experimental design and decision-making over a semester. Students exhibited shifts toward more expert-like views of experimental physics and improved capabilities make experimental decisions. The activities designed in this study drew on existing frameworks in the PER community, but no such curricular frameworks exist at the graduate level. Thus, the next study used semi-structured interviews and textual analysis to identify physics instructors' expectations for problem-solving in graduate coursework, as well as how problem-solving is implemented and assessed in practice. Instructors expected far more problem-solving skills than were practiced in coursework. The final study discusses the development of an assessment to measure a broader range of problem-solving skills in quantum mechanics and describes evidence for the face validity of the assessment. Together, these chapters expand our understanding of problem-solving in classrooms and build the groundwork for future interventions.

Acknowledgments

As with most modern science problems, this collection of studies was not completed alone. There were more challenges than just the science on the road to completing this dissertation. Thank you too all my colleagues, friends, and family who supported me in you own ways: guidance, food, challenging problems, joy, competition, encouragement, adventures, comfort, and countless others. Your support means more to me than you will know, and I'm deeply grateful for you.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Problem solving is an essential skill at all levels of the physics curriculum and in professional scientific practice. Consequently, a considerable amount of literature in physics education research (PER) is focused on problem-solving. Much of the early research in problem solving began by describing how students solved problems through a cognitive approach. These investigations pertained to students' and experts' mental representations, the strategies and heuristics used, and investigations on cognitive load during problem-solving (for a comprehensive review, see [1]). Later investigations built on these characterizations by developing and evaluating interventions to enhance students' problem-solving skills. For example, developing problem-solving by providing steps, incorporating groupwork, and introducing computers to serve as tutoring systems (for a review see [2]). Despite the decades of research there are two limitations in this research: (1) problem solving remains not well-defined and most of these studies are focused on introductory physics courses for STEM majors.

The majority of PER that examines problem solving focuses on textbook-style problems. This approach treats problem solving in the literal sense, often providing procedures or strategies to solve these well-defined problems. However, this style of problem bears little resemblance to the problems faced by professional physicists, which often lack clearly defined initial conditions, goals, and intermediate steps. Furthermore, having students practice solving problems that are not well-defined has been shown to improve problem-solving performance (e.g., [3, 4]). Students must make important decisions determining how to manage the ill-defined aspects, which better resembles physics in practice. Indeed, Wieman [5] argues that ill-structured

problem-solving can be framed in terms of making a series of decisions with limited information.

Prior investigations of ill-structured problems were primarily done in introductory physics courses, which is common in PER [6]. The following chapters aim to expand on the understanding of decision-making in curriculums. In Chapter 2, an intervention to increase decision-making in introductory labs for non-science majors is discussed. In Chapter 3, the decision-making skills practiced in graduate physics coursework are investigated. These skills are then compared to instructor expectations. Finally, the development of a graduate quantum mechanics assessment is described and evidence for face validity is discussed in Chapter 4.

Chapter 2

Traditional introductory physics labs frequently focus on demonstrating and reinforcing lecture concepts despite evidence showing that they are ineffective at these goals. We transformed our semester of traditional physics labs for non-science majors to focus on experimental decision-making. Students practiced making experimental decisions with guidance that was faded over the semester. This transformation used the same lab equipment and maintained the number of individual labs and topics. We found students improved their ability to make experimental decisions and had a shift towards more expert-like beliefs about experimental physics.

Chapter 3

There is currently little physics education literature examining thinking and learning in graduate education, and even less literature characterizing problem-solving among physics graduate students despite this being an essential professional skill for physicists. Given reports of discrepancies between physics problem-solving in the undergraduate classroom and "real-world" problem-solving, we sought to investigate whether this discrepancy exists at the graduate level. We first investigate the problem-solving skills present in first-year graduate physics assignments. A recent framework that characterizes problem-solving as decisions-to-be-made was used. Assignments were taken from the four core courses of one academic year at one research-intensive university and coded by two researchers. We found that only four of the twenty-nine

decisions in the framework were present in most of the assignments. We then interviewed eleven instructors from three universities and asked which decisions they expected of first-year graduate students. Eleven decisions were expected by eight or more of the participants, but only four of these decisions were commonly practiced on assignments. Therefore, there seems to be a mismatch between instructor expectations and practice of problem-solving on assignments. This suggests that graduate physics courses may not be aligned with the problem-solving skills that physics graduate students will need in their research or future careers.

Chapter 4

One of the greatest weaknesses of physics education research is the paucity of research on graduate education. While there are a growing number of investigations of graduate student degree progress and admissions, there are very few investigations of learning at the graduate level. Additionally, existing studies of learning in physics graduate programs frequently focus on content knowledge rather than professional skills such as problem-solving. Given that over 90% of physics PhD graduates report solving technical problems regularly in the workplace, we sought to develop an assessment to measure how well graduate programs are training students to solve problems. Using a framework that characterizes expert-like problem-solving skills as a set of decisions-to-be-made, we developed and validated such an assessment in graduate quantum mechanics (QM) following recently developed design frameworks for measuring problem-solving and best practices for assessment validation. We collected validity evidence through think-aloud interviews with practicing physicists and physics graduate students, as well as written solutions provided by physics graduate and undergraduate students. The assessment shows strong potential in differentiating novice and expert problem-solving in QM and showed reliability in repeated testing with similar populations. These results show the promise of measuring expert decision-making in graduate QM and provide baseline measurements for future educational interventions to more effectively teach these skills.

References

[1] Leonardo Hsu et al. “Resource Letter RPS-1: Research in problem solving”. In: *American Journal of Physics* 72 (9 2004). ISSN: 0002-9505. DOI: 10.1119/1.1763175.

[2] Jennifer L. Docktor and José P. Mestre. “Synthesis of discipline-based education research in physics”. In: *Physical Review Special Topics - Physics Education Research* 10 (2 Sept. 2014). ISSN: 15549178. DOI: 10.1103/PhysRevSTPER.10.020119.

[3] Patricia Heller, Ronald Keith, and Scott Anderson. “Teaching problem solving through cooperative grouping. Part 1: Group versus individual problem solving”. In: *American Journal of Physics* 60 (7 1992). ISSN: 0002-9505. DOI: 10.1119/1.17117.

[4] Vazgen Shekoyan and Eugenia Etkina. “Introducing Ill-structured problems in introductory physics recitations”. In: *AIP Conference Proceedings*. Vol. 951. 2007. DOI: 10.1063/1.2820930.

[5] Carl Edwin Wieman. “Expertise in university teaching & the implications for teaching effectiveness, evaluation & training”. In: *Daedalus* 148 (4 2019). ISSN: 15486192. DOI: 10.1162/DAED_a_01760.

[6] Stephen Kanim and Ximena C. Cid. “Demographics of physics education research”. In: *Physical Review Physics Education Research* 16 (2 2020). ISSN: 24699896. DOI: 10.1103/PhysRevPhysEducRes.16.020106.

Chapter 2

Development and Evaluation of Decisions-Based Laboratories for Non-science Majors

Michael E. Robbins*, Eric W. Burkholder

Department of Physics, Auburn University, Auburn AL 36830 USA

*mer0031@auburn.edu

Accepted at The Physics Teacher

2.1 Introduction

Scripted content-focused labs do not seem to produce conceptual learning benefits [1, 2]. The traditional format is procedural and does not allow students to make important experimental decisions like how many trials to conduct, which measurements to collect, and how to set up the experimental apparatus [3]. We implemented a laboratory transformation in our course for non-science majors which produced inquiry-based, skills-oriented labs. We focused on a subset of lab skills from the investigative science learning environment (ISLE) labs [4]. These skills overlap with AAPT's six major themes for undergraduate physics labs [5]. The new format was inspired by structured qualitative inquiry labs (SQILabs), which focus on iteratively improved design [6]. Instead of providing step-by-step procedures, students were prompted to make experimental decisions. Since students are likely unfamiliar making these decisions, the labs slowly introduce new experimental decisions to make and provide scaffolding to do so. This scaffolding was faded away throughout the semester.

In introductory physics courses, student attitudes and beliefs shift away from expert-like beliefs over the course of a semester [7–10]. A study found an expert-like shift in student attitudes in an experimental-skills focused lab [3]. Some courses were able to negate this shift or attain a shift towards expert-like beliefs by specifically targeting student attitudes and beliefs [9, 10]. We found a shift towards expert-like beliefs in student attitudes and an increase in student interest over the semester without explicitly targeting these attitudes.

Previous literature suggests skill-focused labs and inquiry in labs can lead to students addressing problems in a more expert-like manner [11] and an increase in experimental design improvements and follow-up on [12]. Transforming a traditional lab into an inquiry-based, skills lab has been found in other forms [13]. The transformation we describe here differs from the aforementioned transformation in two ways. First, the labs we present here are each completed in a single two-hour period, allowing instructors to align their lab and lecture content more easily and to fit this transformation into a more traditional introductory lab schedule (i.e., not needed multiple class periods for a single laboratory activity). Second, and more important,

the labs we present here provide students the opportunity to fail and learn from their "productive failure" [14] rather than being explicitly told whether or not a design will work by an instructor.

2.2 Implementation

We transformed a one-semester laboratory course taken in conjunction with an introductory physics course for non-science majors at a large research university. The students were predominantly business and aviation majors. Each lab section had 24 students who worked in groups of two to three, led by one graduate teaching assistant (GTA). Each of the nine labs was individually transformed to address skills-focused learning objectives while operating within the same constraints on time, topic, and available equipment.

The transformation shifted the focus of labs from physics concepts to teaching experimental skills from four AAPT lab themes: designing experiments, developing technical and practical laboratory skills, analyzing and visualizing data, and communicating physics. Experimental design consists of designing an apparatus to address a particular goal while considering available equipment, the quality of data that can be collected, and improvements made upon reflecting on collected data. Developing technical and practical laboratory skills consists of minimizing experimental error and understanding the limitations of the experiment. Error can be minimized by the selection of independent and dependent variables measured, the range and granularity of the independent variable, and the number of trials. Analyzing and visualizing data consists of interpreting collected data to support a conclusion or address a lab goal. Analysis can include interpreting graphs, understanding the implications of a percent difference, and identifying and addressing unexpected results. Communicating physics consists of clearly and concisely sharing the findings and conclusions of the experiment and describing the experimental decisions made for the purpose of grading. An example of how traditional labs provided the experimental decision and the transformed labs asked students to make these decisions is shown in Table 2.2.

Table 2.1: A subset of example tasks from a Newton's 2nd Law lab. These tasks are shown for a traditional lab and how they could be transformed to an inquiry-based lab. The first three tasks fall under the measurement theme and the fourth task under the analysis theme.

Traditional	Inquiry-Based
Measure displacement and time.	Determine which values to measure to calculate experimental acceleration.
Use hanging mass values 20, 50, 75 and 100 grams.	Determine which hanging mass values to use.
Repeat each trial three times.	Determine how many trails to conduct.
Compare the experimental and theoretical accelerations.	Do your results support Newton's 2^{nd} Law?

In a simple case, these themes first appear during lab following the "approximate order of conducting experiments," shown in Figure 2.1: designing experiments, technical and practical lab skills, analyzing and visualizing data, and finally communicating physics. In practice, the sequence is not a straight progression. The iterative process requires revisiting previous experimental decisions, which would include not only targeting individual decisions such as communicating physics, but all the foundational decisions that build up to that skill. For example, while completing the lab in Table 2.2, students may (1) determine values to measure, (2) determine mass values to use, (3) determine the number of trials needed, (4) conduct a few trials, (5) notice an issue then change the values to measure in step 1, and (6) conduct more trials. Each lab has a set of learning goals which align with a lab-skills theme. Students spend most of their time making experimental decisions corresponding to the focused theme and sequentially following themes. Subsequent tasks are left open-ended because an experimental decision may drastically change all following decisions.

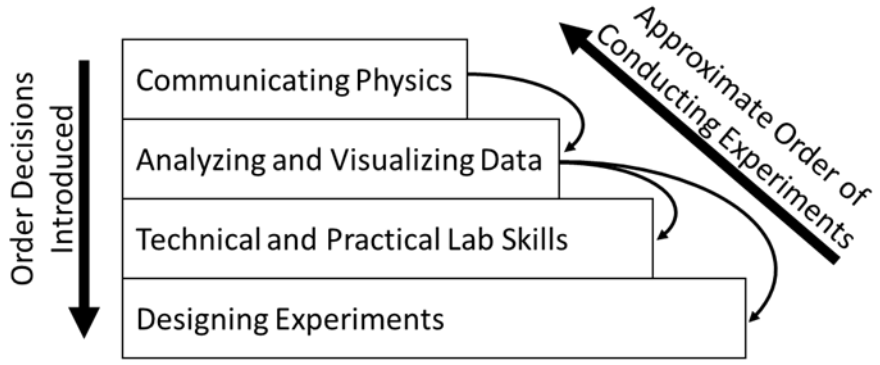


Figure 2.1: Experimental decision progression. Approximate order of conducting experiments is the simplest order these experimental decisions appear in practice. In practice this progression includes iteration, where previous decisions can be revisited as new information is added. The curved arrows show a non-exhaustive list of ways this iteration can manifest. Order decisions introduced is the order these skills were introduced and taught in our lab sequence.

The sequence of skill themes used throughout the semester is shown in Table 2.2. The first time a skill theme was introduced the lab included guiding questions, which encouraged students to consider important aspects and helped break down the open-ended questions to sub-questions. All subsequent times a skill theme appeared, there would be fewer or no guiding questions accompanying the open-ended question. For example, in week 5 students designed an experiment to determine a friction coefficient with guiding questions: (1) which values are needed for this calculation, (2) which of these values can be controlled and measured. Later they were prompted to evaluate their initial answer. Whereas, in week 10, students were asked to determine a spring constant with no guiding questions or prompt to evaluate the initial answer.

Table 2.2: Focused skill theme sequence. The week and the focused theme of the lab are detailed. Weeks 2 and 7 focus on a specific skills as opposed to the more general practice of that theme.

Lab Topic	Focused Theme	Non-focused Theme Required
1: 1-D motion	Communicating physics	-
2: 1-D motion	Skill: standard error	Communicating physics
3: Newton's 1^{st} Law	Analyzing and visualizing data	Communicating physics
4: Newton's 2^{nd} Law	Developing technical and practical lab skills	Communicating Physics, Analyzing and visualizing data
5: Friction Coefficient	Designing experiments	Communicating Physics, Analyzing and visualizing data, Developing technical and practical lab skills
6: Momentum	Developing technical and practical lab skills	Communicating Physics, Analyzing and visualizing data
7: Static Equilibrium	Skill: graph interpretation	Developing technical and practical lab skills
8: Parallel and Series Springs	Designing experiments	Communicating Physics, Analyzing and visualizing data, Developing technical and practical lab skills
9: Spring Constant	Analyzing and visualizing data	Communicating Physics, Analyzing and visualizing data
10: Circuits	Designing experiments	Communicating Physics, Analyzing and visualizing data, Developing technical and practical lab skills

2.2.1 Iterative and Deliberate Practice

The transformation aimed to help students learn these skills with deliberate practice [15]. Deliberate practice consists of providing students with a specific task or goal and including feedback loops during the practice. In this process, the effort is rewarded rather than the success [16]. The lab structure followed an iterative process in which students made an experimental

decision, tested the decision, reflected on the results, and improved upon their decision. Both the GTA and the student group served as sources of feedback. Students were prompted to reflect on their results and assess their decision's effectiveness and the GTA provided targeted feedback during the lab. This feedback structure provided students room to have "productive failure" and still successfully complete the lab. This "productive failure" could occur because students were prompted to make decisions to open-ended questions with minimal guidance. This led to groups of students that had complex discussions and struggled to answer the question, occasionally, not producing a successful solution, failure. Either the students, through the iterative process, or the GTA identified this failure. The GTA then provided feedback and additional scaffolding so the students could work towards a successful solution in the same lab period. The GTA provided more detailed feedback while grading the lab. The GTA feedback was partially in the form of a grade, which focuses on effort and understanding rather than the accuracy of the data. For example, detailing a good experimental design but having inaccurate data can score well if the anomaly or source of inaccuracy is thoroughly discussed by the students. The difficulty of the physics calculations was reduced, such that most students could independently solve the calculations so time could be focused on the skills-focused learning objectives rather than solving procedural textbook-like problems.

2.3 Measuring the Impact

Students from the first semester of the transformation completed the E-CLASS as a pre and post survey to measure student attitudes compared against expert-like attitudes [17]. The pre- and post-test were administered for bonus points during the first lab meeting and as homework on the last week of the semester, respectively. Of the 95 students, 27 student survey pairs were kept, due to incomplete surveys or students being flagged for not reading the questions. The percentage of expert-like responses to each student-reflection question was calculated and denoted as percent favorable (e.g., percent agreement with expert-like responses). The difference between the pre- and post-test percent favorable was calculated as the change across the semester and compared to the historical averages of E-CLASS for similar courses (introductory college physics courses).

Our class's average pre-test percent favorable was 64% (SD = 26.9%), compared to the historical average of 68%. Over the semester our class's percent favorable increased by 10% (d = 0.25), whereas the historical average is -1.0%. Questions with the largest percent favorable increases are shown in Table 2.3. The increase regarding understanding of concepts may arise because students make experimental decisions, requiring an understanding of the concepts instead of simply plugging into an equation. Students also showed an increased interest in physics. This increase is particularly striking, since nearly all students in this class are non-science majors. The GTAs reported student comments about enjoying the decision-making aspect of the labs and freedom to test their own ideas. The increase in attention to systematic error, intention of iterative improvement, and analysis methods align with the learning objectives of the transformation.

Table 2.3: E-CLASS question prompts and their respective percent favorable changes for our class and historical averages. Percent favorable is the percent of responses that agreed with expert-like answers.

		Percent Favorable Increase	
Question Number	Prompt	Transformation	Historical
3	When doing a physics experiment, I don't think much about sources of systematic error.	33%	0.2%
7	I don't enjoy doing physics experiments.	19%	-2%
21	I am usually able to complete an experiment without understanding the equations and physics ideas that describe the system I am investigating.	26%	3.8%
25	A common approach for fixing a problem with an experiment is to randomly change things until the problem goes away.	22%	-2.9%
27	When doing an experiment, I just follow the instructions without thinking about their purpose.	30%	-4.9%
29	If I don't have clear directions for analyzing data, I am not sure how to choose an appropriate analysis method.	19%	-2%

Additionally, the quality of student submissions seemed to improve throughout the semester despite the labs providing less guidance and structure as the semester progressed. Students seemed to address differences between predictions and their data better throughout the semester, agreeing with prior work [3].

2.4 Teaching implications

Students can improve their lab skills by making experimental decisions associated with these skills with timely feedback from the instructor. Most of a student's time should be spent on the learning objectives of the lab: making and reflecting on experimental decisions. These lab activities can be extended across multiple lab sessions or contained to a single session. A skills-focused lab adds steps and therefore time to the lab. A lab's length can be shortened by reducing the difficulty or frequency of calculations, minimizing the time spent learning new equipment or software, or reducing the steps in the lab. This reallocation of time shifts the focus to the learning objectives of the lab.

The transformation we described is highly customizable. It can be applied to individual labs or for a semester-long course. This transformation is compatible with any physics concepts and any lab skills, see AAPT's themes for recommendations [5]. To apply this transformation, instructors can select labs to transform, select lab skills to incorporate, match lab skills with labs, and let students make the associated experimental decisions. We recommend introducing only one skills theme per lab and introducing a new theme following the "order decisions introduced" shown in Figure 2.1. The sequence does not need to exactly mirror the sequence shown in Table 2.2.

The largest investment in the transformation is the lab development. Additionally, the instructor load during lab is larger with this transformation. The instructor is expected to be an expert in this course to make evaluating novel student solutions second nature. The instructor's role expands to assisting students make experimental decisions. We found each of the two graduate teaching assistants was able to fill this role successfully with only the GTA training offered for traditional labs. This transformation may make grading time intensive, depending

on the assessment format. The oral presentation used in our transformation reduced grading time outside of lab.

2.5 Acknowledgments

We thank Adam Pfeifle for his contributions to the development of the lab procedures.

References

[1] N. G. Holmes et al. “Value added or misattributed? A multi-institution study on the educational benefit of labs for reinforcing physics content”. In: *Physical Review Physics Education Research* 13 (1 2017), pp. 010129–0101240. ISSN: 24699896. DOI: 10.1103/PhysRevPhysEducRes.13.010129.

[2] Emily M. Smith et al. “Direct measurement of the impact of teaching experimentation in physics labs”. In: *Physical Review X* 10 (1 2020), pp. 011029–011039. ISSN: 21603308. DOI: 10.1103/PhysRevX.10.011029.

[3] Carl Wieman. “Comparative Cognitive Task Analyses of Experimental Science and Instructional Laboratory Courses”. In: *The Physics Teacher* 53 (6 2015), pp. 349–351. ISSN: 0031-921X. DOI: 10.1119/1.4928349.

[4] E Etkina and a Van Heuvelen. “Investigative Science Learning Environment – A Science Process Approach to Learning Physics”. In: *PER-based reforms in calculus-based physics* (2007).

[5] Joseph Kozminski et al. *AAPT Recommendations for the Undergraduate Physics Laboratory Curriculum Subcommittee Membership*. Tech. rep. 2014.

[6] N. G. Holmes, Carl E. Wieman, and D. A. Bonn. “Teaching critical thinking”. In: *Proceedings of the National Academy of Sciences of the United States of America* 112 (36 2015), pp. 11199–11204. ISSN: 10916490. DOI: 10.1073/pnas.1505329112.

[7] Steven J. Pollock. “Transferring transformations: Learning gains, student attitudes, and the impacts of multiple instructors in large lecture courses”. In: *AIP Conference Proceedings*. Vol. 818. 2006, pp. 141–144. DOI: 10.1063/1.2177043.

[8] Edward F. Redish, Jeffery M. Saul, and Richard N. Steinberg. “Student expectations in introductory physics”. In: *American Journal of Physics* 66 (3 1998), pp. 212–224. ISSN: 0002-9505. DOI: 10.1119/1.18847.

[9] Kara E. Gray et al. “Students know what physicists believe, but they don’t agree: A study using the CLASS survey”. In: *Physical Review Special Topics - Physics Education Research* 4 (2 2008), pp. 020106–020115. ISSN: 15549178. DOI: 10.1103/PhysRevSTPER.4.020106.

[10] K. K. Perkins et al. “2004 Physics Education Research Conference: Sacramento, California, 4-5 August 2004”. In: *2004 Physics Education Research Conference* 790 (2005), pp. 61–64. ISSN: 0009-4978. DOI: 10.5860/choice.43-4715.

[11] Eugenia Etkina et al. “Design and reflection help students develop scientific abilities: Learning in introductory physics laboratories”. In: *Journal of the Learning Sciences* 19 (1 2010), pp. 54–98. ISSN: 10508406. DOI: 10.1080/10508400903452876.

[12] N. G. Holmes, Dhaneesh Kumar, and D. A. Bonn. “Toolboxes and handing students a hammer: The effects of cueing and instruction on getting students to think critically”. In: *Physical Review Physics Education Research* 13 (1 2017), p. 010116. ISSN: 24699896. DOI: 10.1103/PhysRevPhysEducRes.13.010116.

[13] Steven Frederick Wolf and Mark W. Sprague. “Introductory Physics Labs: A Tale of Two Transformations”. In: *The Physics Teacher* 60 (5 2022), pp. 372–375. ISSN: 0031-921X. DOI: 10.1119/5.0032370.

[14] Manu Kapur. “Productive failure”. In: *Cognition and Instruction* 26 (3 2008), pp. 379–424. ISSN: 07370008. DOI: 10.1080/07370000802212669.

[15] K. Anders Ericsson, Ralf Th Krampe, and Clemens Tesch-Römer. “The Role of Deliberate Practice in the Acquisition of Expert Performance”. In: *Psychological Review* 100 (3 1993), pp. 363–406. ISSN: 0033295X. DOI: 10.1037/0033-295x.100.3.363.

[16] Daniel L. Schwartz, Jessica M. Tsang, and Kristen P. Blair. *The ABCs of How We Learn: 26 Scientifically Proven Approaches, How They Work, and When to Use Them*. W. W. Norton & Company, July 2016.

[17] Bethany R. Wilcox and H. J. Lewandowski. “Students’ epistemologies about experimental physics: Validating the Colorado Learning Attitudes about Science Survey for

experimental physics". In: *Physical Review Physics Education Research* 12 (1 2016), p. 010123. ISSN: 24699896. DOI: 10.1103/PhysRevPhysEducRes.12.010123.

Chapter 3

Decision-Making in Graduate Physics Coursework: What Is Being Assessed Versus What Is Expected

Michael E. Robbins*, Nathan D. Davis, Eric W. Burkholder

Department of Physics, Auburn University, Auburn AL 36830 USA

*mer0031@auburn.edu

Manuscript in press at Physical Review Physics Education Research.

3.1 Introduction

Despite decades of physics education research (PER) focused on improving students' conceptual understanding of important physics concepts [1–3] there remain few studies of thinking and learning among physics graduate students. In more recent years, there have been some important studies of graduate students' conceptual understanding of quantum mechanics [4], quantum measurement [5], and graphical representation of wave functions [6], which have shown that physics graduate students still struggle to learn some important fundamental ideas in physics. Similar to what we have seen in undergraduate education, some studies have shown that implementing student-centered pedagogical strategies like groupwork [6, 7] can improve students' conceptual understanding of difficult ideas in quantum mechanics.

Conceptual understanding is an important aspect of physics education. However, there are also professional skills to develop like critical thinking, problem-solving, and scientific communication. Practicing scientists routinely encounter complex, ill-structured problems [8] in their work that require them to collect additional information, consider external constraints, and continuously reflect on their solutions [9–12]. Indeed, this type of "authentic" problem-solving is cited as one of the most important technical skills required of recent physics graduates [13]. Frequently, however, employers and education researchers report that graduating students are not prepared to solve these kinds of problems [14]. This aligns with studies in physics which suggest that graduate coursework is mainly focused on mathematical skills, rather than higher order reasoning and problem-solving skills [15].

There are several studies of problem-solving and critical thinking skills among undergraduate physics students (e.g., [16–18]). For example, a recent study at an elite private university found that ill-structured problem solving was rarely assessed (and hence, rarely practiced) in physics coursework [19]. A notable exception in that study was a capstone course focused on solving ill-structured problems, like the types of problems a student might encounter in research. However, there has been almost no work in this area at graduate level. Leak et al. characterized the strategies used by graduate students to address their research problems and

found that, if students encountered routine problems such as those they might see in coursework, they would simply consult external resources for the answer—most of their effort was spent on attempting to solve the open-ended problems for which there was no known answer or solution path [20]. Thus, they suggested instructors provide coursework that goes beyond routine problems, to provide better preparation for problem-solving students will face in their future careers.

Earlier studies have shown that physics assessments can often be solved with a routine application of equations and without conceptual understanding [21]. Yet, as part of the "hidden curriculum" of physics [22], students are expected to develop more robust problem-solving skills that transfer to novel situations, despite the well-documented difficulties with transfer (e.g., [23–26]). Indeed, the way physics is taught is often drastically different from the way physics is done in practice [27]. Physics teaching tends to be rooted in positivist epistemology [28], which frames the knowledge being delivered as absolute truth. Physics teaching has a "strong framing," meaning there is little flexibility in what ideas are disseminated and how [29]. The ideas are also often clearly defined and distinct from other topics, which Bernstein refers to as a "strong classification," and taught in a particular hierarchical sequence [30, 31]. This leaves little room for integrating ill-structured problems with less clearly defined parameters and solution spaces. One of the primary difficulties with teaching problem-solving is that process is typically not well-defined or is rooted in an implicit model that an instructor may have difficulty articulating (e.g. [32]). Though there are many models of problem-solving in physics (e.g., [33]), they are typically prescriptive rather than empirically derived. We will adopt the framework proposed by Price et al. [34], which describes problem-solving skills as 29 decisions to-be-made (see Appendix A). These decisions range from selecting a problem to presenting the solution. Some examples include identifying gaps in the field to find a problem to solve (Decision 2 [D2]), deciding on the best way to represent and organize information (D17), and determining the audience for communication, (D28). This framework found commonalities in the decisions made by "experts" in science, engineering, and medicine, where experts are described as successful practitioners with considerable experience working as faculty in highly rated universities or in technical positions in successful companies, but not necessarily the

most exceptional performers in their fields. These expert decisions seem to overlap with the strategies graduate students employ in research as identified by Leak et al., [20]. Notably, this framework for problem-solving does not reduce the process to a prescriptive, linear procedure. The decisions may not be made in a particular order, and the solver may revisit decisions multiple times in the problem-solving process because these decisions are typically made with limited information.

Given the apparent disconnect between assessment of problem-solving in physics classes and how physicists solve problems in practice, our research questions were:

1. What decisions are physics students being asked to make in their core graduate coursework?

2. What decisions do graduate physics instructors expect students to be able to make following their coursework?

3.2 Decisions in Graduate Physics Coursework

3.2.1 Methods

A sample of assignments from the first-year graduate physics curriculum in the 2021-2022 academic year at Auburn University was used to evaluate which problem-solving decisions were being assessed and, thus, that instructors were implicitly communicating as important to their students. Three of the course instructors were theorists and one was an experimentalist. These assignments constitute a significant portion of the practice these graduate students received in making these decisions in their first year, as they typically have not begun doing research at this point. The assignment types included in this sample were questions from the graduate doctoral exam (GDE; a written qualifying examination, see [35]) offered at the end of the first year, in-class exams (including final exams), in-class quizzes, and homework. Each course had 1 to 2 GDEs (1 if all students pass the first), 3 to 4 exams, 3 to 5 quizzes, and 5 to 11 homework assignments, with homework comprising the bulk of the problems students received for practice. Within each course, a given assignment type typically had the same number of problems each time, but the number of problems on each assignment type was different across courses. For

example, an exam in Classical Mechanics often had fewer questions than an exam in Statistical Mechanics. Within each course, assignments of the same type appeared to practice similar skills with different concepts, so we randomly selected one assignment to investigate from that category. For example, a homework assignment in Quantum Mechanics on the hydrogen atom had a similar structure to a homework assignment on spins. This sampling was done for each of the four core courses - classical mechanics (CM), electricity and magnetism (EM), quantum mechanics (QM), and statistical mechanics (SM) – for a total of 56 problems (Table 3.2.2) across 14 assignments.

3.2.2 Assessment Coding

Table 3.1: Assessment problems by assessment type and course (Classical Mechanics, CM; Electricity and Magnetism I, EM; Quantum Mechanics I, QM; and Statistical Mechanics, SM). A dash indicates the assessment was not used in that course.

Course	GDE	Exam	Quiz	Homework	Total
CM	5	4	7	2	18
EM	5	3	3	2	13
QM	5	3	-	4	12
SM	5	5	-	3	13
Total	20	15	10	11	56

The presence of each decision was coded for each problem. A decision was coded as either "required," "eliminated," "prompted," or "absent." Problems which require the solver to apply the decision to arrive at a solution were coded as required. Below is an excerpt from a problem requiring decision 16 (which calculations and data analysis are needed):

> *Steady current, I, flows down a long solid cylindrical wire of radius, R. Find the magnetic field inside and outside the wire if the current is uniformly distributed over the outside surface of the wire.*

Problems which make the decision for the solver were coded as eliminated. Below is an excerpt from a problem which eliminated decision 10 (which approximations and simplifications are appropriate):

Assume that the cavity is small enough so that P,E_0,and D_0 are essentially uniform.

Problems that instruct the solver to make a decision were coded as prompted. Below is an excerpt from a problem prompting decision 26 (how good is this solution):

Show that your results approach the classical solution at $k \leq k_B T/(\hbar c)$.

Decisions which were not relevant (i.e., not required, eliminated or prompted) to the problem were coded as absent.

Problems with multiple parts were treated as one problem. If a decision was present in only one part of the problem, the entire problem was given the corresponding code. For example, if part A required decision xx and decision xx was absent in part B, decision xx was coded as required for the problem. We found only one example where a decision was required in one part but received another code in a later part: a classical mechanics homework problem required applying decision 16 for four of five parts, but the decision was eliminated for one of five parts. Therefore, this problem was coded as required to represent the majority of the problem.

M.E.R and N.D.D. independently coded 10 problems and discussed discrepancies until an agreement was reached. There was an initial 75 percent interrater agreement. The disagreements were then discussed to a consensus. The remaining 46 problems had an average initial interrater agreement of 92 percent. While the initial interrater agreement increased as more problems were completed, decisions 4 and 5 continued to have low initial interrater agreement, averaging 45 percent across all 56 problems reported. This is not surprising, as Price et al. discussed how these decisions frequently co-occur [34]. As we detail below, this did not affect the overall results. The researchers thus discussed any differences until an agreement was reached for all codes.

Both codes, prompted and required, identified a problem where decisions had to be made to arrive at a solution. However, a decision was identified as prompted only 6 times compared to 147 times for required. Therefore, the codes prompted and required were condensed into a single code assessed, aligning with the categories “encountered” and “removed” used in Montgomery et al. [19].

3.2.3 Results

Per problem investigated, a median of 3 decisions were assessed, 0 decisions were eliminated, and 26 decisions were absent (Figure 3.1). Across all problems the maximum and minimum number of decisions assessed was 6 and 0, eliminated was 3 and 0, and absent was 29 and 25. Four decisions were assessed in at least two thirds of the problems: deciding on a specific plan for getting information (D15), deciding which calculations and data analysis (D16), deciding what predictive framework to use (D5), and deciding what are important features and information (D4). Pairs of decisions frequently appeared together because some decisions are closely related and often made at the same time. For example, D4 and D5 often occur together because the important features and information can indicate which predictive framework would be appropriate for the problem. The need to determine a plan for getting information (D15) and decompose a problem into sub-problems (D11) were commonly eliminated by providing multiple parts which guided the solver along a solution path. A need for evaluating one's solution (D26) was typically eliminated by problems instructing a student to check a particular limit but not explain why that supported or did not support their answer. Similarly, (D10) what approximations or simplifications are appropriate, was typically eliminated by stating assumptions for the student.

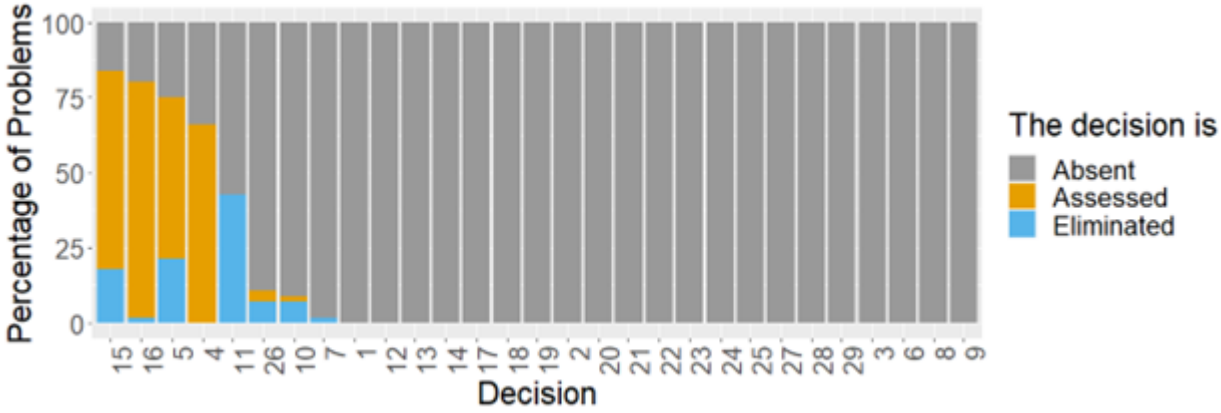


Figure 3.1: Percentage of problems versus decision. The decisions are numbered according to Price et al.'s framework. [34] Decisions were either absent, assessed, or eliminated. Assessed indicates the decision needed to be made to solve the problem. Eliminated indicates the problem statement made the decision for the solver. Absent indicates the decision was not assessed or eliminated.

Problems which required the solver to determine a plan for getting information (D15) typically also required the solver to determine which calculations are appropriate (D16). The statistical mechanics homework problem in Figure 3.2 was characteristic of a typical problem investigated, assessing three of the most common decisions while eliminating one. Important features (D4) needed to be identified to properly build the partition functions. Parts a and b are sub-problems necessary to solve part c. Therefore, breaking the problem into sub-problems (D11) has been eliminated for the solver. A plan (D15) and necessary calculations (D16) are required to solve part c.

(4 points) Hydrogen molecules can exist in ortho and para states. The spins of the two hydrogen nuclei couple to form a triplet state known as orthohydrogen, and a singlet state known as parahydrogen.

a. The rotational Hamiltonian of a parahydrogen molecule is given by $\mathcal{H}_p = \frac{\hbar^2}{2I} l(l+1)$, where $l = 0, 2, 4, \cdots$. Given that each energy state has a degeneracy of $2l + 1$, write down the rotational partition function of a parahydrogen molecule, Z_p, and evaluate its low- and high-temperature limits.

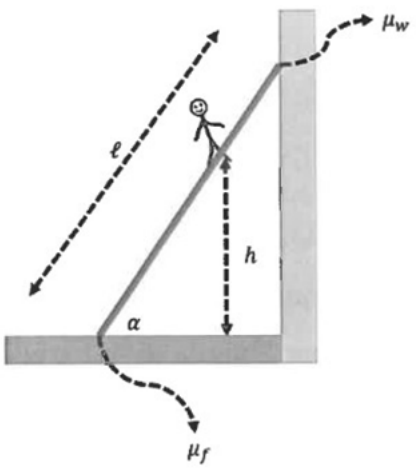
b. The rotational Hamiltonian of an orthohydrogen molecule is given by $\mathcal{H}_o = \frac{\hbar^2}{2I} l(l+1)$, where $l = 1, 3, 5, \cdots$. Given that each energy state has a degeneracy of $3(2l + 1)$, write down the rotational partition function of an orthohydrogen molecule, Z_o, and evaluate its low- and high-temperature limits.

c. Assuming the parahydrogen and orthohydrogen are in equilibrium at a temperature T, and their center-of-mass or vibrational partition functions of one molecule are the same, show that at the high-temperature limit, the ratio of the number density of orthohydrogen to parahydrogen is $n_o : n_p = 3 : 1$. (Hint: Helmholtz free energy is minimized in this equilibrium.)

Figure 3.2: A typical problem from the data set in terms of number of decisions assessed and eliminated. This Statistical Mechanics exam problem assessed three decisions while eliminating one.

In problems requiring the solver to select an appropriate predictive framework, D5, it was typically necessary to identify key features, D4. The classical mechanics qualifying exam problem shown in Figure 3.3 assessed the most decisions of any problem investigated, assessing four and eliminating one. (This problem was atypical because it was written by E.W.B.) In the problem statement a point mass approximation for the person, D10, is provided. Important features, D4, and a predictive model, D5, are necessary to begin solving the problem. The problem prompts checking multiple limits of the solution, D26.

Problem 1. Consider a person of mass M standing on a ladder of mass m (uniformly distributed) and length ℓ. The ladder is leaning against the wall, making an angle with the ground α. The coefficient of static friction between the ladder and the wall is μ_w and the coefficient of static friction between the floor and the ladder is μ_f. Treat the person as a point mass and consider the gravitational force.



(a) How high off of the ground (h) is the person able to stand before the ladder slides down the wall? Express your answer only in terms of the given variables and appropriate constants $h(\mu_w, \mu_f, \ell, \alpha)$.

(b) Verify that the ladder can still support the weight of the person in the limit of $\mu_w \to 0$ and $m \to 0$. What force supports the weight of the person in this case?

(c) What does your answer in (a) predict in the limit $\mu_f \to 0$? Does the prediction match what physically would happen? If not, please describe why.

(d) What does your answer in (a) predict in the limit $\alpha \to \pi/2$? Does the prediction match what physically would happen? If not, please describe why.

(e) What does your answer in (a) predict in the limit $\alpha \to 0$? Does the prediction match what physically would happen? If not, please describe why.

Figure 3.3: Atypically decision-rich problem. This Classical Mechanics qualifying exam problem assessed five decisions while eliminating one.

The classical mechanics quiz problem shown in Figure 3.4 assessed no decisions. This problem probed a conceptual topic that doesn't require any decisions to solve, as it is simply recalling a fact and not engaging in skill development. While not all quiz questions were multiple choice, they shared this focus on conceptual topics. Homework and quizzes were the only two assessment types with problems assessing no decisions.

If the coordinate q is cyclic, then:

The generalized momentum and the generalized velocity are both conserved.
The generalized momentum is conserved but the generalized velocity is not.
The generalized velocity is conserved but the generalized momentum is not.

Figure 3.4: Example problem from a Classical Mechanics quiz requiring no decision-making.

A Fisher's exact test was conducted to identify differences in the percentage of absent, assessed, and eliminated decisions by course (Appendix B). We did not find any statistically

significant variations (at the α = 0.05 level) in the percentage of absent, assessed, and eliminated decisions when disaggregated by course (p = 0.09). This means that all four subject areas assessed any of the decisions, eliminated any of the decisions, or didn't address any of the decisions at approximately equal rates. We also conducted a Fisher's exact test to investigate differences across assessment types; these differences were statistically significant (p< 0.001).

We then conducted a post-hoc analysis to investigate pair-wise differences between assessment types, shown in Table 3.2.3 (also Appendix B). We used a Bonferroni correction to account for inflated type 1 error, which resulted in threshold of α = 0.008 for statistical significance. A pair that is statistically significant suggests the percentages of absent, assessed, and eliminated problems is different between the two assessment types; whereas no statistical significance suggests the pair has a similar distribution. There were statistically significant differences between exams and quizzes (p < 0.001), homework assignments and quizzes (p < 0.001), and quizzes and the qualifying exams (p < 0.001). The difference between in-class exams and qualifying exams was not statistically significant (p = 0.35). There was no statistically significant difference between homework assignments and the qualifying exams (p = 0.027) or in-class exams and qualifying exams (p = 0.52). The difference between in-class exams and homework assessments was also not statistically significant (p = 0.015). The homework assignments tended to eliminate more decisions compared with other assessment types because they were intended to practice specific skills or concepts. This focus was achieved by providing additional direction (e.g., identifying assumptions to make or assigning a framework to apply); thereby eliminating more problem-solving decisions. Quizzes had more absent decisions than other assessment types. The quiz questions tended to probe basic conceptual knowledge and calculation skills, which limited the need for decision-making.

Table 3.2: Post-hoc analysis with Fisher's exact test. A Fisher's exact test was conducted for each pair of assessment types. The p values are reported. A Bonferroni correction was used resulting in a threshold of α = 0.008 for statistical significance. * indicates the difference between the pairs was statistically significant with a Bonferroni correction.

Assessment Differences	Qualifying Exam	Exam	Quiz
Qualifying Exam	-	-	-
Exam	0.52	-	-
Quiz	$< 0.001^*$	$< 0.001^*$	-
Homework	0.27	0.015	$< 0.001^*$

3.2.4 Discussion

First-year graduate students were assessed on four decisions frequently, and these decisions were similar across the core courses. These four decisions can readily appear in textbook-style problems, which represent most of the assessments investigated and aligns with the findings of Montgomery et al. [19]. Twenty of twenty-nine decisions were not practiced at all. This finding is consistent with the lack of problem-solving skills present in undergraduate physics [19] and the emphasis on content knowledge in physical science courses [36].

In-class exams and the qualifying exam had similar distributions of decisions, while quizzes differed from each assessment type. The quiz problems investigated typically probed one specific conceptual or mathematical proficiency, which was likely due to the shorter timeframes allotted for quiz problems. Therefore, more decisions were absent in homework problems. On the other hand, homework problems were typically broader problems which had the scopes limited to specific concepts or mathematical proficiencies. While these differences were not statistically significant, homework problems seemed to eliminate more decisions to target a specific goal. While these assessments have qualitative differences, the general absence of decision-making opportunities is a commonality and was consistent across all subject areas.

While many of these decisions were not required to solve the problems we analyzed, the problems could have been reworked to allow more room for decision-making, even within more traditional assessment formats. For example, considering how good the solution is (D26) can be applied in nearly all cases by asking students to explain the limitations of their solution at

the end. Students could also be offered an opportunity to explicitly explain where they had trouble (D25) rather than simply lose points. Many of these decisions may require a more intentionally designed question or assessment structure than that of a textbook-style problem. For example, determining appropriate conclusions based on data (D21) may require problem statements or solutions more involved than one numerical answer. To improve the practice of decision-making, courses may need to alter the existing course structure, assessment types, or types of problems provided. While decision-making can be practiced simultaneously with other course goals, time spent on these decisions needs to be diverted from other areas. Instructors must determine how to distribute practice to reflect the importance of their course goals.

3.3 Decisions Expected of Students

3.3.1 Methods

To investigate which decisions that instructors of graduate physics courses expect their students to be able to make, semi-structured interviews were conducted with 11 instructors. All instructors were unpaid volunteers who taught a core graduate-level physics course at a U.S. university in the five years prior to the interview. The instructors were recruited through the professional networks of the authors and were all tenured faculty at large research-intensive institutions. The research interests of the two experimentalists and nine theorists included atomic, molecular and optical (AMO) physics, condensed matter, particle physics, plasma physics, and space physics. Though this is heavily biased toward theoretical physicists, the framework of decision-making was derived to be applicable across theory, experiment, and applications among many STEM disciplines. The courses included CM, EM, QM, and SM. Three instructors taught CM, three taught EM, three taught QM, and five taught SM. Some instructors had experience teaching more than one of these courses. Though the course content varied slightly by institution and instructor, we were focused more on the expectations of the instructors for problem-solving skills among first-year graduate students. Additionally, we note that CM was only a required graduate course at one of the institutions.

M. E. R. conducted all interviews, each of which lasted between 45 and 75 minutes. The interview consisted of an open-ended and closed-ended portion. The open-ended portion was

conducted to probe which learning goals the instructors had in their courses that were not related to content knowledge (see protocol in Appendix C). This portion was conducted first to avoid biasing the responses towards our chosen theoretical framework of problem-solving as decision-making with limited information. This allowed the researchers to identify whether any of the instructors' stated goals for the introductory courses could not be categorized in terms of Price et al.'s framework [34]. After the instructor exhausted their ideas, the closed-ended portion began. In this portion, instructors were given the definitions of each of the decisions identified by Price et al. and asked to identify whether they would expect a physics graduate student to be able to successfully make that decision following their course. The interviewer clarified the meaning of each decision as needed. There was sometimes discussion of the decision between the interviewer and the instructor that resulted in the instructor's answer being interpreted differently than stated or an instructor changing their answer. Most instructors gave a clear binary (Yes or No) answer, but in some cases, the authors needed to infer this choice from the reasoning the instructors gave.

M.E.R. and E.W.B. independently coded each interview transcript; they did not identify any of the instructors' stated objectives in the open-ended portion that could not be characterized in terms of decisions-to-be-made. For the close-ended portion, each decision was coded as "expected of," (i.e. the instructor would expect a student to be able to make that decision at the end of their course) "not expected of," "unrelated," or "unanswered." The code unrelated occurred when instructors gave a response which did not address the current decisions, which was typically relevant to another decision. The code unanswered occurred when instructors did not provide answers about certain decisions, due to time constraints of the interviewee (one interview concluded after discussing only 10 of the 29 decisions) or when instructors stated they did not want to give a definitive answer after reading and discussing the details of the decisions with the interviewer.

3.3.2 Results

The instructors provided definitive answers for a median of 24 of the 29 decisions (i.e., expected-of or not-expected-of). Of the remaining five decisions, instructors gave responses that were

unrelated for two, and were unsure of the remaining three (both median values). A decision was only coded as unrelated or unanswered if no related response was found in either the closed- or open-ended portion of the interview. Five instructors received at least one unanswered code, leaving a median 2.5 decisions unanswered. 20 unanswered decisions were from one interview which ended early, hence our choice to use median values to reduce the impact of this outlier. No notable differences were noted by research interest or course taught. Below the decisions are discussed by varying levels of support by faculty. The results, shown in figure 3.5, did not produce clearly delineated groups of decisions. However, groups were used to report themes and present the results concisely.

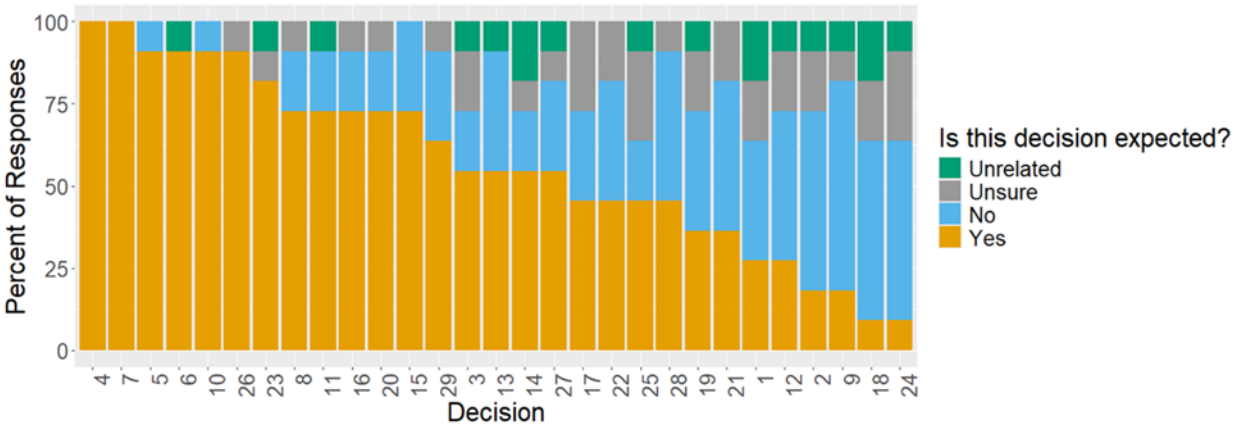


Figure 3.5: The percentage of codes for instructor responses to each decision.

High Support

The decisions which most of the instructors agreed that a first-year graduate student should be able to make following their core courses are given in Table 3.3.2. The instructors were unanimous that deciding on important features of the problem (D4) and deciding how the current problem is related to previous problems the students have solved (D7), were expected of first year graduate students. For D4, the general agreement was that identifying relevant versus irrelevant information and key features (e.g., symmetries or the curved nature of a system) is how problems are started to gain an understanding of what the problem is about. One instructor stated,

"You want to look at a physical situation and use your basic knowledge to say, what is this problem? What do I need to try and understand to come up with a solution?"

Table 3.3: Codes for high-support decisions. A number and description are included from Price et al.'s decision framework. [34] Expected of codes indicate the instructor expects students to make this decision; whereas expected of indicates the instructor does not. Unrelated was assigned for responses unrelated to the decision. Unsure or unanswered was assigned if the instructor chose to not give a response after discussing the decision with the interviewer.

Decision	Decision description	Code count			
		Expected	Not expected	Unrelated	Unsure or unanswered
4	What are the important underlying features or concepts to apply?	11	-	-	-
7	What are related problems or work seen before, and what aspects of their solving process and solutions might be useful in the present context?	11	-	-	-
5	Which potential predictive framework to use?	10	1	-	-
6	How to narrow down the problem?	10	1	-	-
10	What approximations or simplifications are appropriate?	10	1	-	-
26	How well does the solution hold?	10	-	-	1
23	Are previous decisions about simplifications and predictive frameworks still appropriate?	9	-	1	1
8	What are potential solutions?	8	2	-	1
11	How to decompose the problem into more tractable sub-problems?	8	2	1	-
16	What calculations and data analysis are needed?	8	2	1	-
20	If a result is different than expected, how should you follow up?	8	2	-	1
15	What is the specific plan for getting additional information?	8	3	-	-

This was similar to the reasoning of the other interview participants and aligns with Price et al.'s [34] description, which includes finding a suitable abstract representation of core ideas such as an equation in physics. When considering similar problems (D7), one participant remarked:

> *"It might be a problem you haven't experienced and or seen [a similar] problem that might get you started."*

Another participant qualified this response, distinguishing between memorizing problem-types and matching solution methods:

> *"Yes, with some caveats. You're not trying to pattern match and memorize a previous problem. You're trying to pattern match how did you solve that problem."*

These sentiments are similar to what Price et al. [34] describe as identifying what aspects of a related problem's problem-solving process and solutions might be useful.

Decisions 5, 6, 10, 23, and 26 had near-unanimous agreement, with 9 to 10 of 11 faculty indicating they would expect graduate students to be able to make these decisions following core coursework. what "predictive framework" or mental model is appropriate for the question (D5) was often justified by responses such as

> *"Yeah, when I teach physics, I ask students to have a model in their mind. It can be different than you thought, but you have to ask yourself why your approach was wrong. Don't just trust what you're deriving."*

This response mentions both the need for a guiding model or theory, but also the flexibility to revisit the appropriateness of that model later (which is a different decision, discussed below). One faculty member disagreed because they thought of predictive frameworks as "givens" in the context of coursework:

> *"You know, what you do at a core-graduate course is usually kind of the opposite. You have a framework, and you choose the problems that illustrate the frame. But you definitely want to tell them why the framework applies to this particular problem. So, what is it about this particular problem that I can take this tool that I have and use it on this. And so that when you see it again, you'll know what to do."*

This faculty member suggests that, in coursework, students should apply a given framework rather than selecting the appropriate framework for a given problem. Price et al., includes creating or deciding among possible predictive frameworks as making this decision, which is not conducted if the framework is provided [34]. However, Price et al.'s description aligns with the views of the faculty members in support.

> *"sometimes the approximations don't even have to be accurate. They just have to give you insight. So, an approximation can be inaccurate, but you can still learn something by analyzing under that approximation."*

Second, faculty suggested that students should understand when an approximation is appropriate or invalid:

> *"Start identifying situations where one approximation is good, or one approximation is not good. For example, if I have something that is near a minimum, it looks like a parabola, if I give too much energy and now it's not... near that minimum, will that solution still be the same?"*

One faculty member was hesitant to let students select approximations, since student-selected approximations may lead to solutions which obscure the desired findings of the problem:

> *"if I don't tell them take this approximation... They often will get started on things that seem not very interesting to me."*

Price et al., describes this decision as determining which approximations are appropriate, aligning with the second idea described by faculty [34]. The first idea describes a potential application of this decision.

D6, how to narrow down the scope of the problem, had a high consensus with 10 faculty saying they expected this of students and no faculty disagreeing. One faculty member attempted to relate this to a degrees-of-freedom analysis:

> *"We go from six degrees of freedom to a one-dimensional problem. Or learning about symmetries, if I know that this half is the same as this half then I only solve half of the problem."*

However, this description of taking a complicated problem and simplify it to something solvable is more closely aligned with D10. Price et al. suggest that D6 is more accurately characterized by behaviors such as focusing on one piece of a system; for example, a soft-matter physicist might only focus on the results of the motion of self-propelled particles rather than the chemical, optical, or electromagnetic stimuli that direct that motion [34].

Checking whether assumptions or simplifications are appropriate (D23) was identified as expected by all 9 faculty who answered the question. The responses support reflecting on work and evaluating approximations made, with one participant describing this process as having,

> *"something that looks like an appropriate solution, but when you think about it more carefully, they miss something. And the thing you've missed can lead you in a while different direction."*

Incorrect approximations occur, so identifying the approximation as incorrect during the solving process is necessary. Another response describes the utility of this skill in identifying false pattern matching:

> *"If it pattern matches it might be wrong to match it for a particular reason... it looks like you can do this simplification here; it worked fine in this case, but it doesn't work in this case."*

These statements, along with the responses for D10, suggests that faculty value students' abilities to not only make assumptions and simplifications, but to reflect on those choices and identify whether they were reasonable.

Similarly determining how well a particular solution holds (D26) was identified by all 10 faculty members who answered that question as something expected of graduate students following coursework. The responses were unified around the idea that checking your solutions, which can manifest as checking limits, checking units, proportional reasoning, etc., is an essential skill for physicists. One response described it as,

"look at your result; make sure the units make sense. Then check if the concepts are consistent with what you learned. Maybe it's when the distance goes up the quantity should go down, but [if] your quantity goes down, there's something wrong. So, you go back and check your calculations."

This aligns with the definition of Price et al., which includes exploring failure mode and limitations or applying a does it "make sense" test [34].

Decisions 8, 11, 16, and 20 were generally considered to be expected of graduate students, though each decision had two participants who dissented. Participants generally thought that identifying potential solutions (D8) was important because having an approximate answer in mind is beneficial:

"have an idea what you expect the answer to be before you solve the problem. Not the number but roughly ballpark." Others saw value in trying to predict the solution method: "maybe see a particular geometry and it brings up spherical harmonics."

The difference between predicting an answer or solution depends on the problem statement, but those in support see value in this approximate predictive ability. One faculty member who disagreed said,

"That works for the freshman level when you have two-step problems. At the graduate level, when you want to measure the probability of the system in a certain state... When you go to the graduate level it's generally a multi-step problem,"

suggesting that considering potential answers is too challenging for certain problems to be expected of students. Price et al., describes this decision as determining potential solutions based on experience or identified criteria of key features, aligning with the sentiments of those in support [34]. The faculty member opposed to the decision describes a case where this expertise may not exist.

Deciding how to decompose a problem into subproblems (D11) was generally supported because faculty believed that students should be able to both break up large problems into a set of smaller problems and then solve each smaller problem. One participant said,

"You know that to solve these problems you will have to break it into steps. You should be able to design this process."

Those opposed said that students are typically given the subproblems, removing the need to exercise this decision:

"generally the things you do in the first-year graduate courses, it's already [a subproblem]."

Price et al., describes this decision as decomposing a problem to tractable subproblems which can be independently solvable pieces, aligning with faculty in support [34].

Faculty generally expected students to be able to determine which calculations and data analysis are needed (D16), though some were thrown by the mention of "data analysis," which is not typically considered in lecture-based physics courses. The responses in support described representing the system in equations as important:

"You have to be able to relate [the mathematical techniques] to the physical concepts... Math is how we talk. You have to speak precisely. Only when you connect [math and physics concepts] are you a good physicist."

Responses opposed focused more on the data aspect of the decision:

"In our class it's only conceptual. Some teachers may ask them to do real world, give them a real set of data... it may not be that relevant."

These responses suggest that physics faculty view calculations and data analysis as separate ideas, despite the original problem-solving framework identifying them as different manifestations of the same cognitive process. Price et al., includes determining and conducting the necessary calculations to determine a solution, which is not exclusively data analysis [34]. Faculty responses align with the calculation aspect of this decision, but not the data analysis.

Faculty again were generally in support of students being able to identify significant anomalies in their answers (D20) Faculty in support expressed interest in students being able to determine the next steps after encountering an unexpected result:

"You rethink what you expected and look at the math, maybe your calculation is wrong... Then check if the concepts are consistent with what you learned. Maybe it's when the distance goes up the quantity should go down, but your quantity goes down, there's something wrong, so go back and check your calculations."

The instructors mentioned identifying areas in the work which could lead to this unexpected result, questioning the concepts applied and reassessing the methods used are expected follow-up strategies. While the exact methods suggested differed between faculty members, these methods attempt to identify the cause of the unexpected result. Price et al. describes this decision as identifying if the anomaly fits within an acceptable range of the predictive framework or acceptable levels of uncertainty and determining how to follow up on this anomaly. [34] This description aligns with what the faculty members expect of students. On the other hand, two responses suggested this decision is more appropriate for research or upper-level classes:

"This is really sort of a research skill... That would be nice to try in a quantum 2 course where there's more freedom. In quantum 1 there is some pressure to get through some material."

Their concern was the contexts that exist in the course are not involved enough to warrant making this decision.

Determining a specific plan for getting information (D15) was largely supported by faculty; however, 3 faculty indicated this decision was not expected of this level. One participant expressed their expectation is not for students to

"recognize immediately that I expect them to use this equation, but to have to think through for themselves what the [question] means [and] in which techniques and methods they should use."

A general expectation among those in support is for students to determine a set of steps and tools to apply to solve a problem. Faculty members who were opposed believed there was no additional information to solve for. One participant expressed this as,

"when we give [first-year graduate students] a problem, we give them enough information to solve it, so I don't think we do that in graduate course."

The main difference among the groups was the interpretation of "additional information." Price et al., suggest the solution to a problem is "additional information," aligning with faculty in support of this decision [34]. Responses opposing this decision were focused on a specific aspect of this decision.

Moderate Support

Nine decisions were expected by most instructors, but there was a more substantial number of opposing viewpoints compared to the previous decisions reported (Table 3.3.2). Instructors occasionally described ways these decisions were or could be conducted in certain course structures. For example, an instructor who incorporates a group project in their course shared their expectation for determining the best way to present work (D29) as

> *"in a [project-based assignment] they will need to clearly present their project to someone else... their peer would need to understand."*

However, a different instructor did not see a need to develop D29 before students began publishing papers:

> *"Those are important things when we train graduate students to write a paper."*

Not all responses depended on if the decision was already present in the instructor's current course. An instructor said that identifying the information needed to distinguish between potential solutions (D13) is

> *"something that should be a part of [the course] but probably isn't."*

They proceeded to describe an example to incorporate this decision, which involved distinguishing between perturbation theory and the variational method. Instructors opposed to these decisions believed these decisions were more appropriate in upper-level courses or research.

Table 3.4: Codes for moderate-support decisions. A number and description are included from Price et al.'s decision framework. Expected of codes indicate the instructor expects students to make this decision; whereas expected of indicates the instructor does not [34]. Unrelated was assigned for responses unrelated to the decision. Unsure or unanswered was assigned if the instructor chose to not give a response after discussing the decision with the interviewer.

Decision	Decision description	Code count			
		Expected	Not expected	Unrelated	Unsure or unanswered
29	What is the best way to present the work to have it understood and its correctness and importance appreciated?	7	3	-	1
3	What are your goals for this problem?	6	2	1	2
13	What information is needed to solve the problem?	6	4	1	-
13	What to prioritize among many competing considerations?	6	2	2	1
27	What are the range and contexts where solution applies, and what are the broader implications? What outstanding problems in field might it solve? What novel predictions can it enable?	6	3	1	1

Low Support

The remaining eight decisions received low support (i.e., a minority of instructors indicated they were expected of first-year graduate students; see Table 3.3.2). These decisions also had a higher number of "unsure" and "unrelated" responses. Instructors more commonly stated these decisions are not appropriate for a first-year course, may be appropriate for upper-level courses, and are valuable in research. A frequent justification for practicing these decisions later was the challenge of incorporating practice into a first-year course, due to the types of problems assigned. One instructor responded on their expectation of students identifying the best solution (D22) as

"to some degree, yes. But I don't think we have too many options... I don't think there's [much] room to choose [the] best."

The types of problems this instructor assigned didn't lend themselves to having multiple solutions. The instructor had an interest in students making this decision but thought incorporating this decision would be challenging. Similarly, this difficulty can arise from the course not including necessary datasets to practice representing and organizing information (D17):

"In order to really do something with them, you need more comprehensive data for them to analyze... not just one number in the end."

One faculty member stated that determining if additional knowledge is needed (D24) did not appear because

"in a course the problems are more thought out and should be solvable." "in a course the problems are more thought out and should be solvable."

This idea of the decision not being compatible with first-year problems was also used against evaluating new information (D19); determining appropriate conclusions form data (D21); determining if a problem is solvable (D9); and determining if the information is believable (D18). Arguments against the remaining decisions were that these decisions were beyond what is expected of students. For example, an instructor discussed identifying important questions and determining where the field is heading (D1),

"to be successful, as a physicist, you have to be thinking about question number one all the time, but for the course that's not the emphasis."

These decisions were again identified as important, but not at this level.

Table 3.5: Codes for low-support decisions. A number and description are included from Price et al.'s decision framework[34]. Expected of codes indicate the instructor expects students to make this decision; whereas expected of indicates the instructor does not. Unrelated was assigned for responses unrelated to the decision. Unsure or unanswered was assigned if the instructor chose to not give a response after discussing the decision with the interviewer.

Decision	Decision description	Code count			
		Expected	Not expected	Unrelated	Unsure or unanswered
17	What is the best way to represent and organize available information to provide clarity and insights?	5	3	-	3
22	What is the best solution?	5	4	-	2
25	How well is the problem-solving approach working and does it need to be modified including do the goals need to be modified?	5	2	1	3
28	What is the audience for communication?	5	5	-	1
19	As new information comes in, particularly from experiments or calculations, how does it compare with expected results?	4	4	1	2
21	What are appropriate conclusions based on the data?	4	5	-	2
1	What are important questions or problems? Where is the field heading? Are there advances in the field that open new possibilities?	3	4	2	2
12	Which are areas of particular difficulty and/or uncertainty in the solving process?	3	5	1	2
2	If and where are the gaps/opportunities to solve in field?	2	6	1	2
9	If the problem is solvable and its solution worth pursuing given the difficulties, constraints, risks, and uncertainties?	2	7	1	1
18	Is information valid, reliable, and believable?	1	6	2	2
24	Is additional knowledge/information needed?	1	6	1	3

3.3.3 Discussion

All the instructors interviewed agreed that each of the 29 decisions were important for graduate students to learn at some stage of their career. As this set of decisions is theorized to be a complete description of the scientific problem-solving process, this suggests that scientific problem-solving is an important goal of physics graduate education. This is unsurprising given the extensive testing of the decision-making framework that Price et al. conducted with a panel of scientists, engineers, and physicians during their coding process [34]. We identified a subset of 11 decisions that the instructors thought a physics graduate student should be able to make by the end of their first year. This subset includes important features and information (D4), what predictive framework (D5), related problems (D7), potential solutions (D8), approximations or simplifications (D11), how to decompose into sub problems (D11), specific plan for getting information (D15), which calculations or data analysis (D16), any significant anomalies (D23), and how good is the solution (D26). Montgomery et al. (2023) found that typical undergraduate physics courses typically only offer room for making three of the 29 decisions. As our faculty identified 11 of the decisions as appropriate for first-year graduate students, this suggested the possibility that graduate physics courses may incorporate more decision-making than undergraduate physics courses. However, it is also not uncommon for instructors' intended learning outcomes to be misaligned with their assessments [21].

The remaining 18 decisions did not receive a strong consensus in support; however, instructors highlighted the value of these decisions as professional or upper-level courses. Some decisions were not expected of students in a classroom setting because they were either not seen as appropriate for first-year students, or they were not seen as feasible to implement in the classroom. For example, how to represent and organize information (D17) was mentioned as hard to incorporate without introducing comprehensive data to analyze. Additionally, some instructors did not know how some of these decisions could be taught or practiced in a class, e.g., an instructor mentioned the importance of additional knowledge needed (D24), and the difficulty of trying to incorporate it because their first-year problems contain all the relevant information. Their idea of typical first-year problems either removed or did not include necessary elements to make these decisions necessary. For some decisions, instructors were hesitant to

give a definitive answer (see Low Support above), which could have been due to fatigue (being later in the interview) or trouble understanding the decision itself. This is likely a limitation of our methodological choice to not define the decisions for the faculty members, or to provide examples from Price et al. [34]. Because the purpose of this study is not to quantify which decisions are supported and by how many faculty more generally, it does not affect the main takeaways. Implications for future research are discussed below.

This subset of expected decisions is consistent with solving problems that have clearly defined goals, all necessary information provided, and answers requiring limited analysis. The responses from the instructors in this study seem to support previous findings which argue that physics classes tend to have "curricular coherence," which is where students learn in a structured fashion that builds layers of knowledge sequentially [30, 31]. Bernstein argues that physics classes exhibit both "strong classification" and "strong framing," meaning the boundaries between core physics concepts and the organization of curricular structure in physics tends to be rigid [29]. Indeed, many studies have shown that physics classroom instruction often causes students to shift their epistemologies more toward classroom physics and away from expert conceptions of doing physics [37]. Sin posits that graduate students begin to experience the uncertainty inherent in physics knowledge, amidst a social-constructivist epistemology [38], when working on research projects [27].

Sin claims that physics instruction exhibits a positivist epistemology [27], which places knowledge as absolute truth with little room for interpretation. The positivist epistemology stands in stark contrast to how physics is practiced as a profession, which involves making sense of nature amidst uncertainties. Indeed, Price et al. define an "authentic problem," as one not only where the solution path is not known to the solver, but one where it is not known whether a solution even exists [34]. Their framing of problem-solving as making decisions with limited information, which suggests the possibility of multiple solutions and/or solution paths, clashes with notion of the physics classroom as positivist. One could interpret our findings as in support of this framework of physics instruction, in that there is little room for decision-making. However, as many of the instructors knowledge, there are often logistical and practical challenges that may prevent the incorporation of decision-making into these classes, such as

limited instructional time, the difficulty this would cause with grading, or limited resources for the instructors to develop such activities.

3.4 General Discussion

We attempted to identify the extent to which graduate students were able to practice decision-making in their core coursework. We found that four decisions were regularly practiced, but that eleven decisions were identified by instructors as something they would expect of their first-year students. These findings suggest there is a mismatch between the expected learning outcomes and the practice conducted in first-year courses. This is similar to findings in instructional design literature, wherein faculty seem to have expectations for student outcomes that are not being explicitly assessed (and therefore, not practiced). This is consistent with Redish's notion of the "hidden curriculum" of physics – there are parts of physics enculturation and learning that are not made explicit [22]. Many experts are clearly developing these decision-making skills despite not being made explicit. These findings highlight first-year courses as an area to better develop these skills, especially for those who otherwise may not have figured out the "hidden curriculum." Explicitly assessing these skills may allow instructors to better gauge incoming preparation and adapt teaching and practice to their students' needs. Future investigations should also focus on what graduate instructors expect students to be able to do coming into their courses versus leaving their courses.

Both analyses reinforce the strong framing and strong classification associated with physics instruction. The instructors' notions that some problem-solving skills are only important at later stages in students' career development shows the rigid conceptions of the progression of learning physics (strong framing). Similarly, the thoughts that some decisions were impractical to incorporate into coursework shows evidence of strong classification, e.g., what should or should not be in a physics course. This is reminiscent of the trajectory of expertise development illustrated by Schwartz et al. [39] They conceptualized "adaptive expertise" (closely related to Price et al.'s framework [34]) as a trajectory in a two-dimensional space of "efficiency" and "innovation." In this model, students need to both be fluent in basic skills like calculations and concepts and be able to apply that in novel and creative ways to become 'adaptive experts.'

They suggest an approach where you make incremental progress along both axes in small increments (e.g., become fluent in solving differential equations, then using that to build a model of a novel electromagnetic system). However, the rigid nature of physics curricula seems to suggest that students need to become completely efficient before being engaged in innovation. That is, there is a core set of physics knowledge, mathematical techniques, and canonical problem types that must be mastered before applying those ideas to novel situations which require decision-making.

The assessments analyzed appeared to be content focused, including limited problem-solving decisions to be made, consistent with this positivist notion of physics learning. The similarities between decision-making in first-year graduate courses and upper-level undergraduate courses are not entirely surprising, as they are often one to two years apart, and at this institution, may have substantial content overlap. However, faculty's agreement that all the decisions were important things for graduate students to eventually learn – either through advanced coursework or research – emphasize the need for students to develop proficiency with these decisions. While the instructors seemed skeptical of doing this within courses, Montgomery et al. illustrated how a capstone course at a highly selective university offered much more room for decision-making (26 of the 29 decisions) [19]. There are also other examples, including one from engineering that show so-called "cornerstone" design-based courses can also develop decision-making skills [40].

3.4.1 Limitations and Future Work

There are limitations to this investigation. First, the assessments investigated were from a single institution in one academic year, and thus may not represent the full extent of decision-making in graduate physics coursework more broadly. However, these assignments were written by physics faculty from a wide variety of subfields, experience with teaching, and graduate degrees from a wide variety of institutions across the world. Similarly, we have no reason to believe that the extent of decision-making practice would be any higher or lower at this institution compared with others. Indeed, similar results from Stanford's upper-level undergraduate courses were reported by Montgomery et al. [19] Secondly, a small sample of eleven faculty

were interviewed for the second piece of this study, again reflecting a limited set of viewpoints about the place of decision-making in graduate coursework. Though this sample of faculty represented multiple institutions, the precise estimate that 11 decisions are expected is likely not an accurate number. A more reasonable conclusion is that instructors do not expect first-year graduate students to be completely proficient scientific problem-solvers by the end of their first year, but that these expectations are likely shaped by a pervasive positivist standpoint in physics education, rather than reflecting any limitations in the capabilities of first year graduate students or the feasibility of incorporating more decision-making into coursework. However, this does suggest that instructors expect more problem-solving skills of their students than they are given the students opportunities to practice or demonstrate.

It would be valuable to extend this study to a wider population of graduate physics faculty nationally as a way of characterizing the state of graduate physics education. This is particularly important as emphasis in education shifts toward adaptive expertise and problem-solving skills, rather than mathematical routines which may now be largely automated. An extended study may also investigate instructor course goals more broadly to understand the relative importance of problem-solving skills compared to mathematical fluency, conceptual understanding, etc. Given some of the uncertainty and unrelated answers among faculty respondents, we would recommend that the survey not only provide the definition of the decision, but one or two examples of how that might manifest in physics.

There remain substantial challenges to redesigning graduate physics instruction to emphasize decision-making. First, educational change literature indicates that unseating the idea that physics instruction is supposed to be done a certain way will be difficult (e.g., [32, 41]). Second, designing static tasks that practice and assess many of these decision-making skills is difficult for instructors to learn how to do. This is again, in part, due to their mindset that physics classes are executed in a certain way. Wieman and Price have provided general guidance for writing such problems in a recent book chapter [42]. We are also in the process of developing and testing such problems to use in graduate quantum mechanics to provide more concrete examples to physics faculty, and plan to eventually develop a suite of assessments and activities suitable for a graduate-level physics curriculum.

3.5 Acknowledgements

We would like to thank all the instructor participants for volunteering their time.

References

[1] Jennifer L. Docktor and José P. Mestre. “Synthesis of discipline-based education research in physics”. In: *Physical Review Special Topics - Physics Education Research* 10 (2 Sept. 2014). ISSN: 15549178. DOI: 10.1103/PhysRevSTPER.10.020119.

[2] Lillian C. McDermott and Edward F. Redish. “Resource Letter: PER-1: Physics Education Research”. In: *American Journal of Physics* 67 (9 1999). ISSN: 0002-9505. DOI: 10.1119/1.19122.

[3] Chandralekha Singh and Emily Marshman. “Review of student difficulties in upper-level quantum mechanics”. In: *Physical Review Special Topics - Physics Education Research* 11 (2 2015). ISSN: 15549178. DOI: 10.1103/PhysRevSTPER.11.020117.

[4] Chandralekha Singh. “Student understanding of quantum mechanics at the beginning of graduate instruction”. In: *American Journal of Physics* 76 (3 Mar. 2008), pp. 277–287. ISSN: 0002-9505. DOI: 10.1119/1.2825387. URL: https://pubs.aip.org/ajp/article/76/3/277/1056828/Student-understanding-of-quantum-mechanics-at-the.

[5] Guangtian Zhu and Chandralekha Singh. “Improving students’ understanding of quantum measurement. I. Investigation of difficulties”. In: *Physical Review Special Topics - Physics Education Research* 8 (1 2012). ISSN: 15549178. DOI: 10.1103/PhysRevSTPER.8.010117.

[6] C. D. Porter and A. F. Heckler. “Graduate student misunderstandings of wave functions in an asymmetric well”. In: *Physical Review Physics Education Research* 15 (1 2019). DOI: 10.1103/physrevphyseducres.15.010139.

[7] L. D. Carr and S. B. McKagan. “Graduate quantum mechanics reform”. In: *American Journal of Physics* 77 (4 2009). ISSN: 0002-9505. DOI: 10.1119/1.3079689.

[8] Herbert A. Simon. “The structure of ill structured problems”. In: *Artificial Intelligence* 4 (3-4 Dec. 1973), pp. 181–201. ISSN: 0004-3702. DOI: 10.1016/0004-3702(73)90011-8.

[9] BET. *Criteria for accrediting engineering programs 2019-2020*. 2019.

[10] Clive L Dym. *Design, Systems, and Engineering Education*. Tech. rep. 2004, pp. 305–312.

[11] Clive L. Dym et al. "Engineering design thinking, teaching, and learning". In: *Journal of Engineering Education*. Vol. 94. Wiley-Blackwell Publishing Ltd, 2005, pp. 103–120. DOI: 10.1002/j.2168-9830.2005.tb00832.x.

[12] David Jonassen, Johannes Strobel, and Chwee Beng Lee. "Everyday problem solving in engineering: Lessons for engineering educators". In: *Journal of Engineering Education* 95 (2 2006), pp. 139–151. ISSN: 10694730. DOI: 10.1002/j.2168-9830.2006.tb00885.x.

[13] Mulvey P and Pold J. *Physics Doctorates: Skills Used and Satisfaction with Employment. AIP Statistical Research Center*. 2020.

[14] Quacquarelli Symonds. "The global skills gap in the 21st century". In: *Retrieved July* 20 (2018), p. 2021.

[15] Chandralekha Singh and Alexandru Maries. "Core graduate courses: A missed learning opportunity?" In: *AIP Conference Proceedings*. Vol. 1513. 2013. DOI: 10.1063/1.4789732.

[16] Patricia Heller, Ronald Keith, and Scott Anderson. "Teaching problem solving through cooperative grouping. Part 1: Group versus individual problem solving". In: *American Journal of Physics* 60 (7 1992). ISSN: 0002-9505. DOI: 10.1119/1.17117.

[17] N. G. Holmes, Carl E. Wieman, and D. A. Bonn. "Teaching critical thinking". In: *Proceedings of the National Academy of Sciences of the United States of America* 112 (36 2015), pp. 11199–11204. ISSN: 10916490. DOI: 10.1073/pnas.1505329112.

[18] Eric Burkholder, Lena Blackmon, and Carl Wieman. "Characterizing the mathematical problem-solving strategies of transitioning novice physics students". In: *Physical Review Physics Education Research* 16 (2 Nov. 2020). ISSN: 24699896. DOI: 10.1103/PhysRevPhysEducRes.16.020134.

[19] Barron J. Montgomery, Argenta M. Price, and Carl E. Wieman. “How traditional physics coursework limits problem-solving opportunities”. In: *2023 Physics Education Research Conference Proceedings*. American Association of Physics Teachers, Oct. 2023, pp. 230–235. DOI: 10.1119/perc.2023.pr.Montgomery. URL: https://www.per-central.org/items/detail.cfm?ID=16588.

[20] Anne E. Leak et al. “Examining problem solving in physics-intensive Ph.D. research”. In: *Physical Review Physics Education Research* 13 (2 2017). ISSN: 24699896. DOI: 10.1103/PhysRevPhysEducRes.13.020101.

[21] Leslie O Dickie. “Approach to learning, the cognitive demands of assessment, and achievement in physics”. In: *The Canadian Journal of Higher Education* 33 (1 2003).

[22] Edward F. Redish. “Introducing students to the culture of physics: Explicating elements of the hidden curriculum”. In: *AIP Conference Proceedings*. Vol. 1289. 2010, pp. 49–52. ISBN: 9780735408449. DOI: 10.1063/1.3515245.

[23] Jose P. Mestre. *Physics Education You may also like Expanding physics learning beyond classroom boundaries-a case study*. Tech. rep. 2001, p. 44.

[24] Kara E. Gray and N. Sanjay Rebello. “Transfer between paired problems in an interview”. In: *AIP Conference Proceedings*. Vol. 790. Sept. 2005, pp. 157–160. ISBN: 0735402817. DOI: 10.1063/1.2084725.

[25] Darryl J. Ozimek et al. “Retention and transfer from trigonometry to physics”. In: *AIP Conference Proceedings*. Vol. 790. Sept. 2005, pp. 173–176. ISBN: 0735402817. DOI: 10.1063/1.2084729.

[26] Joanne Lobato. *Alternative perspectives on the transfer of learning: History, issues, and challenges for future research*. 2006. DOI: 10.1207/s15327809jls1504_1.

[27] Cristina Sin. “Epistemology, sociology, and learning and teaching in physics”. In: *Science Education* 98 (2 2014). ISSN: 00368326. DOI: 10.1002/sce.21100.

[28] Norman G. Lederman. “Students’ and teachers’ conceptions of the nature of science: A review of the research”. In: *Journal of Research in Science Teaching* 29 (4 1992), pp. 331–359. ISSN: 10982736. DOI: 10.1002/tea.3660290404.

[29] Basil Bernstein. “On the classification and framing of educational knowledge”. In: *On the classification and framing of educational knowledge*. Vol. 3. Routledge and Kegan Paul, 1975, pp. 85–115.

[30] Lisa R. Lattuca and Joan S. Stark. “Will Disciplinary Perspectives Impede Curricular Reform?” In: *The Journal of Higher Education* 65 (4 1994). ISSN: 00221546. DOI: 10.2307/2943853.

[31] Janet G. Donald. “Knowledge Structures: Methods for Exploring Course Content”. In: *The Journal of Higher Education* 54 (1 1983). ISSN: 00221546. DOI: 10.2307/1981643.

[32] Charles Henderson. “The challenges of instructional change under the best of circumstances: A case study of one college physics instructor”. In: *American Journal of Physics* 73 (8 2005). ISSN: 0002-9505. DOI: 10.1119/1.1927547.

[33] F. Reif and Joan I. Heller. “Knowledge Structure and Problem Solving in Physics”. In: *Educational Psychologist* 17 (2 1982). ISSN: 15326985. DOI: 10.1080/00461528209529248.

[34] Argenta M. Price et al. “A detailed characterization of the expert problem-solving process in science and engineering: Guidance for teaching and assessment”. In: *CBE Life Sciences Education* 20 (3 2021). ISSN: 19317913. DOI: 10.1187/cbe.20-12-0276.

[35] Shiva Basir and Eric Burkholder. “Investigating faculty perspectives on written qualifying exams in physics”. In: *Physical Review Physics Education Research* 20 (1 Jan. 2024). ISSN: 24699896. DOI: 10.1103/PhysRevPhysEducRes.20.010139.

[36] John M. Braxton, Nick Vesper, and Don Hossler. “Expectations for college and student persistence”. In: *Research in Higher Education* 36 (5 1995). ISSN: 03610365. DOI: 10.1007/BF02208833.

[37] W. K. Adams et al. “New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey”. In: *Physical Review Special Topics - Physics Education Research* 2 (1 Jan. 2006). ISSN: 15549178. DOI: 10.1103/PhysRevSTPER.2.010101.

[38] A Sullivan Palincsar. *SOCIAL CONSTRUCTIVIST PERSPECTIVES ON TEACHING AND LEARNING*. Tech. rep. 1998, pp. 345–75.

[39] Daniel L Schwartz, John D Bransford, David Sears, et al. "Efficiency and innovation in transfer". In: *Transfer of learning from a modern multidisciplinary perspective* 3.1 (2005), pp. 1–51.

[40] Eric Burkholder, Lisa Hwang, and Carl Wieman. "SUPPORTING AUTHENTIC PROBLEM-SOLVING THROUGH A CORNERSTONE DESIGN COURSE IN CHEMICAL ENGINEERING". In: *Chemical Engineering Education* 55 (3 2021). ISSN: 21656428. DOI: 10.18260/2-1-370.660-126222.

[41] Melissa Dancy et al. "Physics instructors' knowledge and use of active learning has increased over the last decade but most still lecture too much". In: *Physical Review Physics Education Research* 20 (1 Jan. 2024). ISSN: 24699896. DOI: 10.1103/PhysRevPhysEducRes.20.010119.

[42] C Wieman and A. Price. "Assessing complex problem-solving skills through the lens of decision making". In: *Innovating Assessments to Measure and Support Complex Skills*. Ed. by N Foster and M Piacentini. OECD Publishing, 2023.

Chapter 4

An Assessment of Expert Decisions in Graduate Quantum Mechanics

Michael E. Robbins*, Gabriel J. DiQuattro, Eric W. Burkholder

Department of Physics, Auburn University, Auburn AL 36830 USA

*mer0031@auburn.edu

Manuscript in press at Physical Review Physics Education Research: Investigating and Improving Quantum Education through Research [Special issue]

4.1 Introduction

Graduate education has received relatively little attention in physics education research compared with the vast literature on undergraduate education. Most research on physics graduate programs consists of national statistics on PhD completion rates [1], admissions data [2, 3], demographics of degree recipients, and career trajectories of recent graduates [4]. While these data provide valuable insight, they say little about the quality of graduate teaching – i.e., what are our students learning in their graduate education? There have been a few studies looking at graduate students' conceptual understanding in Quantum Mechanics (QM), which have shown that, like undergraduate education, didactic instruction results in small learning gains compared with research-based instructional methods [5].

Despite the static nature of the goals and structure of graduate programs, Harshman [6] highlights that those may not be serving students well. In particular, graduate education may not prepare students for careers outside of academia [7]. According to national statistics, only 49% of physics PhD recipients will be in an academic career after graduating, including postdoctoral appointments [4]. Thus, the current structure of graduate programs is tailored to only a fraction of the student population. Furthermore, among those who work in the private sector, only 34% of graduates do work related to physics. Many may end up in finance or data science, for example. However, over 90% of PhD recipients, regardless of sector, report that they solve technical problems daily [4]. We thus posit that the true goal of graduate programs should be to teach students transferrable real-world problem-solving skills. Some may argue that programs are already doing that, but there is little to no rigorous research to support such claims. Developing a practical and accurate way to measure problem-solving is a key step in designing activities that develop this skill. Instructors need to (1) identify learning goals, (2) design ways to measure whether students have achieved those goals, and (3) design activities to support those goals. In the context of scientific problem-solving, we have addressed point (1) in a previous study [8]. The focus of the current study is point (2), which will lay the groundwork for the design of educational interventions (3).

For the current study, we focus primarily on an assessment of problem-solving in QM. The assessment measures the ability to make expert decisions (discussed in Theoretical Framework) and is intended for first-year graduate students in QM. Researchers have recently argued that problem-solving cannot be measured independently of disciplinary knowledge [9, 10] because authentic problem-solving requires the application of discipline-specific knowledge [11–13]. Physics somewhat uniquely among other science disciplines maintains a core set of canonical knowledge in graduate education ("strong framing" [14]): classical mechanics, statistical mechanics, electromagnetism, and QM. Students often have difficulties in learning QM compared with these other subjects because of its drastically different knowledge structure compared to classical systems [15]. As the United Nations International Year of Quantum declaration indicates, being able to use and apply QM is going to be an essential skill for future physicists. We thus focus this study on problem-solving in QM but discuss the implications in relation to other physics domains at the end of this article.

4.2 Literature Review

4.2.1 Definition of Problem Solving

Problem-solving has long been studied in discipline-based education research (DBER) and cognitive science [12, 16–20]. It is defined in the 2012 National Academies report on DBER as being "required whenever there is a goal to reach, and attainment of that goal is not possible either by direct action or by retrieving a sequence of previously learned steps from memory" [21, 22]. Researchers further draw a distinction between well-defined and ill-defined problems [19]. In well-defined problems the initial conditions, goal, and constraints on the solution are all specified clearly. In ill-defined problems, students may have to define problem-components on their own [20] and the means of generating the solution may be unclear. The latter are the kinds of problems students will have to solve when they enter the scientific workforce.

PISA [23] defines problem-solving competency as "an individual's capacity to engage in cognitive processing to understand and resolve problem situations where a method of solution is not immediately obvious. Heller defines problem solving as "the cognitive process of moving towards a goal when the path is uncertain" [16]. Similarly, Adams and Wieman define

problem-solving as "cognitive processing directed at achieving a goal when no solution method is obvious to the problem solver" [17]. All three definitions emphasize that for true problem solving to occur, the solver does not know, at least initially, how to solve the problem. This suggests an important distinction between "authentic problems" and "exercises". Typical textbook problems are "exercises" because to solve them, students are required to apply a known procedure. The designation of a task as an "authentic problem" is thus dependent on the solver: if they know how to solve it, it is, by definition, not an authentic problem.

To understand how these real-world problems contrast with a typical physics problem ("exercise") see Table 4.2.1. Though the textbook problem is difficult (and at the graduate level), it offers no room for students to make decisions. The students are told which model to use (first order degenerate perturbation theory), what the goal of the problem is (solve for corrections to the eigenvalues), and, though the procedure is difficult and involves multiple integrals, the solver is simply executing a well-defined procedure. Indeed, Price et al. [13] extend the definition of authentic problem-solving to include the possibility that a solution may not exist at all, unlike previous frameworks.

Table 4.1: Example textbook problem (left) versus an authentic problem written by an atomic physicist (right).

Exercise	Authentic Problem
Using first order degenerate perturbation theory evaluate the correction to the eigenstates of the n=2 levels of hydrogen in a constant electric field that is directed along the z-axis of the atom.	You wish to look for evidence of nuclear decay of elements generated in supernova explosions. The large amounts of neutral cobalt isotopes generated in the supernova will undergo nuclear decay into nickel via beta radiation. If the cobalt atoms are in the ground state, what eigenstates would you expect for the resulting nickel atoms after the nuclear decay and how might you detect their presence?

4.2.2 Expert-Novice Differences in Problem Solving

Much of the work on physics problem-solving has studied the differences between novice and expert problem-solvers [16, 17, 24–28]. Many studies have shown that students often learn how to solve quantitative problems by plugging values into algorithmic equations and pattern

matching, and thus are not developing the necessary skills to transfer their understanding to unseen situations [29–34]. There is also substantial research on the use of representations in physics problem-solving [19, 25, 35, 36]. Representations describe the depth of consideration of the problem-solver (are they representing only the surface features of the problem or are they representing the deeper physics?). Representations also refer to ways in which the problem-solver is depicting the problem; pictures, free body diagrams, equations, graphs [25]. Experts typically exhibit a pattern of representations whereby they first construct a pictorial (or diagrammatic) representation, then a physical representation (like a free body diagram), and finally a mathematical representation. Experts have been found to be particularly skilled in transitioning between these various representations [35]. Experts engage in more analysis while "novices are more likely to behave mechanically or algorithmically, producing multiple representations without being able to make much use of them" [25].

Several studies have found differences in both the knowledge structure and the problem-solving strategy typically used by experts compared to novices. Reif & Heller found that the main difference between novices and experts was how they organized and used their knowledge in the context of solving a problem [34]. Experts organize their knowledge in a structured, cohesive way and can activate these knowledge structures when they are needed. Novices typically do not have these knowledge structures; rather their knowledge consists of random facts and equations that are context-specific and lacking in conceptual meaning [37, 38]. Experts redescribe the problem and use qualitative arguments to plan solutions before describing the details of the problem from a mathematical perspective. Novices tend to rush to quickly string together various, miscellaneous mathematical equations [34].

Adams and Wieman found that, when solving problems, experts spend more time analyzing, planning, and managing their own behavior than novices and generally, demonstrate a more holistic and systemic approach. Heller [16] found that experts usually follow specific steps when solving a problem: understanding the problem, determining the concepts, making the plan, solving the problem, and evaluating the outcome. Novices, on the other hand, first try to solve problems by using mathematical expressions. Expert problem solvers take more time

to understand the problem and the concepts involved as well as to explore the relationship between concepts. Novice problem solvers cannot establish these relationships, especially when the problem is complex [16].

4.2.3 Measuring Problem Solving

In 2012, the National Academies identified a "pressing need" for measurement tools that assess problem-solving skills at scale. At that time, there were some existing measurement tools (based on prescriptive problem-solving models) in physics [39] and chemistry [40], but they were not widely used. Since then, several assessments that measure problem-solving or critical thinking have been developed, but most share an assumption that problem-solving can be measured independently of content knowledge.

Some researchers argue that expertise involves the application of discipline-specific knowledge, and thus, higher-order skills like problem-solving and critical thinking cannot be measured independently of disciplinary knowledge and skills [9–11]. Most existing assessments implicitly rely on some basic content knowledge such as static equilibrium and would thus only measure problem-solving in that context. To our knowledge, there is no research on how well these measured problem-solving skills correlate with problem-solving in other domains. To say that problem-solving cannot be measured independently of content knowledge is not to say that it is a non-transferrable skill. Indeed, many problem-solving frameworks focus on how the solver draws upon their disciplinary content knowledge to solve the unknown problem at hand. Thus, we are arguing that problem-solving skills are transferable, but whether a student will be able to transfer those skills to a new scenario will depend on them possessing the relevant content knowledge. That is, a physics student skilled in problem-solving would be able to transfer these skills from QM mechanics to statistical mechanics provided that the requisite content knowledge is there.

Most studies involving the measurement and assessment of problem-solving use interviews and/or think-aloud protocols. Students and experts are interviewed by researchers who pose questions designed to understand the approach being taken to solving a problem. Think-aloud is a semi-structured cognitive interviewing method in which a person is asked to verbalize

their thought process as they do a specific task, during which they are recorded (on paper, audio or video) for further analysis [41]. In the study published by Ali et al., 21 students were asked to narrate their thinking while solving physics problems and all data were recorded [42]. Afterwards, qualitative interviews were held with the students. Interviews (or think-aloud recordings) are typically transcribed, and the text is analyzed and categorized (or coded) by several experts in the subject-matter domain. Interrater reliability is monitored by measuring the level of agreement between the expert coders to ensure validity. The interviews and generation of codes are often iterative. Price et al., for example, interviewed 22 science and engineering experts about problem-solving expertise in their fields [13]. From the interviews, they generated a list of decisions commonly made by experts while they are solving a problem. This list was then refined and validated via a set of 31 semi-structured interviews. These interviews were then coded for the decisions represented to characterize the problem-solving process.

There are some examples of assessments which use puzzle scenarios outside the realm of disciplinary knowledge in STEM to measure problem-solving, such as the Colorado Assessment of Problem-Solving (CAPS), developed by Adams [43], which measures strengths and weaknesses on the forty-four components of problem-solving identified by Adams and Wieman [17]. CAPS does not rely on discipline-specific knowledge and thus does not measure problem-solving expertise in the way we conceptualize it, and it is unclear how these measurements correspond to problem-solving aptitude on authentic problems.

Rubrics are another common measurement tool used to evaluate students' written work and are often used as part of the methodology in problem solving studies. Halim et al. [44] studied students' ability to apply problem solving strategies in physics. Students were asked to solve routine problems on paper and rubrics were used as the measurement tool. Burkholder et al. developed a solution template for students to use as they solve problems [45]. Responses to students' problem solving using this template were collected and then the template was used as a rubric for measuring how expert-like students' problem-solving thinking was. Docktor and Heller developed a rubric for assessing problem-solving skills in physics for which there is evidence of validity and reliability [39]. This instrument is based on the prescriptive problem-solving process outlined in Heller et al. [46]. Indeed, this instrument has also been shown to be

effective when assessing "context-rich" problems, which more closely align with the authentic problems students will have to solve upon graduation. The drawback to using rubrics is that they are static, and thus do not capture how students react when provided with new information about the problem they are solving.

4.2.4 Student Difficulties with Quantum Mechanics

Many studies on QM identify conceptual difficulties among physics undergraduates across a variety of institutions [47]. Singh [15] reviews many of these conceptual difficulties for upper-level undergraduate students. For example, students incorrectly thought particles lost energy during quantum tunnelling [48], had difficulties identifying the possibility of bound or scattering states for a given potential energy [49], had difficulties with the concept of time dependence in relation to expectation values [50], and had difficulties equating different representations in Bra-Ket notation [51]. More recent studies identified additional difficulties in a conceptual understanding of quantum entanglement [52], difficulties in understanding the Fermi-Dirac distribution and fermi energy for both undergraduate and graduate students [53], and difficulties representing a system of non-interacting identical particles [54]. These conceptual difficulties may be hidden behind good performance on algorithmic problems due to strong mathematical skills. However, authentic problem-solving would require both good foundational understanding of these QM concepts as well as well-developed problem-solving skills.

4.3 Theoretical Framework

Our framework for operationalizing real-world problem-solving builds upon previous work which defines problem-solving as a set of decisions that expert scientists make when approaching a typical problem in their work. In that study, researchers interviewed over fifty experts (here defined as an accomplished scientist, engineer, or physician, but not necessarily the most exceptional person in their field) in science, engineering, and medicine [13]. Adapting the critical decision method of cognitive task analysis [55], they asked the experts to recount how they solved a recent, routine problem in their work. This might be a difficult, but not exceptional

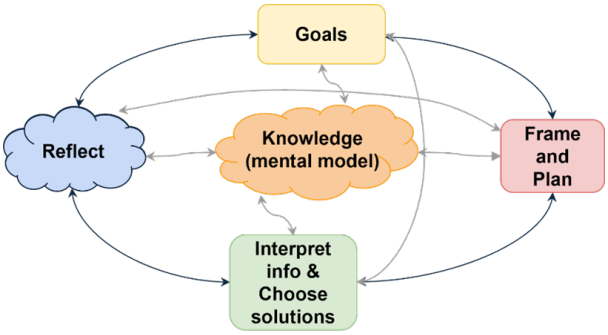


Figure 4.1: Illustration of the problem-solving process in [13]. Illustration provided by Argenta Price.

medical case, a challenging theoretical calculation, or the design of a new product or experimental apparatus. From these interviews they identified a set of 29 expert decisions, all of which were made by at least 80% of the experts interviewed in some combination [13]. For example, they might decide what the goals of the problem are (decision 3), what assumptions they can make (decision 10), what calculations are necessary (decision 16), whether information gathered from experiments is reliable (decision 18), and what the potential failure modes of the solution are (decision 23). This process is synthesized into a non-linear and iterative model, illustrated in Figure 4.1. Examples of the expert decisions and how they were made by an experimental condensed matter physicist trying to determine if a given material is a topological superconductor are given in Table 4.3.

With the key expert decisions of the problem-solving process identified, Price et al. suggest that measuring problem-solving will thus require presenting the solver with a realistic scenario, and then asking them questions that require them to make a subset of the 29 different expert decisions [12]. Pellegrino et al. and the National Research Council in Knowing What Students Know outline a framework for instructionally relevant assessments known as the assessment triangle [56]; this framework defines assessment as a process of making reasonable inferences about what students know from certain observations. This is represented by the three vertices of the triangle: (1) a model of student cognition in the relevant domain (2) a set of assumptions about the observations that will provide evidence of competence relative to that model; and (3) a process of interpreting that evidence.

Table 4.2: Selection of expert decisions from Price et al. [13] and an example of each expert decision from a discussion with a condensed matter physicist.

Expert Decision	Example from Physics
3) Goals, criteria, constraints?	We set out to design a new type of experiment that would tell us "topological or not topological?" but not anything about details.
4) Important features and info?	If it's a topological superconductor, these are the numbers that are theoretically allowed to change when you go through a phase transition. Other features relating to the microscopic mechanisms of superconductivity don't matter for this problem.
7) Related problems?	People have tried looking for this before, but by comparing six experiments, which made the systematics overwhelming - uncertainty gets blown up.
11) Decompose into sub-problems.	Approach individual technical problems separately, like how to transmit/measure sound through cables without also bringing heat from room temperature?
14) Priorities.	Need to prioritize approach to get some information quickly even though it won't be perfect, because building the perfect experiment from scratch will take too long and we may have missed something obvious that we'd learn about from the less perfect approach.
17) Represent and organize information.	Measuring the sound velocity as a function of temperature, he saw something that kind of looked like this.
18) How believable is information?	Does the data look right – does it match with what I expect this type of measurement to look like?
20) Any significant anomalies?	We identified (some quantifiable, others not) a bunch of different systematic effects in our experiment but were able to determine that they're not responsible for our result.
22) What is the best solution?	We had to redesign some experiments and eventually got some data that looked appropriate, and it turns out that the right combination of them had the right number of drops to tell us that this is a topological superconductor.
24) Additional knowledge needed?	We had to go back into the literature and find other examples of where people have solved similar problems using sound velocities... (also related problems)
26) How good is solution?	We checked something called Ehrenfest relations and they showed us that yes, our answer probably was correct.

Our framework synthesizes Pellegrino's assessment triangle with the model of cognition described by Price et al.'s framework of problem-solving as scientific decision-making, following the set of decisions described above [13]. In their second paper, Price et al. outline an assessment framework for observation and interpretation based on their model of cognition [12]. The assessment framework begins with a problem scenario that contains both relevant and irrelevant (as well as incomplete) information and requires some domain-specific content knowledge to solve. Importantly, the problem context does not eliminate any decision-making possibilities, for example, by specifying assumptions for the solver to make. To collect observations which align with this model of cognition, solvers are then asked guiding questions that target higher-order decisions like key features of the problem, what information is needed or what some potential solutions might be. It is essential that these beginning questions be general to allow the participants space to make these expert decisions without explicit prompting (e.g., write a plan for how you would solve this). The questions then get more targeted and may explicitly ask about information related to specific factors in case more novice solvers did not make that expert decision earlier in the assessment. Another key feature of this assessment is that it is not static – the solver is provided with more information (some of which they may not have identified) and then must interpret that information and reflect on their problem-solving process and solution. The assessments may iterate on these steps several times.

Price et al. present a novel way of interpreting the observations collected in these assessments [12]. The scoring, they argue, should be based on (1) whether a student makes an expert decision before being explicitly prompted through the built-in scaffolding and (2) whether the student applies the correct reasoning and relevant information when making those expert decisions. This is different from typical assessments which might only focus on the final answer the student arrives at, and implicitly assumes that the assessment responses must be interpreted in the context of current expertise in the field. This assessment structure guides the student through an approximation of a complex problem-solving process and there is information at each step about how a student made critical decisions. Another novel way of interpreting student responses is that Price et al. argue that one should score the students by comparing their problem-solving process to that of "experts" in the field. This could provide additional validity

evidence by showing separation between experts and different levels of students. This approach represents a dramatic shift from how typical assessments work. Rather than the primary objective being to rank students relative to one another, the focus becomes comparing students to an empirically derived standard of mastery. This provides more meaningful targets for curricular design – as we can specify a level of mastery expected of students at the end of a program (like standards-based grading [57]). These assessments are also inherently measuring multiple expert decisions, so they can highlight individual problem-solving skills that need more focus from students and instructors.

4.4 Assessment Design and Data Sources

In this work, we applied the framework of Price et al. to design an assessment of problem-solving for graduate QM [12]. The definition of authentic problems (see above) is essential for the assessment design: we cannot use textbook problems to probe authentic problem-solving because the solution path for textbook problems is often well-defined and known (particularly by experts [20,21]). Following Price et al., authentic problem-solving is defined as the reasoning processes undertaken by the solver as they make these expert decisions with limited information under the guidance of a sophisticated, discipline-specific mental model [13].

The context for the assessment is that students are working with a group of peers (fictionalized) and need to troubleshoot a variation on the Mach-Zehnder interferometer and determine the probability of detecting a photon at a certain location given an arrangement of beam-splitters and mirrors (see Figure 4.2). This is a device that many physicists would be familiar with, but students may not have encountered before. We introduced a unique variation to the design (a scattering beam splitter, inspired by the Quantum Bomb thought experiment [58]) to make the context more authentic for experts (i.e., the experts wouldn't know the solution in advance). According to Price et al., 2021, the ability to make decisions cannot be measured independently of concepts because problem-solving fundamentally requires use of a sophisticated, discipline-specific mental model. Thus, strong problem-solving is predicated on strong conceptual understanding. While the Mach-Zender interferometer is only one apparatus, it does include a large portion of concepts covered in QM 1. However, it remains an open

question as to how strongly correlated assessments like this one may be across different content areas.

An assessment template developed by Price et al. [12] and the QM assessment sequence are outlined in Figure 4.3. Price et al, developed the template to outline the structure used in their assessments which aimed emulate authentic problem-solving. This structure includes subsets of the decisions they identified in previous work [13], some of which occur in combination. For example, the "key features?" component is often accompanied by picking a predictive framework to use (decision 5), because the relevant predictive framework depends on key features. In the QM assessment, the solver was first presented with background information on a standard Mach-Zehnder interferometer (item one). This information detailed the basic working principles of the interferometer and the expected output for a given configuration. The background information was provided and remained available throughout the assessment to minimize the effect of familiarity with interferometers on the score. Next, the solver was presented with the design of the system and then asked how they would determine the expected probability of detecting light at a certain point in the system (item two). They were then asked what information they would need to solve the problem before being iteratively provided with more information about the system (item three). After several iterations of providing students with new information and asking them to interpret it, students were presented with a potential answer and then asked to determine whether the answer was reasonable (items seven and thirteen). We provided them with hypothetical measurements and asked them to diagnose what could be wrong with the system (items eighteen and nineteen). The use of Dirac notation and the calculation of expected values is commonly seen in undergraduate QM, reducing the impact of calculation ability on the assessment score. The complete list of questions asked as they correspond to Figure 4.3 may be found in the appendix.

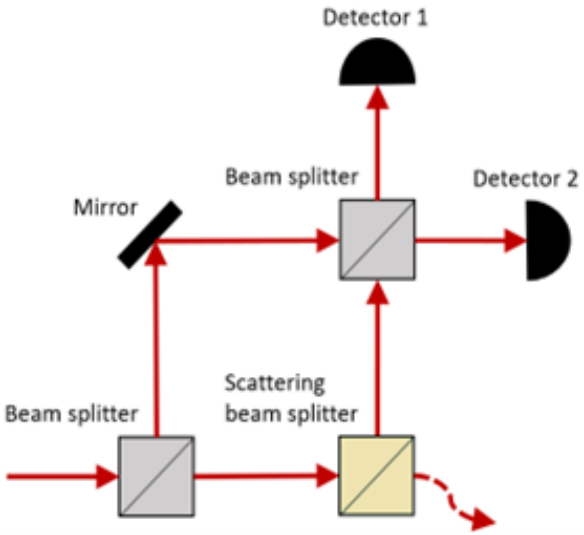
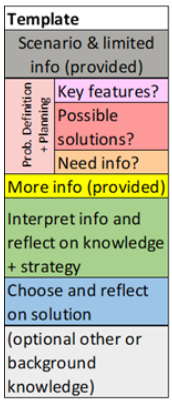


Figure 4.2: Illustration of the modified Mach-Zehnder interferometer used in the QM assessment.

The structure of providing new information allowed the assessment to probe decision-making at different stages of the problem-solving process while narrowing the solution space. Additionally, this new information provided a standardization to the assessment, which allowed solvers with unfavorable plans to complete the rest of the assessment. In places where the solution space was potentially very broad (e.g., what is the appropriate mathematical basis), questions were asked in a way that required the solver to evaluate the potential solutions of their hypothetical peers. This assessment probed nine of the 29 expert decisions identified by Price et al: [8]: developing a plan for solving the problem (decision 15), what potential solutions are (decision 8), whether the answer makes sense (decision 26), any significant anomalies (decision 20), what calculations are needed (decision 16), what constraints are on the solution (decision 3), what are important feature and information (decision 4), what predictive framework (decision 5), and how to decompose into subproblems (decision 11).



Item	Item Type	Prompt
1		A Mach-Zehnder interferometer... Figure...
2		A variation of the Mach-Zehnder... Figure...
3		Write a plan to theoretically determine the probability of detecting the photon
4		What results do you expect? Constraints?
5		What additional information... assumptions?
6		Determine basis. Asim thinks there are three states...
7		Do you think Asim/Nina's basis will allow you to determine the probabilities? Why? What changes would you make?
8		Which of the proposed bases would you pick?
9		Basis provided.
10		What should your group do next?
11		Determine operators. Which components in our design should be represented as operators?
12		Components specified.
13		Is provided operator acceptable? Why? What changes would you make?
14		Transformations provided.
15		Determine the probability of detection at each detector, normalized to the input.
16		Is your solution reasonable?
17		Theoretical and Experimental probabilities provided.
18		What are a few likely explanations for the difference between experimental and theoretical results?
19		What are likely explanations...: closed response selection and justify.
20		What are additional tests you would suggest running to determine why experimental and theoretical results differ?

Figure 4.3: General template for designing problem-solving assessments developed by Price et al. (left) and how that maps to the QM assessment (right). Each item reflects the information provided or a question asked. Items seven and thirteen contain two and three questions, respectively. The item type is color coded. Items with multiple color codes involved multiple expert decisions.

First, to ensure the questions were being interpreted in the desired way, four senior graduate students piloted the assessment in think-aloud interviews. These students were compensated for their time at a rate of $15 per hour. The interviewer (M.E.R.) asked the students to identify unclear questions or information. A lack of clarity prompted a discussion between the interviewer and student to find alternate wording. The questions were then refined to increase clarity. For example, the description of the scattering beam splitter was made more explicit. In the first two pilot tests there was an assumption that the provided experimental errors were not a possibility (i.e., no physicist would make this mistake). Therefore, the introduction of an experimental team in the assessment was changed to explicitly be identified as graduate students. After the change, this assumption was not present in the final two pilot tests. Additionally, the reasoning behind each answer was discussed either while the student was solving or after an answer was reached. This discussion confirmed each question was probing the desired skills.

Nine expert physicists from five universities and twelve first-year graduate students (none of whom were in the pilot data) from one university then completed the assessment following a think-aloud interview protocol. The students were compensated for their time at a rate of $15 per hour, but the experts were not compensated. The interview was open-note and open-book. While some interviewees brought these resources with them, none of the interviewees chose

to use these resources. We considered professional researchers in atomic or condensed matter physics to be experts, as they are most likely to use QM in their day-to-day activities. These experts were recruited through the authors' professional networks and through email advertisements to the American Physical Society mailing lists for the Division of Condensed Matter Physics and the Division of Atomic, Molecular, and Optical Physics. All graduate students participated in the interviews within one month of passing their graduate QM qualifying exam in Fall 2022, as this was roughly the target level of understanding the assessment was designed for (i.e., students who had completed a graduate QM course). Twelve of the invited fourteen graduate students from Fall 2022 participated. Participants in the think-aloud interview completed the assessment, submitting their typed answers, while discussing their thought process solving each problem. The interviewer (M.E.R.) asked for clarification of thoughts or answers as needed. Additional testing was conducted with undergraduate students who completed the assessment the month prior to their QM final exam in Fall 2022 and graduate students within one week of their QM final exam in Spring 2023. Students were invited to complete the assessment online via Qualtrics as an open book, open note, individual assignment. The assessment was not time restricted, and students could complete it across multiple days. Students could not alter previous answers once advancing pages, since additional information is provided throughout the assessment. Students were encouraged to contact M.E.R. with questions regarding the assessment. However, no questions were asked. Eight of the nineteen invited undergraduate students and ten of the seventeen invited graduate students from Spring 2023 participated and received bonus points in their QM course. In the results below, we refer to the experts and graduate students from the initial testing phase (think-aloud interviews) as "Expert" and "Fall 2022," respectively. The additional testing with undergraduate and graduate students through written responses only is referred to as "Undergrad," and "Spring 2023," respectively.

4.5 Rubric Development

M.E.R. used the expert data to create a scoring rubric for the assessment based on the Expert and Fall 2022 responses. This is quite different from typical problem-solving tasks in which the problem's author establishes the rubric a priori. In this case, inductive coding was used

to characterize the expert responses for each question. A list of codes was generated to cover the entire answer space of the Expert and Fall 2022 responses. A new code was added when a response could not be characterized by the existing codes. The existing codes were then merged to represent common or fundamental themes in the responses. Codes were merged if there was no distinction between them in responses. However, if some responses clearly included an idea which others did not, the code was not merged. For example, in item three, some responses described tracking the intensity of the beam along the paths, but other responses described tracking the intensity and the phase. There were multiple responses that made this distinction and multiple that did not. Therefore, the codes were not merged. The granularity of codes was flexible. Extremely coarse codes may lead to identical scores across responses, and extremely fine codes may lead to seemingly random scores independent of response quality. (We investigate the discriminant validity below.) The codes generated mainly relied on the typed responses, as the verbal component of the think-aloud interview did not significantly change the codes identified. What defines a code depends on the response. In item thirteen codes represented distinct issues in the operator or solutions to these issues which included mathematical responses. Whereas in item sixteen, codes represented broader ideas or methods used to check a solution.

Following Price et al. [12], we wanted to establish a scoring scheme relative to expert consensus. Price et al., argued that the problem-solving of experts reflects adaptive expertise, which is a measure of mastery which indicates the ability to solve novel problems with no known solution or solution path [12]. Research on novice problem-solvers suggests that novices more often pattern-match, which may be effective when solving problems with known and well-defined solutions but is not a sufficient or desirable strategy in more authentic contexts. Thus, in line with Price et al. [12] we identified "mastery" codes and "non-mastery" codes based on expert responses. Codes which the majority (five of nine) experts mentioned were deemed mastery codes, with the exception of two questions for which the experts diverged into two distinct approaches. This choice of a simple-majority threshold balances two competing ideas. One on extreme end, requiring unanimity among experts to define "mastery" codes suggests that there is only one correct way to approach the problem, which is at odds with

the very definition of an authentic problem. One the other extreme end, naming all statements given by experts as "mastery" codes (e.g., a threshold of one of nine) implies that the problem space is likely not well-defined enough to be used for reproducible and scalable assessment purposes. Responses to items eight, eighteen, and twenty did not have a clear consensus, suggesting the answer space is too broad to provide useful discrimination between experts and students. Therefore, the questions associated with these items are not scored. Item fifteen was a calculation question, unlike the expert-decision-probing questions, and was not scored as measuring accuracy of calculations is not the primary goal of this assessment. Item fifteen was included because item sixteen requires reflecting on this calculated answer.

During the scoring process, we conducted a sensitivity analysis which computed assessment scores using all possible mastery thresholds. The mastery threshold has a direct effect on the number of mastery codes, which consequently affects the score in two ways. (1) The number of mastery codes was used in the score calculations, discussed in the next section. (2) Questions were only scored if it had at least one mastery code. Therefore, a lower mastery threshold results in more questions being included, while a higher mastery threshold would exclude more questions. For example, a mastery threshold at eight of nine experts leads to only five questions containing at least one mastery code; therefore, only five questions could be used in scoring. While all consensus thresholds (one to nine) provided a clear delineation between experts and students, we found that five of nine provided a balance which included some aspects of the problem-solving process which were well-defined enough to achieve expert unanimity, as well as other aspects where experts may have given a range of reasonable answers due to the limited information available at that point in the assessment. For example, we found that two questions had distinct groups of expert responses because there were multiple reasonable ways of answering the questions. The details of how these were handled are in the Scoring section below.

4.6 Assessment Scoring

The assessment was scored by awarding points for each expert code and deducting points for each non-expert code. Each expert code was awarded one point while each non-expert code

resulted in a half-point deduction. This non-expert penalty was selected to maximize the delineation of experts and students. While student answers that do not align with expert response may not be wrong, these answers potentially neglect the more holistic or efficient approach of experts. For example, students may name a lot of different features which are all relevant, but not properly prioritize [59], which would artificially inflate their scores without reflecting their ability to prioritize (Decision 14 in Price et al., [13]). Each question was thus scored using the following equation

$$questionscore = \frac{C_M - 0.5 * C_N}{C_M}$$

where CM is the number of mastery codes and CN is the number of non-mastery codes, which resembles a scoring equation used in a heat transfer problem-solving assessment [22]. Two questions were scored differently to account for multiple unique solution paths of the experts. For these questions, each mastery code aligned with only one path, while non-mastery codes aligned with no path. Responses were given a score, using the scoring equation, for each solution path. The highest score was used as the question score for the response. This effectively amounts to a "branching" of the solution space at this point in the assessment which could result in multiple ways to exhibit mastery.

We note that items fifteen, eighteen, and twenty (see Figure 4.3), were not included in the assessment score. Item fifteen required the solver to conduct a procedural calculation, which is not the focus of this assessment. However, this calculation was required for item sixteen, which required them to reflect on their solution (Decision 26 [13]). Items eighteen and twenty did not have any codes that met our threshold for mastery codes, suggesting the question was too open or vaguely worded to provide useful information. Therefore, the responses are excluded from the assessment score. All experts and students in this sample still answered these questions to account for their potential influence on later answers, but in the future these questions will be removed.

We provide some specific examples of assessment coding and scoring below. We report the median scores for individual questions. The median and average for individual questions was occasionally quite different, due to the granularity of scoring with codes. However, the median and average for the overall assessment score similar. Item three in the assessment

asked for the solver to outline a plan for determining the probability of detecting a photon at each detector. Five total codes were identified in Expert and Fall 2022 responses (Table 4.6). Three codes, track path intensity, track path phase, and interference, were identified as mastery codes. The median score on this question was 50% for experts and 33% for Fall 2022. If we had used a low threshold for mastery (e.g., one or two of nine experts), the median expert scores would have been 60%, and the Fall 2022 median would have been 20%. Conversely, if we had used a high mastery threshold (seven or more of nine), we would not have been able to get any information out of this question.

Table 4.3: Codes identified in question 1: "Write a plan (bullet points or steps) to determine the probability of detector 1 and detector 2 each detecting the photon, normalized to the input." The code and a description are included. The Experts column shows the fraction of experts who received this code. Mastery codes are marked with an asterisk.

Code Label	Details	Experts
Modify Background	Compare the Modified interferometer to the Mach-Zehnder interferometer provided in the background.	2/9
Track path intensity	Track the wave intensity along the path.	5/9*
Track path phase	Track the wave phase along the path.	5/9*
Validate method with background	Test method with standard Mach-Zehnder interferometer	2/9
Interference	Account for the interference in some form	6/9*

Item thirteen provided potential transformations for a mirror operator for the solver to choose from. The behavior of the mirror was provided in the background, including details on a phase shift. A basis was provided at this point in the assessment. The solver was asked to determine if the proposed transformations were reasonable and provide modifications if they were not. One code, the irrelevance of the transformation on state three, was identified as a mastery code (Table 4.6). Concern or disapproval of the imaginary term in the transformation was only found in student response and in zero expert responses. The median score on this question was 100% for experts and 0% for Fall 2022. This question would have been included

for almost any choice of threshold, but if the threshold was lowered to one of nine, the expert median would have been 33% and the Fall 2022 median would have been 33%.

Table 4.4: Codes identified in question 9: "Do you think Nina's mirror operator is acceptable for the agreed basis? Why? If not, what changes would you make?" The code and a description are included. The Experts column shows the fraction of experts who received this code. Mastery codes are denoted with an asterisk. † Code mentioned by students only.

Code Label	Details	Experts
Yes	All elements are correct and necessary.	1/9
$\hat{A}\|3\rangle$ is irrelevant	This transformation is irrelevant to the problem. (Does not require an assessment of the transformation accuracy.)	8/9*
$\hat{A}\|1\rangle$ is irrelevant	This transformation is irrelevant to the problem. (Does not require an assessment of the transformation accuracy.)	1/9
Imaginary term issue	The imaginary term is unnecessary or incorrect.	0/9†

One question in item seven provided a potential basis to describe the modified Mach-Zehnder interferometer system. This basis assigned two states which related to the photon's direction of travel. Codes were identified to characterize the responses of experts and Fall 2022 (Table 4.6). There were three solution paths discussed by experts: (1) Three experts stated the basis was reasonable because the scattering which was neglected from the basis could be mathematically handled, coded as math scattering. (2) Three experts considered the basis unreasonable because the basis did not account for scattering, coded as scattering. Two of these experts stated the basis needed an additional state to represent the scattering state, coded as scattering state. (3) The remaining three experts believed the basis was unreasonable because the basis did not include information about the phase, phase information. Two of these experts proposed using states which include the phase information, coded as phase basis. To represent these solution paths, the subset of codes associated with a solution path were identified as mastery codes. Solution paths two and three include one code for identifying an issue and a second code for a solution. The median score on this question was 100% for experts and 50% for Fall 2022.

Table 4.5: Codes identified in question 5: "Do you think Nina's basis will allow you to determine the probability of detector 1 and detector 2 each detecting the photon? Why? If not, what changes would you make?" The code and a description are included. The Experts column shows the fraction of experts who received this code. [1] Mastery codes for solution path one. [2] Mastery codes for solution path two. [3] Mastery codes for solution path three. [†] Code mentioned by students only.

Code Label	Details	Experts
Math Scattering	The scattering can be mathematically handled with this basis.	3/9[1]
Complete	This basis is complete.	0/9[†]
Scattering	Identifying the basis does not account for scattering	3/9[2]
Uniqueness	Each state is not unique. It represents more than one state.	1/9
Phase information	Phase information is not included in the basis.	3/9[3]
Scattering state	An additional state should be added to account for scattering	2/9[2]
Phase basis	The states should be altered to include phase information	2/9[3]

Once all question scores were determined the total score was then calculated as a weighted average across all questions. The weighting was proportional to the average expert score, which awarded more points to questions in which experts had a high consensus and lower points for a question with a low consensus, using the equation

$$totalscore = \sum_i \frac{E_i}{E_{total}} Q_i$$

Where i is the question index, E_i is the average expert question score for question i, E_{total} is the sum of the average expert question scores across all questions, and Qi is the question score for the response.

An inter-rater reliability check was conducted with G.J.D as the second rater. The second rater was a graduate student with experience in qualitative coding, but no involvement in the development of the assessment or codes. The second rater conducted two rounds of scoring. In the first round, 7 responses were scored across the groups: two from Expert, two from FA2022,

two from SP2023, and one from Undergrad. A response was a complete set of answers from one solver. The number of responses per group was determined, and the responses were randomly selected from the group. Any differences were discussed, and an agreement was reached. In the second round, one response from each group for a total of 4 was scored. An inter-rater reliability check using Cohen's kappa suggests a "substantial agreement" using the Landis & Kock guidelines [60], with a kappa value of 0.72. The two raters had an 88% agreement on the additional 4 responses.

4.7 Validity and Reliability Evidence

Below, we provide multiple sources of validity and reliability evidence [56]. These include (1) *discriminant validity* or *face validity* in which we show clear delineation between different groups of students which you would expect based on their level of education, (2) *inferential validity* in which we show that the assessment scores are correlated with conceptual understanding of QM, and (3), *test-retest reliability* in which we show that the assessment scores were nearly identical for multiple similar groups of students.

4.7.1 Discriminant Validity

We created box plots of the assessments scores from each of the testing groups, "Expert" (the nine experts recruited in the rubric development phase), "FA2022" (the graduate students recruited in the rubric development phase), "Spring 2023," (the graduate students who took the assessment after the rubric development), and "Undergrad," (the undergraduate students who took the assessment after the rubric development). Both Expert and FA2022 completed the assessment during the think-aloud interviews. SP2023 and Undergrad completed the assessment as online assessments at home. The scaled aggregate scores are plotted in Figure 4.4. The box encompasses the 25th and the 75th percentiles of each score, with the group median indicated by the solid line. Experts had a median score of M = 80% with an interquartile range of 11%, the graduate students had a median score of M = 37%, interquartile range of 15% (2022) and M = 38%, interquartile range of 27% (2023), and the undergraduate students had a median score of M = 29%, interquartile range of 8%. This provides evidence of discriminant

validity as the experts had much higher scores than the other groups, and the graduate students had higher scores than the undergraduate students: the participants with more experience and education had higher scores. Initial items were less defined and had a wider range of response. As the items became more defined, the responses had a higher consensus. However, expert responses began to converge at earlier item numbers. Additionally, the variation in expert and undergraduate scores was smaller than the graduate students, which Price et al. [12] argue is another metric of validity. That is, the experts provide more unified responses as they are all guided by similarly sophisticated mental models. The graduate students have a wider variation in responses due to wider variations in the students' experiences in solving real problems (e.g., some have begun research, whereas some have not). The undergraduate students did not have a large variation of scores, likely due to a floor effect because the assessment was designed for graduate students.

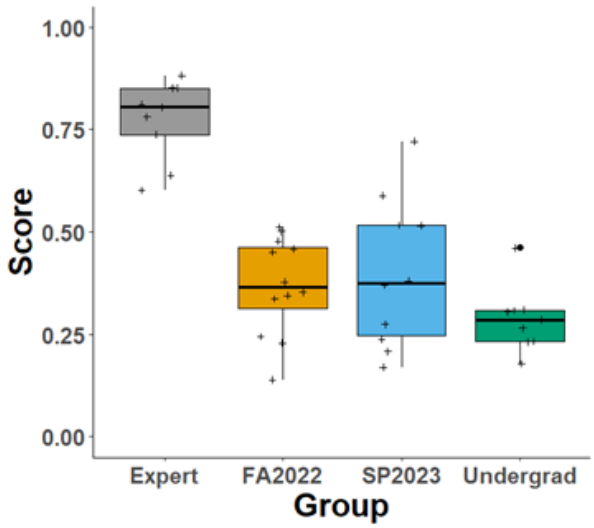


Figure 4.4: Normalized scores (out of 1) on the problem-solving assessment in QM. Gray box is for expert volunteers, Orange is for graduate students who completed the qualifying exam, Blue is for graduate students in QM 1, and green is for undergraduate students.

4.7.2 Inferential Validity

The Spring 2023 graduate student cohort completed a QM Survey [49] during the first week of the QM 1 course as they were serving as a quasi-control for an ongoing intervention study. The QM Survey assesses conceptual understanding of upper-level undergraduate or graduate

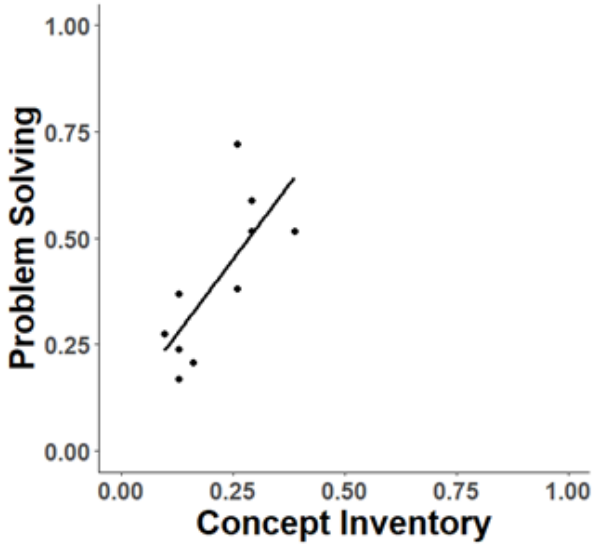


Figure 4.5: Assessment versus concept inventory scores. The Concept Inventory score and problem-solving assessment scores for the spring 2023 cohort are plotted.

students, specifically the formalism of one-dimensional quantum mechanics. The QM problem-solving assessment was completed in the last week of the course. A paired Spearman test was conducted, indicating a strong correlation between the survey and assessment (rho = 0.70, p = 0.02; see Figure 4.5). We would expect this correlation to be strong as Price et al. [12] argue that problem-solving requires domain-specific conceptual knowledge.

4.7.3 Reliability

A comparison between two cohorts of graduate students was used to investigate repeatability. A test-retest reliability of an assessment is typically conducted as multiple measurements on the same group in similar conditions. However, similar conditions could not be recreated for one group. For repeat measurements taken substantially after the initial, students would have received additional practice and instruction, making their test conditions different. For repeat measurements taken shortly after the initial, experts and students are likely to respond differently because completing the assessment reveals information that should only be known to the solver at appropriate stages in the assessment. Therefore, the two cohorts were used as a measure for similar conditions. In terms of preparation, the Fall 2022 cohort had one additional summer of independent study for the qualifiers compared to the Spring 2023 cohort. Both groups had a comparable level of academic preparation. A Mann-Whitney U test was

conducted on the problem-solving assessment scores and suggested the groups did not have statistically different scores ($p = 0.63$).

4.8 Discussion and Implications for Future Work

We discussed the design of the QM problem-solving assessment which focused on measuring the ability to make the expert decisions expected of first-year graduate students in the context of QM. We collected preliminary evidence for discriminant validity, inferential validity, and test-retest reliability of the assessment which suggests that the assessment has the potential for wider use.

Currently, measures of graduate student competencies in QM consist primarily of course grades, which are largely based on textbook-style problems with limited opportunities for decision-making [8, 61]. Leak et al., argue that the graduate classroom needs to move beyond routine problems which have known solutions and strategies toward more authentic problems [61]. Indeed, one of the goals of The International Year of Quantum Science and Technology is training and inspiring "the next generation of quantum pioneers." This assessment offers a potential measurement of the skills which physics graduates will need to "drive quantum technological innovation" that will address the urgent climate crisis, improve health and well-being, and drive economic growth [62]. Given the reports of student difficulties with QM concepts and applications of knowledge, this assessment will provide a metric to be used in conjunction with other research-based instruments to triangulate students' abilities to understand and apply QM to grand societal challenges, and thus inform targets and potentially new methods for instructional reform of graduate quantum education.

The online nature of the assessment allows for more versatile use of this assessment. The lack of strict time and resource restrictions may make this assessment comparable to an authentic problem-solving experience of experts. This assessment is not intended to be used as a pre and posttest, because the information is provided throughout the assessment may alter future responses. While both raters had qualitative coding experience, the interrater reliability check was conducted with someone unfamiliar with these assessment structures and the underlying

theoretical framework. Thus, we expect that a reasonably experienced teaching assistant or instructor would be able to follow the rubric.

While the preliminary evidence supporting this assessment is promising, there are many limitations which will require further research and development. First, the sample size of experts was small, and therefore their responses may not represent the full solution space of experts. These experts did span multiple universities and hold a variety of professional experience and sub-field expertise, but mastery codes could still shift if a larger sample of experts be included. Second, the sample of students was drawn from one university. We have no reason to believe other institutions using common lecture-focused teaching would score much differently, given previous reports about student misconceptions in QM [47]. The QM survey and the decisions assessment were taken at the beginning and end of the semester, respectively. The inferential validity could be made stronger if these measurements were taken at the same time.

Following Walsh et al. [63], the next phase of assessment development would require collecting data from more experts and students nationally and internationally to develop a more robust rubric and provide further validity evidence. Similarly, for this assessment to achieve wider use and undergo further tests of inferential validity, it would ideally be transformed into a closed response format for automatic and large-scale grading. This process will take considerable time and resources but will be an important investment in graduate quantum education, and an important proof-of-concept that a typically nebulously defined skill like problem-solving could be measured at a large scale.

Finally, with respect to the focus of this issue on quantum education, we are planning to develop parallel assessments in other core graduate subject areas (Statistical Mechanics and Electromagnetism), as it remains an open question whether student difficulties solving authentic problems are unique to, or more pronounced in, QM compared with other areas of physics. This is an important question not just in quantum education, but in education research more broadly.

4.9 Acknowledgments

This work was partially supported by the National Science Foundation, Grant No. DGE-2429155. We would like to thank Nathan D. Davis and J. Isaac Garcia for their insights during the development of the assessment, as well as the students and professionals for their participation.

References

[1] ABET. *Criteria for accrediting engineering programs 2019-2020*. 2019.

[2] Clive L. Dym et al. “Engineering design thinking, teaching, and learning”. In: *Journal of Engineering Education*. Vol. 94. Wiley-Blackwell Publishing Ltd, 2005, pp. 103–120. DOI: 10.1002/j.2168-9830.2005.tb00832.x.

[3] Clive L Dym. “Design, systems, and engineering education”. In: *International Journal of Engineering Education* 20.3 (2004), pp. 305–312.

[4] David Jonassen, Johannes Strobel, and Chwee Beng Lee. “Everyday problem solving in engineering: Lessons for engineering educators”. In: *Journal of Engineering Education* 95 (2 2006), pp. 139–151. ISSN: 10694730. DOI: 10.1002/j.2168-9830.2006.tb00885.x.

[5] C. D. Porter and A. F. Heckler. “Effectiveness of guided group work in graduate level quantum mechanics”. In: *Physical Review Physics Education Research* 16 (2 2020). ISSN: 24699896. DOI: 10.1103/PhysRevPhysEducRes.16.020127.

[6] Jordan Harshman. “Review of the Challenges that Face Doctoral Education in Chemistry”. In: *Journal of Chemical Education* 98 (2 2021). ISSN: 19381328. DOI: 10.1021/acs.jchemed.0c00530.

[7] Brittany D. Busby and Jordan Harshman. “Program elements’ impact on chemistry doctoral students’ professional development: A longitudinal study”. In: *Chemistry Education Research and Practice* 22 (2 2021). ISSN: 11094028. DOI: 10.1039/d0rp00200c.

[8] Michael E Robbins, Nathan D Davis, and Eric W Burkholder. “Decision-making in graduate physics coursework: what is being assessed versus what is expected”. In: *in review* ().

[9] Frank Fischer et al. *Scientific reasoning and argumentation: The roles of domain-specific and domain-general knowledge*. 2018. DOI: 10.4324/9780203731826.

[10] Per Kind and Jonathan Osborne. “Styles of Scientific Reasoning: A Cultural Rationale for Science Education?” In: *Science Education* 101 (1 Jan. 2017), pp. 8–31. ISSN: 1098237X. DOI: 10.1002/sce.21251.

[11] Christopher J. Harris et al. “Designing Knowledge-In-Use Assessments to Promote Deeper Learning”. In: *Educational Measurement: Issues and Practice* 38 (2 2019). ISSN: 17453992. DOI: 10.1111/emip.12253.

[12] Argenta Price et al. “An accurate and practical method for assessing science and engineering problem-solving expertise”. In: *International Journal of Science Education* 44 (13 2022). ISSN: 14645289. DOI: 10.1080/09500693.2022.2111668.

[13] Argenta M. Price et al. “A detailed characterization of the expert problem-solving process in science and engineering: Guidance for teaching and assessment”. In: *CBE Life Sciences Education* 20 (3 2021). ISSN: 19317913. DOI: 10.1187/cbe.20-12-0276.

[14] Basil Bernstein. “On the classification and framing of educational knowledge”. In: *On the classification and framing of educational knowledge*. Vol. 3. Routledge and Kegan Paul, 1975, pp. 85–115.

[15] Chandralekha Singh and Emily Marshman. “Review of student difficulties in upper-level quantum mechanics”. In: *Physical Review Special Topics - Physics Education Research* 11 (2 2015). ISSN: 15549178. DOI: 10.1103/PhysRevSTPER.11.020117.

[16] Kenneth Heller and Patricia Heller. *Cooperative Problem Solving in Physics A User’s Manual Why? What? How? Recognize the Problem What’s going on? STEP 1*. Tech. rep. 2010.

[17] Wendy K. Adams and Carl E. Wieman. “Analyzing the many skills involved in solving complex physics problems”. In: *American Journal of Physics* 83 (5 May 2015), pp. 459–467. ISSN: 0002-9505. DOI: 10.1119/1.4913923.

[18] Shima Salehi. *IMPROVING PROBLEM-SOLVING THROUGH REFLECTION A DISSERTATION SUBMITTED TO THE GRADUATE SCHOOL OF EDUCATION AND THE COMMITTEE ON GRADUATE STUDIES OF STANFORD UNIVERSITY IN PARTIAL*

FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY. Tech. rep. 2018. URL: http://purl.stanford.edu/gc847wj5876.

[19] Michelene T.H. Chi, Paul J. Feltovich, and Robert Glaser. "Categorization and representation of physics problems by experts and novices". In: *Cognitive Science* 5 (2 1981). ISSN: 03640213. DOI: 10.1207/s15516709cog0502_2.

[20] Chandralekha Singh. "When physical intuition fails". In: *American Journal of Physics* 70 (11 Nov. 2002), pp. 1103–1109. ISSN: 0002-9505. DOI: 10.1119/1.1512659.

[21] Sam Wineburg. "Reading Abraham Lincoln: An expert/expert study in the interpretation of historical texts". In: *Cognitive Science* 22 (3 1998). ISSN: 03640213. DOI: 10.1207/s15516709cog2203_3.

[22] Jiamin Zhang et al. "Assessing authentic problem-solving in heat transfer". In: *ASEE Annual Conference and Exposition, Conference Proceedings*. American Society for Engineering Education, Aug. 2022. DOI: 10.18260/1-2--40752.

[23] Organisation for Economic Co-operation and Development. *PISA 2012 assessment and analytical framework : mathematics, reading, science, problem solving and financial literacy.* OECD, 2013, p. 261. ISBN: 9789264190528.

[24] Alan H Schoenfeld. *TOWARD A THEORY OF TEACHING-IN-CONTEXT*. Tech. rep. 1998. URL: http://www.gse.berkeley.edu/Faculty/.

[25] Patrick B. Kohl and Noah D. Finkelstein. "Patterns of multipe representation use by experts and novices during physics problem solving". In: *Physical Review Special Topics - Physics Education Research* 4 (1 2008). ISSN: 15549178. DOI: 10.1103/PhysRevSTPER.4.010111.

[26] Pamela Thibodeau Hardiman, Robert Dufresne, and Jose P. Mestre. "The relation between problem categorization and problem solving among experts and novices". In: *Memory & Cognition* 17 (5 1989). ISSN: 0090502X. DOI: 10.3758/BF03197085.

[27] Ton De Jong and Monica G M Ferguson-Hessler. *Cognitive Structures of Good and Poor Novice Problem Solvers in Physics*. Tech. rep. 1986, pp. 27–288.

[28] Jill Larkin et al. “Expert and novice performance in solving physics problems”. In: *Science* 208 (4450 1980). ISSN: 00368075. DOI: 10.1126/science.208.4450.1335.

[29] Eric Mazur. “Qualitative versus quantitative thinking: are we teaching the right thing?” In: *Optics and Photonics News* 3 (2 1992), p. 38.

[30] William J. Leonard, Robert J. Dufresne, and Jose P. Mestre. “Using qualitative problem-solving strategies to highlight the role of conceptual knowledge in solving problems”. In: *American Journal of Physics* 64 (12 1996). ISSN: 0002-9505. DOI: 10.1119/1.18409.

[31] Eric Mazur. “Peer instruction: Getting students to think in class”. In: 2008. DOI: 10.1063/1.53199.

[32] Eunsook Kim and Sung-Jae Pak. “Students do not overcome conceptual difficulties after solving 1000 traditional problems”. In: *American Journal of Physics* 70 (7 2002). ISSN: 0002-9505. DOI: 10.1119/1.1484151.

[33] Edward F. Redish. “Changing student ways of knowing: What should our students learn in a physics class?” In: *Proceedings of World View on Physics Education 2005: Focusing on Change* (August 2005).

[34] F. Reif and Joan I. Heller. “Knowledge Structure and Problem Solving in Physics”. In: *Educational Psychologist* 17 (2 1982). ISSN: 15326985. DOI: 10.1080/00461528209529248.

[35] Jill H. Larkin and Herbert A. Simon. “Why a Diagram is (Sometimes) Worth Ten Thousand Words”. In: *Cognitive Science* 11 (1 1987). ISSN: 03640213. DOI: 10.1016/S0364-0213(87)80026-5.

[36] Jill H. Larkin, Joan I. Heller, and James G. Greeno. “Instructional implications of research on problem solving”. In: *New Directions for Teaching and Learning* 1980 (2 1980). ISSN: 15360768. DOI: 10.1002/tl.37219800206.

[37] Alan Van Heuvelen. “Overview, Case Study Physics”. In: *American Journal of Physics* 59 (10 1991). ISSN: 0002-9505. DOI: 10.1119/1.16668.

[38] Mel S. Sabella and Edward F. Redish. “Knowledge organization and activation in physics problem solving”. In: *American Journal of Physics* 75 (11 2007). ISSN: 0002-9505. DOI: 10.1119/1.2746359.

[39] Jennifer Docktor and Kenneth Heller. “Assessment of student problem solving processes”. In: *AIP Conference Proceedings*. Vol. 1179. 2009. DOI: 10.1063/1.3266696.

[40] Melanie M Cooper et al. *An Assessment of the Effect of Collaborative Groups on Students’ Problem-Solving Strategies and Abilities Chemical Education Research edited by*. Tech. rep. 2008, p. 23. URL: www.JCE.DivCHED.org.

[41] K. Anders Ericsson and Herbert A. Simon. “Verbal reports as data”. In: *Psychological Review* 87 (3 1980). ISSN: 0033295X. DOI: 10.1037/0033-295X.87.3.215.

[42] Marlina Ali et al. “The Importance of Monitoring Skills in Physics Problem Solving.” In: *European Journal of Education Studies* 1 (3 2016). ISSN: 2501-1111.

[43] Wendy Kristine Adams. “Development of a Problem Solving Evaluation Instrument; untangling of specific problem solving assets”. PhD thesis. 2007.

[44] A. Halim et al. “An analysis of students’ skill in applying the problem solving strategy to the physics problem settlement in facing AEC as global competition”. In: *Jurnal Pendidikan IPA Indonesia* 5 (1 2016). ISSN: 20894392. DOI: 10.15294/jpii.v5i1.5782.

[45] Eric Burkholder, Lena Blackmon, and Carl Wieman. “Characterizing the mathematical problem-solving strategies of transitioning novice physics students”. In: *Physical Review Physics Education Research* 16 (2 Nov. 2020). ISSN: 24699896. DOI: 10.1103/PhysRevPhysEducRes.16.020134.

[46] Patricia Heller, Ronald Keith, and Scott Anderson. “Teaching problem solving through cooperative grouping. Part 1: Group versus individual problem solving”. In: *American Journal of Physics* 60 (7 1992). ISSN: 0002-9505. DOI: 10.1119/1.17117.

[47] Chandralekha Singh. “Student understanding of quantum mechanics at the beginning of graduate instruction”. In: *American Journal of Physics* 76 (3 Mar. 2008), pp. 277–287. ISSN: 0002-9505. DOI: 10.1119/1.2825387. URL: https://pubs.aip.org/ajp/article/76/3/277/1056828/Student-understanding-of-quantum-mechanics-at-the.

[48] Michael C. Wittmann, Jeffrey T. Morgan, and Lei Bao. “Addressing student models of energy loss in quantum tunnelling”. In: *European Journal of Physics* 26 (6 Nov. 2005), pp. 939–950. ISSN: 01430807. DOI: 10.1088/0143-0807/26/6/001.

[49] Guangtian Zhu and Chandralekha Singh. “Improving students’ understanding of quantum measurement. I. Investigation of difficulties”. In: *Physical Review Special Topics - Physics Education Research* 8 (1 2012). ISSN: 15549178. DOI: 10.1103/PhysRevSTPER.8.010117.

[50] Emily Marshman and Chandralekha Singh. “Investigating Student Difficulties with Time dependence of Expectation Values in Quantum Mechanics”. In: American Association of Physics Teachers (AAPT), July 2014, pp. 245–248. DOI: 10.1119/perc.2013.pr.049.

[51] Chandralekha Singh. “Student understanding of quantum mechanics”. In: *American Journal of Physics* 69 (8 Aug. 2001), pp. 885–895. ISSN: 0002-9505. DOI: 10.1119/1.1365404.

[52] Michael Brang et al. “Spooky action at a distance? A two-phase study into learners’ views of quantum entanglement”. In: *EPJ Quantum Technology* 11 (1 Dec. 2024). ISSN: 21960763. DOI: 10.1140/epjqt/s40507-024-00244-y.

[53] Paul Justice, Emily Marshman, and Chandralekha Singh. “Student understanding of Fermi energy, the Fermi-Dirac distribution and total electronic energy of a free electron gas”. In: *European Journal of Physics* 41 (1 2020). ISSN: 13616404. DOI: 10.1088/1361-6404/ab537c.

[54] Emily Marshman, Christof Keebaugh, and Chandralekha Singh. “Student difficulties with the basics for a system of non-interacting identical particles”. In: *Physics Education*

Research Conference Proceedings. American Association of Physics Teachers, 2021, pp. 257–263. ISBN: 9780917853487. DOI: 10.1119/perc.2021.pr.Marshman.

[55] Kathleen Mosier et al. "Expert Professional Judgments and "Naturalistic Decision Making"". In: *The Cambridge Handbook of Expertise and Expert Performance: Second Edition*. 2018. DOI: 10.1017/9781316480748.025.

[56] James W. Pellegrino, Louis V. DiBello, and Susan R. Goldman. "A Framework for Conceptualizing and Evaluating the Validity of Instructionally Relevant Assessments". In: *Educational Psychologist* 51 (1 2016). ISSN: 00461520. DOI: 10.1080/00461520.2016.1145550.

[57] Ian D Beatty. "Standards-based grading in introductory university physics". In: *Journal of the Scholarship of Teaching and Learning* 13 (2 2013). ISSN: 1527-9316.

[58] Avshalom C. Elitzur and Lev Vaidman. "Quantum mechanical interaction-free measurements". In: *Foundations of Physics* 23 (7 1993). ISSN: 00159018. DOI: 10.1007/BF00736012.

[59] Eric Burkholder, Lisa Hwang, and Carl Wieman. "SUPPORTING AUTHENTIC PROBLEM-SOLVING THROUGH A CORNERSTONE DESIGN COURSE IN CHEMICAL ENGINEERING". In: *Chemical Engineering Education* 55 (3 2021). ISSN: 21656428. DOI: 10.18260/2-1-370.660-126222.

[60] J. Richard Landis and Gary G. Koch. "The Measurement of Observer Agreement for Categorical Data". In: *Biometrics* 33 (1 1977). ISSN: 0006341X. DOI: 10.2307/2529310.

[61] Anne E. Leak et al. "Examining problem solving in physics-intensive Ph.D. research". In: *Physical Review Physics Education Research* 13 (2 2017). ISSN: 24699896. DOI: 10.1103/PhysRevPhysEducRes.13.020101.

[62] *International Year of Quantum Science and Technology*. 2024. URL: https://quantum2025.org/en/.

[63] Cole Walsh et al. “Quantifying critical thinking: Development and validation of the physics lab inventory of critical thinking”. In: *Physical Review Physics Education Research* 15 (1 May 2019). DOI: 10.1103/physrevphyseducres.15.010135.

Chapter 5

Conclusion

In these chapters, methods for characterizing, assessing, and implementing decision-making were discussed. (Chapter 2) Labs for non-science majors were written to develop decision-making with scaffolding, during which student attitudes shifted towards an expert-like mindset and student interest increased. (Chapter 3) Decisions practiced in graduate coursework and instructor expectations of decision making were investigated to reveal a mismatch. (Chapter 4) A quantum mechanics assessment was developed and validated to measure decision-making in decision-making items which many instructors expected of first-year students.

The chapters together build on the existing literature of problem solving and decision-making in courses. The curricular overhaul of the introductory labs produced clear benefits, strengthening the understanding of effective labs. The characterization of decision-making in graduate work raises awareness of the current state of graduate instruction so that future works may develop interventions or large-scale curricular changes to make graduate coursework for effective in preparing students for research. The development of a decision-making assessment is essential in measuring the effect of decision-making interventions. This assessment will allow us to rigorously evaluate future instructional interventions and will be used as a template to develop parallel assessments for the remainder of the topics in the first-year graduate curriculum.

Continuing this work, an in-progress intervention at Auburn aims to develop problem-solving skills in a graduate quantum mechanics course. The intervention is a collection of practice assignments targeting specific problem-solving skills, and the impact is assessed using the assessment developed in Chapter 4. The specific skills included in the intervention were

identified in Chapter 3 as expected but rarely practiced in coursework. This supplemental intervention (intervention not in the primary lecture course) seeks to develop these skills while requiring little to no course modification to incorporate. How well similar supplemental interventions may work in graduate courses remains to be seen. While supplemental interventions can be effective, to more effectively develop decision-making skills a more integrated approach may be productive. Decision-making is prevalent in real-world problems, and these studies furthered the literature aiming to more effectively develop these skills.

Appendices

Chapter 6

Appendix A: Decision-Making Framework

1. What are important questions or problems? Where is the field heading? Are there advances in the field that open new possibilities?
2. If and where are the gaps/opportunities to solve in field? Given their unique perspectives and capabilities, are there opportunities particularly accessible to them?
3. What are your goals for this problem? Considerations include:
   (a) What are the goals, design criteria, or requirements of the problem or its solution?
   (b) What is the scope of the problem?
   (c) What constraints are there on the solution?
   (d) What will be the criteria on which the solution is evaluated?
4. What are the important underlying features or concepts that apply? Could include:
   (a) Which available information is relevant to solving and why?
   (b) (When appropriate) Create/find a suitable abstract representation of core ideas and information (i.e., physics - equation, chemistry - bond diagrams/potential energy surfaces, biology - diagram of pathway steps)
5. Which potential predictive frameworks to use? (Decide among possible predictive frameworks, or create frameworks)
   (a) As the problem-solving process progresses, how well do predictive frameworks apply in specific problem context?
   (b) Predictive framework is defined as a mental model of key features of the problem and the relationships between the features
6. How to narrow down the problem? Often involves formulating specific questions and hypotheses. (Taking general framework and putting it into more specific predictions.)
7. What are related problems or work seen before, and what aspects of their solving process and solutions might be useful in the present context?
8. What are potential solutions?
9. If the problem is solvable and its solution worth pursuing given the difficulties, constraints, risks, and uncertainties?

10. What approximations or simplifications are appropriate? Test them against established criteria. How can I simplify the problem to make it easier for me to solve?

11. How to decompose the problem into more tractable sub-problems? Decide on pieces which involve their own independent problem solving and sub-goals.

12. Which are areas of particular difficulty and/or uncertainty in the solving process? Could also be deciding:

    (a) What are acceptable levels of uncertainty with which to proceed at various stages?

13. What information is needed to solve the problem? Could include:

    (a) What will be sufficient to test and distinguish between potential solutions?

14. hat to prioritize among many competing considerations? What to do first and how to obtain necessary resources? Considerations could include: What's most important? Most difficult? Addressing uncertainties? Easiest? Constraints (time, materials, etc.)? Cost? Optimization and trade-offs? Availability of resources?

15. What is the specific plan for getting additional information? Includes:

    (a) What are the general requirements of a problem-solving approach, and what general approach will they pursue?

    (b) How to obtain needed information?

    (c) What are achievable milestones, and what are metrics for evaluating the process?

    (d) What are possible alternative outcomes and parts that may arise during problem solving process, both consistent with predictive framework and not, and what would be paths to follow for the different outcomes?

16. What calculations and data analysis are needed? Includes planning how to interpret data.

17. What is the best way to represent and organize available information to provide clarity and insights?

18. Is information valid, reliable, and believable?

19. As new information comes in, particularly from experiments or calculations, how does it compare with expected results?

20. If a result is different than expected, how should you follow up?

    (a) Does potential anomaly fit within acceptable range of predictive framework(s)

    (b) Is potential anomaly an unusual statistical variation, or relevant data? Is it within acceptable levels of uncertainty?

21. What are appropriate conclusions based on the data?

22. What is the best solution? Involves evaluating and refining candidate solutions throughout problem solving process. May include deciding:

    (a) Which of multiple candidate solutions are consistent with all available information and which can be rejected?

(b) What refinements need to be made to candidate solutions?

23) Are previous decisions about simplifications and predictive frameworks still appropriate?

(a) Do the assumptions and simplifications made previously still look appropriate in light of new information?

(b) Does the predictive framework need to be modified?

23. Is additional knowledge/information needed? Could involve:

(a) Is solver's relevant knowledge sufficient?

(b) Is more information needed and if so, what?

(c) Does some information need to be checked?

24. How well is the problem-solving approach working and does it need to be modified including do the goals need to be modified?

25. 26) How well does the chosen solution hold?

(a) Decide by exploring possible failure modes and limitations - "try to break" solution.

(b) Does it "make sense" and pass discipline-specific tests for solutions of this type of problem?

26. What are the range and contexts where solution applies, and what are the broader implications? What outstanding problems in field might it solve? What novel predictions can it enable?

27. What is the audience for communication?

28. What is the best way to present the work to have it understood and its correctness and importance appreciated?

Chapter 7

Appendix B: Supplemental Figures for Assignment Analysis

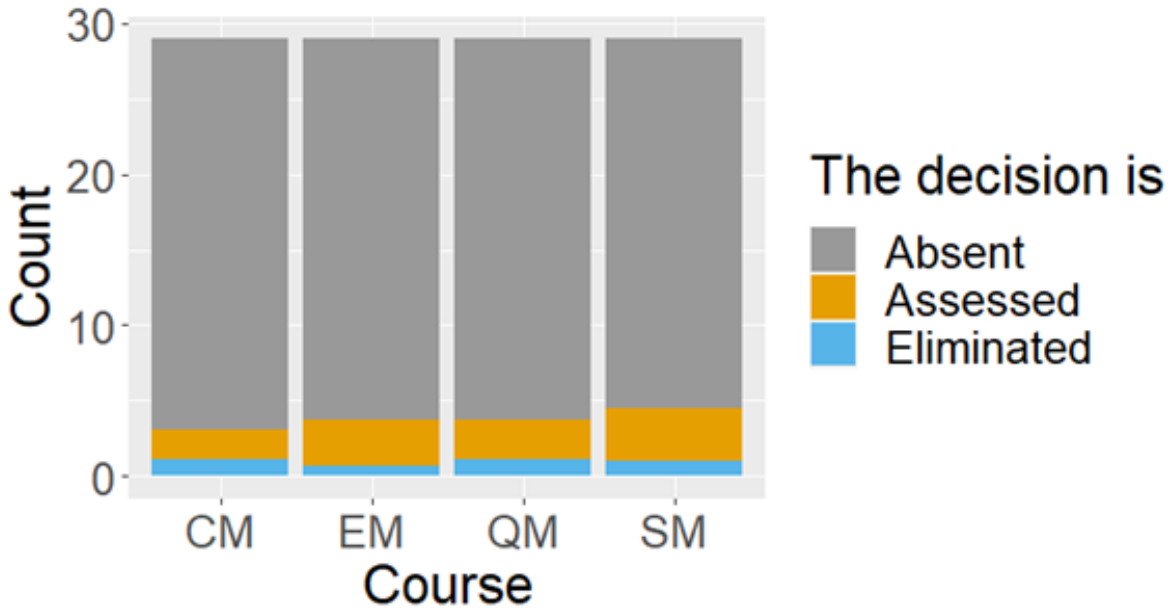


Figure 7.1: Decisions by course. The average number of decisions were shown for each course. The maximum number of decisions was 29. The courses are Classical Mechanics, CM; Electricity and Magnetism, EM; Quantum Mechanics, QM; and Statistical Mechanics, SM. Assessed means the solver must make this decision to solve the problem. Eliminated means the problem statement made this decision. Absent means the decision was neither assessed nor eliminated.

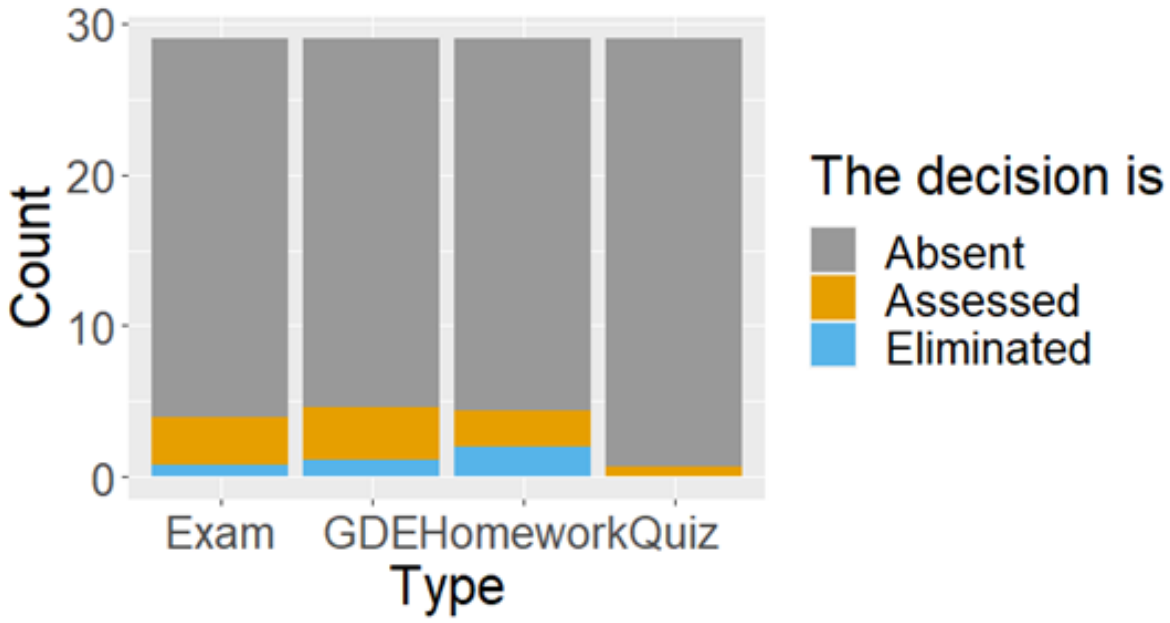


Figure 7.2: Decisions by assessment type. The average number of decisions were shown for each assessment type. The maximum number of decisions was 29. Assessed means the solver must make this decision to solve the problem. Eliminated means the problem statement made this decision. Absent means the decision was neither assessed nor eliminated.

Chapter 8

Appendix C: Faculty Interview Protocal

The following protocol was used to conduct the semi-structured interviews. Questions 1 through 12 were used during the open-ended portion of the interview. Question 13 was used to begin the closed-ended portion of the interview. Not all questions were explicitly asked. The interviewer selected questions as appropriate, during the interview.

Thank you for participating in our study of problem-solving in STEM. My name is [NAME], and I am a researcher on this project. Today we are going to ask you some questions about the learning goals you have for students in your classes and in your research lab. The goal is to eventually use this information to develop better assessments and teaching methods for undergraduate and graduate courses. Thank you for taking the time to talk with us today.

Your responses will be kept confidential by the research team, and it is important that none of you repeat what is said in this room to others. If an issue with a particular instructor or student arises, you can choose not to give a name.

We are going to record this session for research purposes. You have received a sheet outlining our data practices and your rights as a participant. Please indicate verbally whether you consent to have your data used for research.

1. Please briefly describe your research area and what courses you teach.
2. How long have you been...
   (a) Teaching graduate student courses
   (b) Mentoring students in a research capacity
3. What criteria do you use to determine whether a student is successful in your research lab – both short term (by end of PhD) and longer term?

4. What do you do to monitor if students are progressing and developing as intended?
5. What do you want students in your research lab to be able to do once they leave your lab?
6. What kinds of problems do students typically work on for their thesis projects?
    (a) Do students choose their own research projects, or do you assign them to students?
    (b) How do you decide what makes a good research project for a student?
7. Briefly describe the learning goals for your course – what do you want students to be able to do at the end of the course?
8. How do you judge whether a student is successful in achieving these goals?
9. How do you monitor students' progress toward these learning goals?
10. Which learning goals do students most and least successfully acquire?
11. What kinds of problems do students typically solve in your course?
    (a) Are they mainly from textbooks or do you make them up yourself?
        (i) If you make them up – where do you look for inspiration?
    (b) Do any of them require students to make their own assumptions or look up extra information?
    (c) Why do you assign students these particular problems?
    (d) If you had the time and resources, how would you change the kinds of problems you assign students in your course?
12. What is different about problem-solving in the classroom versus in the laboratory?
13. I'm going to give you a list of skills we have previously identified as important elements of good problem-solving. As I read each one aloud, please indicate whether you think this is an important thing for a student to be able to do 1) after completing all their university coursework and 2) after completing their research training.
14. Are there any other skills you want students to be able to acquire that were not on the list I just gave you?