An accurate and practical method for assessing science and engineering problem-solving expertise

Argenta Price, Shima Salehi, Eric Burkholder, Candice Kim, Virginia Isava, Michael Flynn, and Carl Wieman

ABSTRACT

The ability to solve authentic real-world problems in science, engineering, and medicine is an important goal in post-secondary education. Despite extensive research on problem solving and expertise, the teaching and assessing of advanced problem-solving skills in post-secondary students remains a challenge. We present a template for creating assessments of advanced problem-solving skills that is applicable across science, engineering, and medical disciplines. It is based on a cognitive model of the problem-solving process that is empirically grounded in the study of skilled practitioners ('experts') solving authentic problems in their disciplines. These assessments have three key features that overcome shortcomings of current assessment approaches: 1) a more authentic amount and sequence of information provided, 2) opportunities for students to make decisions, and 3) scoring based on comparison with skilled practitioners. This provides a more complete and accurate assessment of authentic problem-solving skills than currently exists for science, engineering, and medicine. Such assessments will be valuable to instructors and curriculum designers for evaluating and improving the teaching of problem solving in these disciplines. We provide examples of problem-solving assessments that illustrate the use of this template in several disciplines.

Introduction

Solving real-world problems is the primary goal of skilled personnel in science, engineering, and medicine in both industry and academic settings. However, identifying effective ways to both assess and teach such problem solving has been challenging. Assessment
and teaching are closely linked, because without accurate ways to measure problem-solving skills, it is impossible to evaluate and improve the methods for teaching these skills.

Different aspects of expertise, problem-solving, and assessment of problem-solving have been studied over many decades. For reviews of the extensive literature, see Frensch and Funke (1995), Csapó and Funke (2017), Dörner and Funke (2017), and Ericsson et al. (2018). The early research on expertise established a general focus on the learning and organising of knowledge in a subject (e.g. Chi, 1978; Siegler, 1978) and identified differences in how experts and novices solve simple problems such as those in physics textbooks (e.g. Chi et al., 1981; Hegarty, 1991; Larkin & Reif, 1979; McCloskey, 1983), computer programming (Card et al., 1983; Kay, 1991), simple mathematical calculations (Resnick, 1982), and more recently chemistry (Randles & Overton, 2015). More recent work (see Farrington-Darby & Wilson, 2006 for review) has brought in many other elements of expertise, including planning, motor skills, situation awareness (Endsley, 2018), awareness and use of relevant resources (Lenzer et al., 2020), and various types of decision making (Mosier et al., 2018).

A criticism of this body of work is that it is not clear whether what has been measured is necessary or complete in capturing what makes someone an expert performer while doing their regular jobs (doing physics research, etc.). Sternberg (1995) stated this as, ‘We need not only study experts, but study these experts doing the tasks that render them experts.’ The assessment template presented here is based on work doing exactly that for many scientists, engineers, and doctors. Price et al. (2021) extensively studied how skilled practitioners (‘experts’) in many disciplines of science and engineering solved problems in their work, and on that basis developed a framework of decisions-to-be-made that characterises their problem-solving process.

In this article we present a template for developing assessments of problem-solving mastery, intended for use in discipline-specific post-secondary educational programmes. Our approach to assessment is organised around the set of twenty-nine decisions that skilled scientists and engineers make, as identified by Price et al. (2021). When solving a typical problem within their discipline, these experts call upon disciplinary knowledge and apply it in consistent ways to make each decision. A notable feature of assessments based on these decisions is that they involve authentic discipline-specific problem scenarios and compare performance (decisions) with that shown by skilled practitioners in the discipline.

Prior assessments of problem-solving have been developed both as research instruments and for standardised assessment of students. An extensive thread in this work has been to explore tasks of varying levels of complexity that do not require specialised knowledge or skills to solve (see Frensch & Funke, 1995 for review). We would refer to these as ‘complex puzzles,’ as they test reasoning skills, but they do not test the appropriate application of specialised knowledge, as is extensively required in every authentic science and engineering problem. Recent work on the standardised assessment of problem-solving skills has recognised the importance of the ‘acquisition and application of knowledge,’ and led to the development of more innovative assessments that do not provide all the necessary information up front (Csapó & Funke, 2017). A second important feature recognised in that work is that the complexity of real tasks involves uncertainty, and the solution process involves ways to reduce the uncertainty (Ch. 3 in...
These research and assessment efforts have contributed a better characterisation of the complexities of ‘complex problem solving.’ The outcomes most relevant to the current paper are the OECD formulation of a framework of complex problem solving and the PISA exams assessing it (Csapó & Funke, 2017; Ramalingam et al., 2017).

The OECD expert group condensed the existing research on cognitive processes required for complex problem solving into four groups, 1) exploring and understanding; 2) representing and formulating; 3) planning and executing; and 4) monitoring and reflecting. The resulting PISA assessments of problem solving are constructed based on this framework. They also recognised that ‘when solving authentic problems, individuals may well need to move beyond a linear, step-by-step model. In practice it is often necessary to cycle through these processes as progress is made towards a solution’ (Ramalingam et al., 2017). Beyond being used for the PISA assessment, this framework has also been applied to analyze written student solutions to problems (Kelly et al., 2016). These results and conclusions are consistent with what we observed in our empirical studies of experts solving authentic science and engineering problems.

While these findings and the PISA assessments have several similar features to our assessment template, there are two significant differences. First, we provide greater specificity regarding the execution of these general problem-solving steps by articulating the specific decisions they involve. The second relates to differences in goals. With PISA, the goal is to measure general problem-solving skills across large populations (secondary school students across many countries). This is done by measuring the performance of tasks of varying levels of complexity that do not require specialised knowledge to solve (see Frensch & Funke, 1995 for review). Our goal is to measure problem-solving skills of students (or employees) who are in training to become experts at solving problems in a specific domain, like the skilled practitioners in science and engineering that we studied. An essential element of their skill is recognising and applying appropriate specialised knowledge with limited information, and therefore that is a basic component of our assessments.

Our work is grounded in several principles for what constitutes good educational assessments. Good assessments are part of good pedagogical strategies, closely aligned with the desired educational outcomes to provide guidance for improving instruction and evaluating individual skills. Many principles of good assessment are discussed in the USA National Research Council study, Knowing What Students Know (National Research Council, 2001). Table 1 lists five principles from Knowing What Students Know that are particularly relevant in this work.

Schwartz and Arena (2013) add an important nuance to these principles with their ‘choice-based assessments.’ They argue that assessments in which students need to make choices are more meaningful than the traditional ‘knowledge-based’ assessments. The template we describe here is designed around having students make problem-solving decisions (a form of ‘choices’).

The shortcomings of traditional assessments in STEM

Schwartz and Arena (2013) and Knowing What Students Know (National Research Council, 2001) provide extensive discussions of why we need to develop non-traditional
assessments. The traditional standards for evaluating the quality of the test questions include 1) face validity, asking disciplinary experts if the facts in the item are correct, and 2) various psychometric tests to evaluate reliability, such as looking at the discrimination index, Cronbach alpha, test-retest repeatability, and factor analysis. These all measure how well the test questions reproducibly rank students, but such psychometric measures say nothing about whether the test is measuring something meaningful, and far too often they do not, as nicely described by Ryoo and Linn (2015).

Conventional assessments generally rely on multiple-choice items that focus on measuring factual knowledge about a single scientific concept, rather than assessing students’ integrated understanding of science (Lane, 2004; Songer, 2006). These multiple-choice items show whether students can select the correct answer, but they do not allow students to demonstrate why they made such a decision or to describe the evidence they used (Ennis, 1993). In addition, most students are only tested on the topics they studied in the latest unit and often quite superficially; there is seldom the expectation of applying concepts learned in prior material.

| Elements of good assessment | Elaboration – quote from Knowing What Students Know | Implementation in our assessment template |
|-----------------------------|-----------------------------------------------------|------------------------------------------|
| 1) Based on a model of cognition | ‘A model of cognition and learning should serve as the cornerstone of the assessment design process. This model should be based on the best available understanding of how students represent knowledge and develop competence in the domain.’ | The assessment template is based on a framework of problem-solving decisions and practices that have been empirically identified (Price et al., 2021; Salehi, 2018). |
| 2) Guided by features of expert-novice differences | ‘Studies of expert-novice differences in subject domains illuminate critical features of proficiency that should be the targets for assessment.’ | The problem-solving decisions were identified by studying the problem-solving process of experts. The goal of the assessments is to measure whether and how well students make these decisions, as compared to the performance of skilled practitioners. |
| 3) Encompass complex aspects of achievement | ‘This body of knowledge [on what constitutes expertise] strongly implies that assessment practices need to move beyond a focus on component skills and discrete bits of knowledge to encompass the more complex aspects of student achievement.’ | Assessments built following the template require the test-taker to complete a large portion of the problem-solving process and engage in most of the complex problem-solving decisions. |
| 4) Involve using knowledge to reason. | ‘What matters in most situations is how well one can evoke the knowledge stored in long-term memory and use it to reason efficiently about current information and problems.’ | All the decisions on which our assessment template draws require problem-solvers to reason with their relevant knowledge and mental models of the problem context. |
| 5) Situate in a real-world context | ‘Traditional testing presents abstract situations, removed from the actual contexts in which people typically use the knowledge being tested. From a situative perspective, there is no reason to expect that people’s performance in the abstract testing situation adequately reflects how well they would participate in organised, cumulative activities that may hold greater meaning for them. From the situative standpoint, assessment means observing and analyzing how students use knowledge, skills, and processes to participate in the real work of a community.’ | The assessment problem scenario requires test-takers to use their knowledge in a more realistic situation, in alignment with the situative perspective. |
A traditional science and engineering assessment question imposes contextual constraints and asks questions in ways that do not probe the test-taker’s ability to make the problem-solving decisions that are required for true mastery. Skilled problem solvers are faced with masses of information in any real problem context, and they can distinguish what is relevant and what is irrelevant. Additionally, they are skilled at deciding what information is needed, and how to obtain that information (National Research Council, 2001). None of these skills are measured by traditional sequestered tests in which only, and all, the needed information is provided.

**Assessment design: principles and guidelines**

**Goal for problem-solving assessments**

In this article, we describe a generalised design for developing problem-solving assessment instruments in science and engineering disciplines. Our goal is to measure students’ progress toward problem-solving mastery in a specific discipline, solving like a skilled practitioner. To do this, we apply the guiding principles for good assessments described above to assessments in which students are given an authentic problem scenario and are then asked to make a set of problem-solving decisions, based on the framework of Price et al. (2021). The general assessment structure is shown in Figure 1 (see ‘problem-solving assessment template’ section). These assessments differ from traditional exams in several ways: 1) the amount and sequence of information provided, 2) the opportunities for students to make decisions, and 3) the scoring based on comparison with skilled practitioners.

While the PISA problem-solving assessments (Ramalingam et al., 2017) also attempt to capture similar features, there are significant differences in approach. For example, the PISA assessments test whether students, when given a computer simulation, can collect from it the information necessary to solve the problem. However, this task does not call on the student to make many of the other decisions that we have seen are critical steps in authentic problem-solving tasks of experts: they do not need to decide what are the criteria for a good solution, what kind of information is needed to solve the problem, nor how to find that information. The first is specified in the problem statement and the second and third are set by the small set of options given. In addition, as problem solving itself is a dynamic process, the PISA assessment developers decided that this process could only be adequately assessed through a dynamic solution process based on computer simulations. While we recognise the strengths of computer simulations and have used them to study and assess problem solving, they have substantial practical challenges for implementation and grading. We believe that our template provides an alternative approach that does not require computer simulations and complex scoring of student actions. It does this by guiding the dynamic solution process around the decisions the student would need to make and comparing those decisions with those made by experts.

Assessments following this template can be used for formative, summative, or programme-evaluation purposes, and they are primarily applicable to advanced undergraduates or post-graduates. While we would not expect students, at least in early years of their training, to perform like skilled practitioners, our goal is to develop assessments that can measure their progress towards this mastery. This serves as the most direct and useful
| Decisions-Based Assessment Template |
|-------------------------------------|
| **Authentic scenario (provided)**  |
| A) What are key features?          |
| Identify, and which are relevant?  |
| B) What are possible solutions?    |
| Propose, or provide feedback on     |
| C) What info needed?               |
| Describe, then choose from options, |
| explain why relevant. May          |
| include plan how to collect.       |
| D) Reflect on assumptions and       |
| applicable knowledge (predictive   |
| framework)                         |
| Have assumptions / simplifications |
| been appropriate?                  |
| **Optional: progressively more     |
| specific questions about A-D        |
| **More info (provided)**           |
| And/or provide more specific       |
| features to consider.              |
| E) How to interpret info?          |
| Draw conclusions and/or provide    |
| feedback. May include data analysis |
| as appropriate.                    |
| F) Reflect on info and strategy    |
| What additional info needed? How    |
| will it be used?                   |
| **Optional: progressively more      |
| specific questions about E-F        |
| **Solution Selection**             |
| G) What is best solution?          |
| Generate, choose from options, or  |
| approve/disapprove a proposed      |
| solution                           |
| H) Reflect on solution             |
| Summarize and justify and/or describe uncertainties. Propose next steps. |
| Optional: Multiple rounds w/ B-F   |

**Figure 1.** Assessment template. The basic components of the template are based on the decisions involved in problem solving practices (Price et al., 2021; Salehi, 2018). 1) provide an authentic problem scenario; 2) Ask a series of questions that require test-takers to make decisions about problem definition and planning how to solve; 3) Provide more information, which includes both information an expert would seek and irrelevant information novices might encounter or seek, and/or more specific criteria needed to define the problem to be solved; 4) Ask a series of questions that require decisions about interpreting information and drawing conclusions; and 5) Have test-takers choose and reflect on their solution. Throughout, the test-taker also must reflect on the planning and interpretation decisions they made. To emulate the problem-solving process used in real situations, there can be multiple rounds of providing new information, asking for interpretation and reflection, and then asking the test-takers to revisit their problem definition decisions (A-D) or plan next steps. In addition, it is useful to ask progressively more specific questions that constrain the planning and interpretation.
metric of how well students are mastering the desired skills. These metrics provide feedback both to the students and the instructor/programme about their progress compared to a meaningful external standard.

**Characterising our assessment construct – measuring mastery in disciplinary problem solving**

We start with a cognitive model for problem solving in a discipline. What does it mean to be a ‘skilled problem solver’ in a discipline and how can this be measured? At one level, it means measuring how often a person can produce the same *final solution* as do experienced practitioners. However, this is not very informative for identifying areas for improvement. A more rich and useful characterisation requires looking at the full complexity of the problem-solving process and measuring how well a person can perform all aspects of that process, not merely the final solution. For academic disciplines such as science and engineering, we argue that the type of expertise needed is ‘adaptive expertise’ (Hatano & Inagaki, 1986), which requires the ability to solve problems in novel contexts that do not have a solution process that is composed of a well-defined set of steps. Such problems are also labelled as ‘ill-structured problems’ because they cannot be solved by deterministically following a set of instructions (Simon, 1973). Therefore, assessments to measure problem-solving in these disciplines need to probe the problem-solving process of ill-structured problems.

We describe this process in terms of the 29 decisions a skilled solver needs to make (Price et al., 2021) and the reflective and execution practices defined by Salehi (2018) (see supplemental table). The set of decisions is common across different fields and includes such items as: deciding if the problem is worthwhile; deciding what are key features of the problem; deciding what information is needed; deciding on priorities; deciding on the believability of information obtained; etc.

An essential element of these decisions is that they are made with limited, i.e. insufficient information. This means the decisions made are never a certainty, but rather a ‘best guess’ based on the knowledge and standards of expertise in the discipline. Skilled practitioners clearly recognise the inherent uncertainties in those decisions and regularly reflect on the accuracy of those decisions as new information or insights become available. All the decisions require the problem solver to recognise relevant knowledge they have and additional knowledge/information they need to obtain to correctly make the decisions in question.

For an assessment to capture a large portion of the problem-solving decisions, the problem should be consistent with the situative perspective outlined in *Knowing What Students Know* and the principles of ‘authentic assessment’ (National Research Council, 2001; Villarroel et al., 2018). That means the assessment tasks should be similar to what a practitioner might encounter in their work, with all the associated complexities and constraints. The assessment should probe the test-takers’ decisions and their justifications for their decisions, to allow these to be compared with the decisions of skilled practitioners. Students can then be evaluated on their problem-solving decisions at two levels: first, whether they engage in the same decisions as skilled practitioners when presented an authentic problem, and second, the quality of their decision making – the reasoning and knowledge behind their decisions – as compared to the skilled practitioners.
Discipline-based assessments using a discipline-independent template

Solving any authentic STEM problem requires correctly recognising and applying specific knowledge (Harris et al., 2019). Thus, while the set of 29 decisions-to-be-made is discipline-independent, the knowledge used to make those decisions is nearly always discipline-dependent.¹ Our general template for assessing problem-solving is used to create a specific instrument by embedding the desired disciplinary expertise in it, including the relevant disciplinary knowledge, through the choice of problem situation and topic-specific prompts. The latter is typically done by providing test-takers with additional information and seeing how they apply it.

We break the assessment down into individual decisions for several reasons. First and most important, it provides a more meaningful assessment. Rather than simply measuring the quality of a final solution, it provides detailed information on specific strengths and weaknesses in the many facets of the problem-solving process. Second, it provides the specificity needed for good formative assessment and feedback that is required for deliberate practice (Ericsson, 2006; Schwartz et al., 2016). Last, it allows students to engage more effectively in the problem by constraining the solution spaces of the individual questions.

Design principles of assessments to probe problem-solving decisions

The process of developing an assessment to measure how well students are learning to make the problem-solving decisions is, in principle, straightforward. Choose an authentic scenario and structure the problem so that the solver is asked to make a relevant subset of the 29 decisions. The chosen problem scenario requires the content knowledge students are expected to know. Depending on the goals, the assessment will often repeat some decisions as more information is provided. The specific flow of questions and information provided is decided based on cognitive task analysis of how practitioners in the discipline solve related problems (Hoffman & Lintern, 2006; Price et al., 2021). We contrast this type of assessment with traditional STEM assessments in Figures 2 and 3.

Figure 2A shows a problem from a chemical engineering textbook (Duncan & Reimer, 2019b) that is difficult and provides an authentic context. However, nearly all the decisions that would be needed to solve an authentic problem are made for the students: they are told what assumptions to make, what information they need (e.g. ‘assume’, ‘ignore’, ‘use table of properties [given],’ …), and they are told how many elements (3) to change. Panel 2B shows how this problem was adapted. Now the students are required to make problem-solving decisions with limited information, like a chemical engineer. The problem is broken down into a set of questions, each of which requires students to make one or a few problem-solving decisions. Additional information is provided at various points, and then students need to act on that information. Finally, the students describe and reflect on their final solution.

Our decomposition of the problem-solving process into a set of decisions may seem similar to the common instructional practice of breaking down a complex problem into parts, to guide students through the solving process. However, our problems are cognitively quite different, as illustrated in Figure 3.
A) Textbook problem: potentially difficult, but almost no decisions.

“Tetrachloroethylene can be synthesized from carbon tetrachloride by the [shown] consecutive pyrolysis reactions at 800 °C. The first reaction is irreversible, but the reaction does not go to completion... [details about reactions & what products would come out in the effluent of reactor]... So rather than recycle CCl₂-CCl₃ to the reactor, the CCl₃-CCl₂ is separated and sent to a second pyrolysis reactor, but at 500°C.

Consider the process flowchart [see Duncan and Reimer, 2019b]. Use the chemical reactions above and the table of physical properties to find at least 3 errors in the design. Ignore errors due to omitted pumps. Assume solid-gas separations are ideal; no gas leaves with the solid. Also, find two ways to improve the process, for example by eliminating a separator and increasing the product yield. Ignore improvements by adding heat exchanges. Finally, change the process flow diagram to fix the errors and incorporate the improvements. Note: solid CCl₂-CCl₃ is soluble in liquid CCl₄-CCl₃.”

B) Decisions-based assessment, troubleshooting format

You are an engineer responsible for designing a process based on the above chemical reaction’s relevant physical properties (click here). You have a particularly ambitious intern who wants to try their hand at process design, so you let them do the preliminary design sketch [see Duncan and Reimer, 2019b]. Remember, you are ultimately responsible for the design and justifying it to your manager.

Subset of Questions:
1) What criteria would you use to assess this design? [List & explain]
2) Given the information you have, does the intern’s design meet all your criteria? Please explain.
3) What modifications are necessary to make this process work?
4) You notice some minor errors in the design and correct them [see image]. What other feedback would you give the intern regarding this design after making these corrections?
5) Given the information you have, is the design optimal with respect to both material and energy consumption?
6) After considering all available info, describe any and all changes you would make to the original design.

Figure 2. Contrast between textbook and decisions-based problems. A is a textbook problem (Duncan & Reimer, 2019a, 2019b) in which most of the decisions are already made for students. B is how we adapted the problem so that test-takers make decisions following our template. B shows a ‘trouble-shooting’ context in which additional information and constraints are added iteratively (Burkholder & Wieman, 2020).

Figure 3A shows a typical physics problem from an introductory physics course. Each step asks students to execute an already-made decision. Important decisions such as how to represent the problem, what assumptions to make, and which concepts apply, are already made, so the student simply follows the appropriate procedure to solve. In contrast, Figure 3B shows our medical diagnosis assessment. Students are prompted to make and justify a series of decisions before being given the outcome of certain decisions (e.g. they are asked what parts of the physical exam they will focus on before being given the results of the physical). Based on the typical progression of a patient diagnosis, they go through several rounds of deciding what information to collect and receiving additional information, eventually making and justifying a diagnosis. Although the disciplines are different for the two problems to provide a greater range of examples, the contrast is clear.
In the assessment of a complex skill, there is a fundamental tension between the validity of what is being measured and the practicality of administering and scoring the assessment. Too much constraint means the important resources and decision processes making up expertise are not probed, while too little constraint results in responses that can vary so much that it is impossible to evaluate them.

In general, the optimal balance between validity and practicality can be achieved by choosing a problem scenario and decision questions that constrain the solver, but not too much. For some disciplines, it is possible to find realistic problems that have a

---

**Figure 3.** Contrasts between a typical ‘scaffolded’ problem and a decisions-based problem. A is a typical problem from a physics course that makes all decisions in the process of breaking down the problem into steps and leaves only procedures to be followed. B is an authentic decisions-based problem following our template, in the context of medical diagnosis, which breaks the problem down into a stepwise series of diagnostic decisions.

**Constraining the solution space enough but not too much**

In the assessment of a complex skill, there is a fundamental tension between the validity of what is being measured and the practicality of administering and scoring the assessment. Too much constraint means the important resources and decision processes making up expertise are not probed, while too little constraint results in responses that can vary so much that it is impossible to evaluate them.

In general, the optimal balance between validity and practicality can be achieved by choosing a problem scenario and decision questions that constrain the solver, but not too much. For some disciplines, it is possible to find realistic problems that have a
sufficiently constrained solution space (the combination of the answer and possible routes to that answer) that the assessments can mirror the actual practitioner problem-solving context closely and be readily scored. Medical diagnoses (see Figure 3B) are examples. For disciplines like engineering design or biochemistry, the solution space for realistic problems is too large. For those cases, we find it works to constrain the solution space by introducing troubleshooting scenarios with proposed possible solutions.

It is straightforward to achieve a desirable balance between validity and practicality using these design principles: 1) Focus on the most important decisions, 2) Provide a ‘troubleshooting’ scenario, 3) Progressively narrow the questions, 4) Use a combination of free-response and closed-response questions. For the populations we have probed (undergraduate and junior graduate students), nearly all their responses are different from the experts, and questions and rubrics based on these principles give good discrimination.

Focus on the most important decisions
In any assessment, there is a tradeoff between length and thoroughness. So, emphasis should be given to probing decisions most important to both the expertise in the relevant discipline and the educational context. Some decisions are not useful to ask students, e.g. ‘what is important in the field?’ Some decisions, notably, ‘how to narrow down the problem?’ are relatively long and difficult to probe and score. We make the pragmatic decision to not include these decisions in our assessment template. In addition, the importance of particular decisions varies with discipline, so some may be probed multiple times in an assessment for one discipline but skipped in another (e.g. ‘what calculations or data representation are needed?’ may be asked repeatedly in an experimental physics problem but not asked in medical diagnosis).

Provide a ‘troubleshooting’ scenario
Where a problem-flow that mirrors real practice is not practical to evaluate, a ‘troubleshooting’ design (Adams & Wieman, 2015; Walsh et al., 2019) is a good alternative. In this design, the test-taker evaluates a solution proposed by a student/intern. Such an evaluation task can still involve mostly the same expert problem-solving decisions and relevant content knowledge, but the solution space is much more limited and so the responses are straightforward to characterise and score. Troubleshooting overlaps well with the set of decisions we seek to probe, because so much of the actual expert problem-solving process involves evaluating and modifying potential solutions, which are the same cognitive processes as used in troubleshooting.

Progressively narrow the questions
Another strategy is to narrow down the focus of questions as the assessment progresses. This provides discrimination across a wider range of problem-solving skill by allowing test-takers with higher levels of mastery to make more decisions early on, but then providing constraint and guidance for those with lesser mastery. For example, in the chemical engineering assessment, test-takers are first asked for their open-ended feedback on an intern’s proposed chemical plant design, then they are asked about specific aspects of the design. This structure allows us to see if the test-taker recognises important factors
without any prompting, as will be the case for high-level expertise, but also to see whether they can apply specific knowledge that they might not have recognised was relevant until being guided to consider it. We have found that experts usually generate a full plan or complete solution before these more specific prompts, while semi-novices need more prompting and information but are able to respond correctly when this is provided, but novices are unable to answer even with prompting. Thus, this structure of asking increasingly specific questions allows for better differentiation of skill level.

**Combination of free-response and closed-response questions**
Free-response questions allow test-takers to respond more authentically, so we choose free-response format, especially for the less-narrowed questions. However, these are time-consuming to score. After an assessment has been piloted by a large number of students, some of these free-response questions can be adapted to multiple-choice for easier computer grading, using student responses to guide item design. To accurately measure the student thinking, our multiple-choice questions are much more complicated than the typical MC format. Our questions have many answer options, largely based on what students say in our pilot testing, and students can select multiple items from the list.

**Assessment design: structure, implementation, and examples**

**Problem-solving assessment template**
We have developed assessments in several areas (medical diagnosis, advanced mechanical engineering design, chemical engineering plant design, natural hazard analysis, and biochemistry). Based on the formats and lessons learned from these assessments, we have created the template shown in Figure 1. Only the most fundamental decisions are shown in the template. However, making these decisions often requires making other related decisions and, depending on the disciplinary context and target population, an assessment designer could choose to ask additional decisions explicitly (see supplement for full decisions list).

**Problem scenario**
The problem scenario should be authentic, with: 1) no obvious single correct solution path, 2) an appropriate level of content knowledge required for the desired test-takers, and 3) a description including both relevant and irrelevant information, but not all the information needed to solve. Finding a good scenario that meets all the necessary criteria is the most difficult part of creating a good assessment.

An engineering problem scenario could be the design of an object or facility to accomplish a particular function, in science it could be to answer a realistically challenging question, and in medicine it could be to diagnose a realistic patient case. See examples in Figures 2B and Figure 3B.

There are a few additional features to consider when designing a scenario. First, avoid implicitly making problem-definition decisions that the student should make themselves (such as what assumptions to make and what information is needed). Second, when designing a troubleshooting scenario, we have found that students are often inclined to assume the mock-student answer is correct and not seriously consider potential
flaws, so at least some flaws need to be blatant. Third, consider the framing of the problem, because scenarios that on the surface look like textbook problems can set students into a mindset of needing to follow a particular known procedure.

**Problem definition and planning**

Providing limited information in the problem scenario means that test-takers are required to make decisions about how to define the problem. The assessment template (Figure 1) guides test-takers through this by asking them to (A) identify important features of the problem (which will help constrain their potential solutions and solving process) and/or (B) suggest possible solutions or provide feedback on a flawed mock-student solution (troubleshooting). Identifying possible solutions often requires students to make other decisions about how well the solution meets the goals of the problem (could lead to refinement of goals) and how to test the potential solutions. These problem-definition questions require students to make decisions about how to proceed, without following a standard procedure.

As examples, in the medical diagnosis problem, test-takers are asked, ‘What are key features in the patient’s history? Explain.’ followed by, ‘At this point, what diagnoses are you considering?’ In the chemical engineering problem, test-takers again identify key features, here in combination with goals, by responding to the question, ‘What criteria would you use to assess this design?’ Initially, they are asked to troubleshoot a mock-student solution, answering ‘Given the information you have, does the design meet your criteria?’ and ‘What modifications are necessary to make this process work physically?’

Later, test-takers are asked to (C) identify additional information they need and (D) reflect on the assumptions they have made and justify those decisions. In the context of deciding which information they need, they may also be asked to plan how to collect that information, or to critique a student’s plan for doing that. For example, there is a vast amount of information that a student might want to collect to solve our biochemistry problem, so to constrain the space, we ask test-takers to provide feedback on a student’s data collection and then rank potential options for additional information to collect. Thus, they must make decisions about priorities and planning in the context of deciding what information is needed.

There are several decisions that test-takers may need to make to successfully answer the problem definition and planning questions but that aren’t explicitly probed in the template. While the template does not specifically probe these decisions, that was an intentional choice made to limit the length of the assessment. If any of those specific decisions are particularly relevant to an instructor’s educational goals, questions about them are easily included in the assessments.

**Provide additional information**

As the test-taker progresses through the assessment, additional information is provided, often in the form of data collected, which the test-takers will need to interpret. The assessments we have created following this template are administered on computers, using survey software to break the assessment into sections so that students make decisions based on limited information before being provided new information. If the assessment includes multiple-choice questions asking what information is needed, survey branch logic is used so that the information given can depend on what students
request. To create an authentic situation in which students see the outcome of their decision, the new information is just the information they selected. An alternative, which can be used in paper-and-pencil or computerised assessments, is to provide a curated set of new information as a rescue point for students who might not have requested critical information. This curated information includes both the essential and some irrelevant information, so students still must decide which information is important.

Sometimes, the new ‘information’ isn’t data but is more specific criteria or features that narrow down the problem. This allows for better discrimination among inexpert solvers. An example of this type of ‘new information’ is in the chemical engineering assessment when test-takers are given multiple rounds of additional criteria (e.g., energy efficiency, safety) to consider in their evaluation of the mock-student solution.

**Data/information interpretation**

The new information provided requires test-takers to engage in data interpretation decisions (E). They interpret and draw conclusions from the new information and refine their possible solutions based on their interpretation. Assessments can probe this by asking questions like (medical example), ‘Given the information now available, what are your top 3 potential diagnoses?’ The extent to which decisions are probed about how to analyze or represent data, or about reliability of the data, can vary depending on the problem scenario. For example, our natural hazard assessment asks the test-takers to analyze data in the form of looking at maps of geologic and geographic features to draw conclusions about which are high-risk areas. The assessments also include questions (F) that probe reflection on information and strategy by asking the test-taker what additional information they need and how they will use it. Having test-takers justify their decisions both requires test-takers to reflect and gives us valuable insight into their decision process.

**Iterations of problem definition, information provided, and interpretation**

It can be useful to have multiple rounds of new information provided. These iterations can be used for different purposes: to break down the problem into particular aspects you would like to probe, to follow a natural sequence of information-gathering that an expert in the field would follow, or to probe other decisions that only come up in specific contexts (such as data analysis or data representation). These iterations also promote reflection and allow reflection decisions to have their natural consequences – reflecting on information can lead the test-taker to reconsider decisions about problem definition and decide to collect more information.

**Solution selection**

Finally, test-takers (G) commit to a final solution and (H) are asked to reflect on that solution by either summarising and justifying their choice, or by describing their remaining uncertainties and proposed next steps. The questions that ask for final solutions are generally direct: ‘What is your working diagnosis?’ in medicine, or ‘Will you incorporate these suggestions [about a chemical process design]?’ in chemical engineering. It is much more informative and authentic to see how test-takers decide on a final solution,
rather than just giving them a correct solution and asking them to justify it, as is done in some traditional medical assessments.

Reflection decisions are critical during problem solving (Salehi, 2018; Winne & Azevedo, 2014). The assessment template includes multiple places for the test-taker to reflect on decisions they’ve made and revise their solution approach, assumptions made, or information needed in response. Reflection on a solution allows review and revision of the problem-solving process. We probe it in two different ways. First, we ask the test-taker to summarise and justify their solution. For example, the chemical engineering assessment asks them ‘Why’ they accept/reject proposed changes, then asks, ‘After considering all available information, describe any and all changes you would make to the original design.’ We also ask test-takers to propose next steps, for example ‘What next steps would you take?’ in the medical assessment, to see if they would offer the confirmatory tests an expert would use. In some problems, we also ask them to reflect by asking if they have uncertainties or other considerations that have not yet been addressed.

The final summary and reflection questions have proven particularly informative, revealing how experts focus on key information while students either leave out part of their process or provide extraneous detail. For example, in the chemical engineering assessment, we found that skilled engineers provided a detailed summary of their final design and how it differed from the original flawed design, while students only summarised the most recent process and forgot modifications and additional information previously suggested. In the medical diagnosis assessment, students often provided a detailed listing of all information available and the final diagnosis, while skilled doctors focused on why key information was important for arriving at and checking the final diagnosis.

**Flexibility of the template**

Our assessment template is intended to be a flexible guide to the overall format of problem-solving assessments. Figure 4 provides a summary of the sequence of questions and information provided in four assessments we have created. All these assessments have the same basic structure, with the decisions moving from problem definition through solution selection and justification. There are also many differences in the details, which illustrate how the sequence of decisions probed will vary by discipline and are guided by examples of how experts in the respective disciplines solve problems. The relevance of particular decisions for different levels of student or the disciplinary context determines the selection of problem scenario and the particular questions asked. For example, in solving problems in biology and earth sciences, decisions about ‘what information do I need and what conclusions can I draw from it’ are particularly important, and hence are emphasised in those assessments. In chemical engineering the emphasis is on identifying and applying appropriate design criteria, so students are repeatedly asked to make decisions about potential designs.

**Assessment development and pilot testing**

After using the generalised template to develop a draft assessment from a disciplinary-specific scenario, we iteratively test and refine the assessment, following the process
described by Adams and Wieman (2011) and Walsh et al. (2019). This is similar to the evidence-centered-design-based approach described in Harris et al. (2019), though differs in the method for developing a scoring rubric. We create a first draft of the problem scenario and questions and ask people with various levels of expertise in the field to review the wording. Then ~5 volunteer student test takers from the intended test population take the test in think-aloud interviews. We revise wording for clarity and consistency of interpretation. Expert consensus is the basis of our scoring and provides evidence of validity, so we also have ~5 skilled practitioners take the test in think-aloud interviews. We evaluate their responses to ensure there is a high level of consistency, revising the questions accordingly if there is not. We then test the revised draft with a few dozen students from the target population. This typically involves a combination of think-aloud and normal administration of the test, mostly the latter. This work was approved by the Stanford Institutional Review Board (IRB no. 48785), and informed consent was obtained from all participants.

This testing further checks the question wording and can lead to small revisions. It can also reveal opportunities for replacing open ended questions with multi-choice options. The latter is preferred where valid, as it makes the scoring easier. We focus on two main questions in this stage of testing:

![Figure 4. Question sequence across assessments used in different disciplines. Each item represents either provided context/information or a question to be answered, colour coded according to the template item type (white in bottom row means assessment has ended). Items with multiple colour codes indicate that multiple decisions were probed in one question (or where half-yellow, information was provided in the context of a question being asked). Reflection decisions are scattered throughout, usually as part of questions in which they make other decisions but not separately represented here in the interest of space.](image-url)
(1) Can the test be completed in an appropriate time?
Our target is between 30 and 60 min, because that is the period of time that we have found instructors are comfortable having their students spend on an assessment, particularly if they administer it during class.

(2) Are any of the questions providing redundant information?
From the combination of interviews and pilot tests, we occasionally find that students, and, less frequently, skilled practitioners, will provide essentially the same information in response to different questions, even though we intended to probe different components of problem-solving. In such cases, we remove the redundant question.

The extent of pilot testing required depends on the goals for the assessment. An instructor developing an assessment for use in their course would have lower requirements than someone developing an assessment to evaluate a departmental programme or to compare across institutions. To have sufficient evidence of validity and reliability for such broader use, more extensive testing with students across multiple populations and levels of experience is required (National Research Council, 2001). If substantial changes (beyond minor wording clarifications) are made during the pilot-testing process, we administer the test to ~5 additional experts for development of the scoring rubric. It is also desirable to have more experts, especially if there are questions on which their responses are variable.

Development of scoring rubric
Our criteria for a good rubric are that it should have skilled practitioners tightly distributed around a high score, while scores of the test population should show a large distribution, and thus indicate a large degree of discrimination. To create such a rubric, we start by looking at the responses of the experts. We select the portion of solutions that are shared by nearly all experts to be the ‘mastery’ answer (Flynn, 2020). Then we score the skilled practitioner and student solutions according to the degree of overlap between their answer and the ‘mastery’ answer. When the response is a list of items, such as the important features to consider when evaluating a solution, we create a score based on the fraction of the ‘mastery’ list they contain, with points subtracted for including items not chosen by experts. In some cases, it is somewhat arbitrary how fine-grained the rubric should be, as there will be a range of specificity and detail across the different skilled practitioner responses. In such cases, the level of detail of an answer that is included in the scoring rubric is determined by looking at the standard deviations of the scores of the experts and the scores of the test-takers. We add details in the rubric if those additions increase the standard deviation of the test-taker scores substantially while making a small fractional increase in the standard deviation of the scores of the experts. If adding a detail to the rubric increases the standard deviation of the expert scores by a significant fraction (greater than 25%), we take that as an indication that there is not a strong enough consensus on that item, and so we do not include it as part of the mastery solution.

Although we want questions and a scoring rubric that provide good discrimination as to the level of expertise across our test takers, we do keep some questions where this is not the case. There are questions with low discrimination because all student test-takers score
very low compared to skilled practitioners, but the questions target problem-solving decisions that instructors or employers will recognise as important but were not adequately taught.

**Assessment validity and reliability**

Our approach to assessment validity is aligned with the approach for evaluating ‘instructionally relevant assessments’ described by Pellegrino et al. (2016). They emphasise collecting multiple forms of evidence to evaluate how well the assessment is measuring its desired goal. Underlying this is the need to explicitly define the knowledge and skills being assessed in terms of the evidence needed to show the skill has been mastered. In our assessments, the skills are defined in terms of the expert decisions and evidence of whether students have mastered making the decision is gathered by comparing their performance to the expert consensus. We use the information collected during the pilot testing with experts and students described above to evaluate the cognitive and inferential validity of the assessment (Pellegrino et al., 2016). We observe how well the assessment captures the expert problem-solving process and how well the students’ responses reflect their process. For all scored questions that survive the initial pilot-testing, the responses of experts have a large degree of overlap and only minor individual differences. This provides additional evidence of validity of the assessment question, as it shows it is measuring a well-defined decision process used in the discipline.

We also use larger-scale testing with students, administered in courses at multiple institutions ideally over multiple years, to further test inferential validity where appropriate. We look to see if students score higher when they have had courses or experiences (research or internships) that involve similar problem-solving decisions in the discipline as those on our assessment. For the assessments we have created to date, the data are limited, and students are generally far from mastery, but the assessments have all met this test of validity. As examples, we have data showing second year medical students have widely varying diagnostic skills, but nearly all are well below those of experienced doctors. In chemical engineering, we have found that students with prior industry design experience scored better on their assessment than students that did not have such experience. The differences were largest for recognising errors in the flawed chemical plant design (a 0.5 standard deviations difference) and suggesting improvements (0.3 s. d. difference) (Burkholder & Wieman, 2020).

During the larger-scale pilot testing we also look at the discrimination of the individual questions. Unlike most traditional assessments, we place less importance on how well the questions rank students relative to each other and instead look at how well they discriminate between students and experts.

We do not determine the reliability of these assessment instruments using standard psychometric tests such as Cronbach alpha and factor analysis, etc. As discussed in Adams & Wieman, 2011, those are unsuitable to this type of exam. Such psychometric tests look at correlations in student responses across different items. Because of our desire to make the assessments as efficient as possible, our questions are specifically designed to avoid overlap and hence correlation, as that provides redundant information. Each question is, to the extent possible, probing a different decision, requiring different background knowledge, and/or a different analysis process. However, all the questions
refer to the same basic context and problem, and hence some of them call on related knowledge. This introduces some correlations between responses to different questions, but such correlations do not have properties appropriate for treatment by psychometric tests. Conforming to such reliability measures would diminish the face validity of the assessment by not allowing it to be formatted around the sequence of problem-solving decisions that we aim to test.

**Discussion and future directions**

This assessment template provides a structure for designing assessments that meet many of the needs described in *Knowing What Students Know* and Schwartz & Arena (National Research Council, 2001; Schwartz & Arena, 2013). Other efforts, largely driven by the K-12 Next Generation Science Standards and the work by OECD, have also pushed for more meaningful assessment at both the K-12 (National Research Council, 2014; OECD, 2020) and undergraduate levels (Laverty et al., 2016; Stowe & Cooper, 2017). While not designed to be related to the Next Generation Science Standards, our assessments have a similar goal of measuring the more meaningful elements of STEM mastery. Our assessments are situated in authentic contexts, they measure complex problem-solving skills based on an empirically defined independent level of mastery, and they provide test-takers with the opportunity to make and justify decisions (‘choices’). These assessment characteristics also align with the elements of ‘authentic assessment’ outlined by Villarroel et al. (2018), for use in undergraduate courses. We add specificity to that work in the form of problem-solving decisions. Assessments based on this template can be used for both formative assessment of individual students and programme/course evaluation.

There are several further directions to this work, in addition to developing assessments for problem-solving across a larger number of disciplines and educational levels. For the assessments we have already developed, the first is conducting more extensive testing and validity studies with larger cohorts of students from multiple institutions. The second is moving to more closed-response questions, such as multiple-choice or ranking to encourage wider use by making scoring easier. As we and others accumulate more test-taker responses from a larger and more varied population, we will be able to identify the extent of variation in the different student answers and hence where closed response questions can provide the same information as the open response versions. For questions involving decisions followed by justifications, we will explore branching options within the test. The justification options provided will depend on their decision as in Walsh et al. (2019).

**Use in instruction**

As noted in the introduction, the best assessment is well aligned with good instruction. Teaching a student to be a skilled problem solver involves them practicing making problem-solving decisions in relevant realistic scenarios, with guiding feedback on how to improve their decision-making. That implies they must also learn during instruction the relevant disciplinary knowledge that is necessary in making such decisions. This type of instruction can be thought of as ‘deliberate practice’ (Ericsson, 2006) to gain problem solving expertise. As discussed in Holmes et al. (2020), the most effective instructional scaffolding is provided by telling the learners what decisions they need to
make, but leaving it to them to make the decision, followed by reflection and feedback on their decision, with opportunities to then revise. Thus, the assessment template, which asks students to make and justify specific decisions, also provides a template for teaching problem solving. The only differences are the conditions under which the student is completing the activity.

During instruction, rather than working individually, students work in small groups answering each question in turn while the instructor monitors their progress. This general format for instruction has been widely used and shown to be effective, with labels like ‘research-based instruction’ or ‘active learning’ (Handelsman et al., 2004; Jones et al., 2015; Wieman, 2019). Here, we only add the explicit focus of the instructional tasks on problem-solving decisions. This teaching approach has now been successfully applied in courses in introductory physics, mechanical engineering design, and clinical reasoning for medical students (Burkholder et al., 2020; Flynn, 2020). This aligns teaching, assessment, and the skills needed in the science and engineering workforce.

**Limitations**

The most important limitation in this work is the lack of comparison between a test-taker’s performance on our assessment and their performance in carrying out problem solving in real life situations. We do have data showing that skilled practitioners score well on these assessments and are quite different from most students, which is better than most assessments. It would be desirable to do longitudinal studies to monitor student progression. Looking at medical students over the course of their training appears to be a particularly good context for such studies. It would also be desirable to do more extensive studies examining the correlation between students’ performance on the assessments and their educational experiences.

**Conclusion**

We have presented a new approach to assessing problem-solving skill across science, engineering, and some fields of medicine. It follows closely the problem-solving approach used by experts in these fields, specifically the set of practices and their associated decisions. The test-takers are given a realistic problem scenario, and then they have to make and justify the decisions an expert problem solver in the discipline would make. We find that such experts make these decisions in a very consistent way, and the student responses are scored according to their degree of overlap with these expert consensus decisions. Such assessments will provide a valuable tool for evaluating the effectiveness of university programmes and courses intended to train engineers and scientists, as well as providing a detailed characterisation of the specific problem-solving strengths and weaknesses of an individual. These assessments also provide a guide for effective teaching of problem solving in science, engineering, and medical disciplines.

**Note**

1. We use the term “discipline” to refer to a defined set of knowledge and skills that experts in a defined area of study use in solving problems and the range of contexts in which they apply
that set. This often, but not always, overlaps with the classification of academic departments and degree programs, but we argue that every “interdisciplinary” program also involves a reasonably well-defined set of skills, knowledge, and problem contexts. As such, they provide a well-defined educational goal which can also be assessed using our approach.

**Disclosure statement**

No potential conflict of interest was reported by the author(s).

**Funding**

This work was supported by Howard Hughes Medical Institute: [Professors grant to CEW].

**ORCID**

Argenta Price http://orcid.org/0000-0002-3692-1683  
Shima Salehi http://orcid.org/0000-0001-6077-3193  
Eric Burkholder http://orcid.org/0000-0001-7420-4290  
Virginia Isava http://orcid.org/0000-0002-3941-1697  
Carl Wieman http://orcid.org/0000-0003-1449-9319

**References**

Adams, W. K., & Wieman, C. E. (2011). Development and validation of instruments to measure learning of expert-like thinking. *International Journal of Science Education, 33*(9), 1289–1312. [https://doi.org/10.1080/09500693.2010.512369](https://doi.org/10.1080/09500693.2010.512369)

Adams, W. K., & Wieman, C. E. (2015). Analyzing the many skills involved in solving complex physics problems. *American Journal of Physics, 83*(5), 459–467. [https://doi.org/10.1119/1.4913923](https://doi.org/10.1119/1.4913923)

Burkholder, E. W., Miles, J. K., Layden, T. J., Wang, K. D., Fritz, A. V., & Wieman, C. E. (2020). Template for teaching and assessment of problem solving in introductory physics. *Physical Review Physics Education Research, 16*(1), 010123. [https://doi.org/10.1103/PhysRevPhysEducRes.16.010123](https://doi.org/10.1103/PhysRevPhysEducRes.16.010123)

Burkholder, E. W., & Wieman, C. (2020). Comparing problem-solving across capstone design courses in chemical engineering. 2020 IEEE Frontiers in Education Conference (FIE), 1–5. [https://doi.org/10.1109/FIE44824.2020.9273820](https://doi.org/10.1109/FIE44824.2020.9273820)

Card, S. K., Moran, T. P., & Newell, A. (1983). *The psychology of human-computer interaction*. Lawrence Erlbaum Associates, Inc.

Chi, M. T. (1978). Knowledge structures and memory development. In R. S. Siegler (Ed.), *Children’s Thinking: What Develops?* (pp. 75–96). Lawrence Erlbaum Associates, Inc.

Chi, M. T., Glaser, R., & Rees, E. (1981). *Expertise in problem solving*. Pittsburgh Univ PA Learning Research and Development Center.

Csapó, B., & Funke, J. (Eds.). (2017). *The nature of problem solving: Using research to inspire 21st century learning*. Educational Research and Innovation, OECD Publishing. [https://doi.org/10.1787/9789264273955-en](https://doi.org/10.1787/9789264273955-en)

Dörner, D., & Funke, J. (2017). Complex problem solving: What it is and what it is not. *Frontiers in Psychology, 8*, 1153. [https://doi.org/10.3389/fpsyg.2017.01153](https://doi.org/10.3389/fpsyg.2017.01153)

Duncan, T. M., & Reimer, J. A. (2019a). *Chemical engineering design and analysis: An introduction. Problem 2-7*. Cambridge University Press. [https://duncan.cbe.cornell.edu/Graphs/Chp2flowsheets/2-7.pdf](https://duncan.cbe.cornell.edu/Graphs/Chp2flowsheets/2-7.pdf)

Duncan, T. M., & Reimer, J. A. (2019b). *Chemical engineering design and analysis: An introduction (2nd ed.).* Cambridge University Press. [https://doi.org/10.1017/9781108377096](https://doi.org/10.1017/9781108377096)
Lenzer, S., Smarsly, B., & Graulich, N. (2020). How do students become experts? An in-depth study on the development of domain-specific awareness in a materials chemistry course. International Journal of Science Education, 42(12), 2032–2054. https://doi.org/10.1080/09500693.2020.1810355

McCloskey, M. (1983). Naive theories of motion. In D. Gentner & A. L. Stevens (Eds.), Mental Models (1st edition, pp. 299–324). Psychology Press.

Mosier, K., Fischer, U., Hoffman, R. R., & Klein, G. (2018). Expert professional judgments and “naturalistic decision making”. In K. A. Ericsson, R. R. Hoffman, A. Kozbelt, & A. M. Williams (Eds.), The cambridge handbook of expertise and expert performance (pp. 453–475). Cambridge University Press.

National Research Council. (2001). Knowing what students know: The science and design of educational assessment. The National Academies Press. https://doi.org/10.17226/10019

National Research Council. (2014). Developing assessments for the next generation science standards. The National Academies Press. https://doi.org/10.17226/18409

OECD. (2020). PISA 2025 Science—Strategic Vision and Direction Paper. https://www.oecd.org/pisa/publications/pisa-2024-assessment-analytical-framework-science-strategic-vision-proposal.htm

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. Educational Psychologist, 51(1), 59–81. https://doi.org/10.1080/00461520.2016.1145550

Price, A. M., Kim, C. J., Burkholder, E. W., Fritz, A. V., & Wieman, C. E. (2021). A detailed characterization of the expert problem-solving process in science and engineering: Guidance for teaching and assessment. CBE—Life Sciences Education, 20(3), ar43. https://doi.org/10.1187/cbe.20-12-0276

Ramalingam, D., Philpot, R., & Mccrae, B. (2017). The PISA 2012 assessment of problem solving. In B. Csapó, & J. Funke (Eds.), The nature of problem solving: Using research to inspire 21st century learning (pp. 75–91). OECD Publishing. https://doi.org/10.1787/9789264273955-7-en.

Randles, C. A., & Overton, T. L. (2015). Expert vs. Novice: Approaches used by chemists when solving open-ended problems. Chemistry Education Research and Practice, 16(4), 811–823. https://doi.org/10.1039/C5RP00114E

Resnick, L. B. (1982). Syntax and semantics in learning to subtract. In T. P. Carpenter, J. M. Moser, & T. A. Romberg (Eds.), Addition and Subtraction: A Cognitive Perspective (1st ed., pp. 136–155). Routledge. https://doi.org/10.4324/9781003046585.

Ryoo, K., & Linn, M. C. (2015). Designing and validating assessments of complex thinking in science. Theory Into Practice, 54(3), 238–254. https://doi.org/10.1080/00405841.2015.1044374

Salehi, S. (2018). Improving problem-solving through reflection. Stanford University.

Schwartz, D. L., & Arena, D. (2013). Measuring what matters most: Choice-based assessments for the digital Age. The MIT Press. https://library.oapen.org/handle/20.500.12657/26061

Schwartz, D. L., Tsang, J. M., & Blair, K. P. (2016). The ABCs of How We learn: 26 scientifically proven approaches, How they work, and when to Use them. W. W. Norton & Company.

Siegler, R. S. (1978). The origins of scientific reasoning. In R. S. Siegler (Ed.), Children’s Thinking: What Develops? (pp. 109–149). Lawrence Erlbaum Associates, Inc.

Simon, H. A. (1973). The structure of ill structured problems. Artificial Intelligence, 4(3–4), 181–201. https://doi.org/10.1016/0004-3702(73)90011-8

Sonner, N B. (2006). BioKIDS: An animated conversation on the development of curricular activity structures for inquiry science. In R. Keith Sawyer (Ed.), Cambridge handbook of the learning sciences (pp. 355–369). Cambridge University Press.

Sternberg, R. J. (1995). Expertise in complex problem solving: A comparison of alternative conceptions. In P. A. Frensch & J. Funke (Eds.), Complex Problem Solving: The European Perspective (pp. 295–321). Psychology Press.

Stowe, R. L., & Cooper, M. M. (2017). Practicing what We preach: Assessing “critical thinking” in organic chemistry. Journal of Chemical Education, 94(12), 1852–1859. https://doi.org/10.1021/acs.jchemed.7b00335
Villarroel, V., Bloxham, S., Bruna, D., Bruna, C., & Herrera-Seda, C. (2018). Authentic assessment: Creating a blueprint for course design. *Assessment & Evaluation in Higher Education, 43*(5), 840–854. https://doi.org/10.1080/02602938.2017.1412396

Walsh, C., Quinn, K. N., Wieman, C., & Holmes, N. G. (2019). Quantifying critical thinking: Development and validation of the physics lab inventory of critical thinking. *Physical Review Physics Education Research, 15*(1), 010135. https://doi.org/10.1103/PhysRevPhysEducRes.15.010135

Wieman, C. E. (2019). Expertise in university teaching & the implications for teaching effectiveness, evaluation & training. *Daedalus, 148*(4), 47–78. https://doi.org/10.1162/daed_a_01760

Winne, P. H., & Azevedo, R. (2014). Metacognition. In R. K. Sawyer (Ed.), *The cambridge handbook of the learning sciences (2nd ed.* (pp. 63–87). Cambridge University Press. https://doi.org/10.1017/CBO9781139519526.006