Handbook of Research on STEM Education

Carla C. Johnson, Margaret J. Mohr-Schroeder, Tamara J. Moore, Lyn D. English

Contemporary Methods of Assessing Integrated STEM Competencies

Publication details
Kerrie A. Douglas, Brian D. Gane, Knut Neumann, James W. Pellegrino
Published online on: 12 May 2020

Accessed on: 04 Sep 2023

PLEASE SCROLL DOWN FOR DOCUMENT

Full terms and conditions of use: https://www.routledgehandbooks.com/legal-notices/terms

This Document PDF may be used for research, teaching and private study purposes. Any substantial or systematic reproductions, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The publisher shall not be liable for an loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
The new vision of STEM education that has emerged from recent reform efforts across the globe envisions students developing a competence in STEM that goes well beyond simply memorizing a wide breadth of factual content. Students are expected to be able to productively engage with key disciplinary content, flexibly apply habits of minds, think critically, develop creative solutions to problems and, most importantly, to be prepared for life-long learning. In essence, competence in STEM refers to the ability to combine knowledge, skills and abilities (KSAs) from one or more disciplines to solve problems across a wide range of authentic contexts (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010; National Research Council [NRC], 2012; Ng, 2008; Pellegrino, 2014). In order to support students in developing such holistic competence, STEM education needs to engage students in STEM practices to explore in depth the relationship between ideas within and across multiple STEM disciplines (NRC, 2014, 2018). Many countries have adopted policies which emphasize students’ need to interact deeply with content from different disciplines, engage in STEM practices, and develop problem-solving and deeper thinking skills (Bernholt, Nentwig, & Neumann, 2012; NGSS Lead States, 2013; OECD, 2017; Roth et al., 2006).

For STEM reform to be successful, all aspects of the educational system (i.e., standards, curriculum, instruction, and assessment) must be aligned to the vision of supporting students in developing integrated competence in STEM (Pellegrino, Chudowsky, & Glaser, 2001). This is particularly important with respect to assessment of such competence, since assessment development and implementation tends to lag behind other aspects of STEM reform and is often controversial in discussions of educational policy and practice (Goldman, Lawless, Pellegrino, Braasch, Manning, & Gomez, 2012).

The purpose of this chapter is to guide researchers and practitioners who are interested in making high-quality inferences about learners’ competence (or competencies) across a variety of STEM domains, including integrating KSAs across multiple STEM domains. Adopting the perspective that detailed and valid inferences about what students know and can do involves a process of evidentiary reasoning, this chapter outlines major issues in STEM assessment design and interpretation, given a variety of contexts of use and intended purposes. In the first section, we discuss various aspects of STEM assessment, including current thinking about the design and validation of assessments of integrated STEM competencies. The second section reviews frameworks for designing STEM assessments. The third section focuses on the types of STEM assessments that might be desired, while the fourth section focuses on the measurement models that can be used with these STEM assessments. This section also discusses the relationship between what is assessed, how it is to be used and what type of measurement approach is most appropriate given the intended interpretive inferences. We
close by looking ahead to the future of assessment and some of the implications for assessment of integrated STEM.

**Assessment in Integrated STEM**

**Role of Assessment in Integrated STEM Reform**

In many ways, the success of integrated STEM reform hinges upon assessments being aligned to the same values and understanding of the reform that are put forward in teacher professional development, curriculum and instruction. Assessments are powerful: they clarify learning expectations for both teachers and students by providing concrete illustrations of the performances expected from students. Assessments are often used as measures of student learning outcomes, teacher effectiveness and school quality. Many decisions of educational and personal consequence are thus informed by student performance on assessments.

One significant concern is that there is a misalignment between the vision of students developing competence in integrated STEM and the traditional vision of STEM assessment, which targets isolated knowledge of STEM facts and procedures (Pellegrino, 2013). For example, several researchers have found that teachers’ common practice is to assign tasks that require students to carry out prescribed activities and memorize content, rather than reason with content and develop deeper understanding (Roth et al., 2006; Weis, Pasley, Smith, & Banilower, & Heck, 2003). While there is a practical need to equip teachers with the capability to assess in ways that are aligned to the reform conceptions of competence (Tekkumru-Kisa, Stein, & Schunn, 2015), teachers are not the only ones struggling to do so for integrated STEM. Recent research found that popular integrated STEM curricular units for lower grades have very few integrated STEM assessments embedded in the curriculum (Douglas, Moore, Merzdorf, Li, & Johnston, 2017). The majority of the assessments were designed to assess memorized content and practices, with very few tasks where students must demonstrate the capacity to combine different KSAs, within or across content domains. In addition, large-scale testing programs have historically been built around multiple-choice items that cover a wide range of topics but with little depth. In response to the adoption of the U.S. Next Generation Science Standards (NGSS), Songer and Ruiz-Primo (2012) argued that assessment must be the new research priority in order to capture the learning targets set out in the standards. Innovative approaches to assessment are needed where students’ reasoning and conceptual understanding are captured and in alignment with the learning objectives set out in the guiding policy documents (Gorin & Mislevy, 2013; Pellegrino, 2013). Next, we discuss design approaches that are aligned to the vision of integrated STEM competence and that should be adopted to generate innovate assessments.

**Design of Assessments Aligned to Integrated STEM Competence: Identifying the Construct(s)**

All assessment design must begin with a clear definition of the construct (or constructs) of interest. Without a clear and precise definition of what is supposed to be assessed, including its scope, there can be no valid inference. In the case of integrated STEM, the construct is *competence* or *integrated competence* in STEM. Ufer and Neumann (2018) discuss in detail the origin and meaning of the term “competence”. For our purposes, it should suffice to say that in the current educational environment, any STEM competency construct must go beyond the recall of isolated, factual knowledge or procedures.

More generally, *competence* in a domain is understood as the integration of all the required KSAs required to solve problems typical to the domain. This may also include problems primarily requiring (deep) knowledge about a particular content or the application of a (complex) procedure.
An alternate, but complementary, way to conceptualize competence in a domain is to consider different dimensions within the domain. This nomenclature is perhaps most prevalent in The Framework for K-12 Science Education (NRC, 2012) but can be also found in other documents for science or STEM education, respectively (for an overview, see Bernholt et al., 2012). The Framework identifies three dimensions that constitute competence in science: knowledge about disciplinary core ideas, knowledge about crosscutting concepts and the abilities and skills required to engage in practices of science and engineering (NRC, 2012). The Finnish Standards for Science Education, for example, similarly identify two dimensions: a knowledge dimension (including declarative, procedural and conceptual knowledge), and the cognitive processes dimensions (including remembering and understanding as well as applying and creating) (Krzywacki, Koistinen, & Lavonen, 2012).

At a broad level, and as discussed in earlier chapters of this book, integrated STEM refers to purposeful learning opportunities which combine aspects of science, technology, engineering and mathematics (Moore & Smith, 2014). Integrated STEM competence therefore refers to the ability to integrate KSAs from across different domains (e.g., mathematics and technology) in order to solve a problem.

**Theoretical Framework for Assessment: Reasoning from Evidence**

Assessment can be understood as a process of reasoning from evidence involving three components that comprise an “assessment triangle”: cognition, observation and interpretation (Pellegrino et al., 2001). The cognition component refers to “a theory or set of beliefs about how students [. . . ] develop competence in a subject domain” (Pellegrino et al., 2001, p. 44). In the assessment of integrated STEM competence, this would be a specific model of: (1) the KSAs constituting competence in the targeted domains, (2) how these develop over time given instruction and (3) the capacity to integrate the KSAs across domains as needed. The observation component specifies the type(s) of assessment tasks likely to yield performance(s) that will provide evidence that allows for drawing conclusions about the targeted competence. Tasks must provide students with the opportunity to demonstrate the integration of selected KSAs to solve problems in STEM. The interpretation component encompasses the procedures employed to reason about students’ competence from the observed performance data. Task scoring rubrics, used to quantify or describe students’ performance, must account for evidence of the degree to which students managed to integrate the KSAs articulated in the cognition component.

The three components of the assessment triangle are inseparably connected. For any assessment to yield reliable and valid information, each component must be consistent with the other two (Pellegrino et al., 2001; NRC, 2014). The Committee on Developing Assessments of Science Proficiency in K–12 identified two major approaches to designing assessments that can help meet this goal: the construct-modeling approach (Wilson, 2005) and the evidence-centered design approach (Mislevy, Steinberg & Almond, 2003). We review each of these approaches next.

**Approaches for Designing STEM Assessments**

Assessments are often designed by simply writing items based on individual interpretations of a set of standards or learning objectives. However, if items are written without a clear specification of the construct to be assessed, including relevant forms of evidence and understanding of how students master the construct, the assessment will often fail to adequately represent the construct and/or provide appropriate and sufficient evidence of the competency. This would limit the reliability and validity of the inferences that can be drawn from students’ performance on the items (for details, see Messick, 1995). In order to ensure adequate coverage and representation of the construct(s) in and across a set of items, as well as reliability and validity of the inferences drawn from students’ performance on those items, Wilson (2005) proposed the construct modeling approach.
The Construct Modeling Approach

The construct modeling approach envisions assessment development as a process of multiple iterations of a cycle through four building blocks: the construct map, the items design, the outcome space and the measurement model (see Figure 19.1). The first building block encompasses the specification of one or more construct maps. A construct map delineates the cognitive construct (e.g., competence in integrated STEM) and its development over time into a hierarchy of qualitatively different levels based on a thorough analysis of the domain given theory, research and empirical data. The highest level defines the maximum level of competence expected from students at a certain stage of schooling, while the lower levels represent intermediate stages through which students typically progress in developing the expected level of competence. The second building block, items design, is about developing assessment tasks that can stimulate responses allowing for observations about the construct or specifying a procedure for the development of such tasks. The third building block, the outcome space, is a description of the possible responses and how they are scored. The last building block, the measurement model, relates the scored responses back to the construct. That is, the delineation of the construct and, more importantly, its specification as a construct map marks the starting point of each cycle and guides the process throughout the cycle: in the beginning, it emphasizes a developmental perspective on the construct; in the end, it provides a point of reference to determine the extent to which students have developed mastery of the construct (Wilson, 2005).

The Evidence-Centered Design Approach

The Evidence-Centered Design (ECD) approach (Mislevy et al., 2003; Mislevy & Haertel, 2006) describes assessment development as a process involving the delineation of three spaces: the claim, evidence and task space (see Figure 19.2) (Pellegrino, DiBello, & Brophy, 2014). The process begins with the delineation of the claim space; that is, specification of the claims that one wants to make about a construct such as students’ competence. This involves unpacking the complexes of KSAs that constitute competence in a domain. Precise formulation of the knowledge students are expected to have and how they are expected to use it is most critical in this step. In the next step, evidence statements are formulated. These evidence statements should clearly specify the features of student performances that will be accepted as evidence that a student has met a claim. Then the types of tasks that are expected to elicit the desired performances need to be specified in terms of essential and optional task features. The entire process—from the delineation of the claims to the specification of the tasks—forms the design part of the assessment argument (Mislevy, 2007).

Figure 19.1 Representation of Wilson’s (2005) Construct Modeling Approach.
Once the tasks are developed, the next step is to define an evidentiary scheme for the tasks. Based on the performances that are expected, one specifies how these performances will be evaluated in the light of the evidence statements and how these evaluations will be combined into evidence supporting (or not supporting) the claims about the construct. A scoring guide that includes a description of all possible performances, precise information on how each performance is scored and how these scores “add up” into a total score ensures that total scores are reflective of students’ competency. This part is the use part of the assessment argument. Together, both the design and use parts of the assessment argument aim to ensure an optimal alignment of claims, evidence and tasks.

Combining Approaches

Both approaches to assessment development, the construct-modeling approach and the evidence-centered design approach, are construct-centered design models that help to establish coherence between the three vertices of the assessment triangle. The construct-modeling approach highlights the developmental aspect of the construct in terms of delineating how students develop competence in STEM into a hierarchy of levels, and the importance of being able to locate students on one of these levels based on their observed performances. The ECD approach highlights the importance of defining as precisely as possible what observations will be considered evidence that students have developed a given level of competence, how these observations will be obtained (by thoroughly specifying task features) and how these observations will be translated into scores reflecting different levels of competence. That is, the construct-modeling approach and the ECD approach emphasize and clarify different aspects of articulation among the vertices of the assessment triangle. The construct-modeling approach emphasizes coherence between the cognition and observation as well as between the cognition and interpretation vertices, whereas the evidence-centered design approach emphasizes coherence between the observation and the interpretation vertex. Both approaches can therefore be profitably combined in order to ensure the development of high-quality assessments of integrated STEM competence.

Unidimensional and Integrated STEM Assessments

In this section we introduce a classification scheme that applies across an array of STEM assessments. Our classification scheme will fit assessment in any of the domains, as the scheme is based on the number and type of dimensions of competency that are targeted by the assessment. Although
instructors may be adopting an integrated STEM curriculum, there may be times when it is appropriate to assess, for example, knowledge of one particular content area. Thus, we review assessments of individual dimensions (of one domain) first, and then turn to assessments that target multiple, integrated dimensions of competency (across multiple domains). For each type of assessment, we review their claims about student competency and the evidence that is marshalled to support those claims.

**Conceptual and Procedural Knowledge of STEM Content**

The first broad class of assessments that we consider are those that focus on conceptual and/or procedural STEM knowledge. In these assessments, one is primarily interested in measuring a single dimension of a STEM domain. Consider an end-of-year course assessment that a secondary school student might take. The implicit claim is that a student has sufficient knowledge of the scientific content presented in a chemistry course. The evidence for (or against) this claim is provided by the student’s performance on the end-of-year course assessment (e.g., a set of selected and constructed response items). Because the desired claimed competency is conceptual knowledge, the evidence must match this claim. The evidence may be composed of the student’s response to fixed-response items, or items that require different levels of reasoning about the content (in order to ensure the item is indeed assessing conceptual knowledge over memorized factual knowledge).

One example of such assessments is that of concept inventories. A concept inventory is an assessment that is focused on measuring conceptual knowledge of a small set of key concepts, while attempting to minimize other requirements such as computation skills (e.g., Streveler et al., 2011; Jorion et al., 2015). The claims in this case are focused on conceptual knowledge and reasoning, and the evidence usually takes the form of the student’s responses to multiple-choice or open-ended prompts. For example, the force concept inventory (FCI) focuses on conceptual reasoning following Newtonian principles and does not require any mathematical computations (Halloun & Hestenes, 1985; Hestenes, Wells, & Swackhamer, 1992). Likewise, the chemistry concept inventory (CCI) focuses on topics learned in a first-year chemistry sequence for post-secondary students (e.g., thermochemistry, bonding, etc.) (Pavelich, Jenkins, Birk, Bauer, & Krause, 2004; Krause, Birk, Bauer, Jenkins, & Pavelich, 2004).

Concept inventories frequently use a multiple-choice format called Distractor Driven Multiple Choice (DDMC) aimed at identifying potentially problematic thinking (e.g., alternative conceptions or misunderstandings). DDMCs have shown some success in identifying when student thinking matches normative scientific thought and/or alternate conceptions that arise from everyday observations and that reflect pre-scientific understandings. Furthermore, some have been designed to show how conceptual reasoning develops over time or experience with the domain (e.g., Herrmann-Abell, DeBoer, 2011; Sadler, 1998). Large test banks (e.g., AAAS 2061; http://assessment.aaas.org/pages/home) and adaptive testing software (e.g., Diagnoser, Thissen-Roe, Hunt, & Minsrell, 2004; http://www.diagnoser.com) have been developed to provide practitioners and researchers with access to these DDMC items. Figure 19.3 shows an example of such a DDMC, where students are reasoning about one key idea in chemistry.

In contrast to a focus on conceptual knowledge and reasoning, other assessments focus on procedural knowledge. Because the claims are different from those claims associated with conceptual knowledge, the evidence must also differ. Using DDMC items will not suffice to reveal students’ procedural knowledge. For instance, if one wants to make the claim that the student has the ability to perform stoichiometric calculations, then part of the evidence should be demonstrations of students’ ability to balance chemical reaction equations. A student might be able to describe the law of conservation of mass but not be able to perform calculations that conform to the law. Likewise, using a student’s performance on problem-solving items as evidence for his or her conceptual knowledge will invariably introduce validation problems. Because stoichiometry primarily requires procedural
knowledge of rules for combining molecules and balancing equations, solving those stoichiometry items might not require one to understand essential conceptual knowledge (e.g., knowledge that in an open system, it might appear that mass is not conserved).

STEM Practices and Literacy

The next broad class of assessments in integrated STEM that we consider are those that measure a student’s ability to engage in a single STEM practice or a student’s STEM literacy (which might involve multiple, related practices). Examples of these competencies include the ability to control variables in planning an experiment or to engage in argument (Osborne, 2010). Examples of these STEM literacies are measures of students’ informational literacy in engineering contexts (Douglas, Fernandez, Purzer, Fosmire, & Van Epps, 2018), or the National Assessment of Educational Progress (NAEP) Technology and Engineering Literacy (TEL) assessment.

In these assessments, one is interested in gathering evidence to support a claim of proficiency about a single construct (e.g., the ability to design an experiment; informational literacy). Because the claim focuses on the single construct, without adding disciplinary conceptual or procedural knowledge, the evidence for the claim must likewise avoid confounding it with disciplinary conceptual or procedural knowledge. As a result, these assessments tend toward minimizing the role of content knowledge in the student’s performance, since doing so is considered a control for construct-irrelevant variance. The evidence to support a claim of this type therefore places constraints on the
design of assessment items because it requires that any disciplinary knowledge required to perform the task be minimized or eliminated.

Consider a prototypical inquiry task that might be used in a chemistry setting in secondary education, one that does not require deep conceptual knowledge of chemistry. This task is called the mystery powders task and has multiple variants, including one implemented in the Principled Assessment Designs for Inquiry (PADI) project (e.g., Mislevy & Haertel, 2006; Seibert, Hamel, Haynie, Mislevy, & Bao, 2006). In the mystery powders task, students are presented with a scenario that can be solved through an inquiry process of investigating, analyzing data and writing an explanation/argument for their conclusions (all broadly conceived of as scientific inquiry practices). The mystery powders task is a “hypothetico-deductive reasoning problem” that requires minimal content knowledge (properties of matter and chemical reactions) (Seibert et al., 2006). Students are presented with a mystery powder that might be a mixture of several powders. Students are to conduct an investigation using standard laboratory procedures to determine which substance(s) make up the white powder. Figure 19.4 shows the first part of the mystery powders task, as implemented by the PADI project. There are multiple investigations that students can simulate (e.g., Heat, Iodine, etc.) and students need to know only about properties of matter (i.e., that there are physical properties that are characteristic of a substance and so can be used to identify the substance) to make sense of the results of some of the investigations. If the task did require deeper knowledge (e.g., the configuration of atoms in a molecule determines the molecule’s properties; an enormous variety of biological, chemical, and physical phenomena can be explained by changes in the arrangement and motion of atoms and molecules; AAAS, 1993), it would cloud the inferences one could make relative to the intended claims. If students lack this deeper knowledge, it might interfere with their ability to demonstrate that they can conduct an investigation and engage in this hypothetical-deductive reasoning.

### Integrated or Multidimensional STEM Competencies

The next broad class of assessments that we consider are those that measure the ability to engage in integrated STEM competencies. In this case, integration means the competency is conceived of multiple dimensions that are assessed simultaneously (Gane, Zaidi, & Pellegrino, 2018). Such an integration could be any combination of conceptual and/or procedural knowledge with STEM practices and/or STEM literacy. In other words, the assessment task must provide students with a chance to demonstrate their knowledge-in-use.

The arguments we make about the claims and types of evidence needed for multidimensional, integrated STEM assessments are similar, no matter which specific dimensions are being integrated. The claim itself is multidimensional (rather than unidimensional), as opposed to the types of claims we have examined in the previous sections (Harris, Krajcik, Pellegrino, & McElhaney, 2016). By saying the claimed competency is multidimensional and integrated, we emphasize that the two (or more) dimensions must be used together by students. The claim is integrated because for a student to engage in the performance, he or she must have KSAs that work together to allow the student to do something with their knowledge. The evidence, like the claim, needs to be multidimensional. This has strong implications for assessment task design, since any task has to be able to elicit multidimensional evidence (Gane et al., 2018; Harris, Krajcik, Pellegrino, & Debarger, 2019).

The example of multidimensional integration that we will focus on is conceptual knowledge integrated with scientific practices. In contrast to the previous section, in which we described scientific practice assessments as attempting to minimize deep conceptual knowledge, the integration does the opposite. It is through engaging in the scientific and engineering practices that one demonstrates deep conceptual knowledge. Consider the example chemistry assessment task in Figure 19.5 (College Board, 2014). This assessment task integrates two dimensions: essential disciplinary knowledge and scientific practices. Note that the learning objectives (Table 19.1) provide the claims that are
Mystery Powders

You have been given a mystery powder consisting of at least one but not more than two powder components. You can determine the components given information from the available experiments.

Current experiment: Taste

Tastes sweet.

Step one: what can you deduce so far? Indicate whether each possible component is In or Out of mixture, or whether that cannot be known yet:

<table>
<thead>
<tr>
<th>In</th>
<th>Out</th>
<th>Can’t tell</th>
<th>Can’t know</th>
</tr>
</thead>
<tbody>
<tr>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
</tbody>
</table>

Step two: which experiment do you want to do next?

Heat Iodine Look Taste Vinegar Water

Figure 19.4 Mystery Powder Task (Seibert et al., 2006).

being made about students, and the assessment task elicits the evidence from students. Each of the two learning objectives are multidimensional, integrating the essential knowledge with a scientific practice. The student has to develop a representation of a homogenous mixture that shows how the H₂O particles and LiCl particles are arranged, including the orientation of both particles. The science practice dimension is integrated with the essential knowledge statement to create a unified learning objective that is a multidimensional claim about students’ competence. The student, in producing the response, must provide the evidence needed to support the multidimensional claim. The performance expectations in the Next Generation Science Standards are examples of three-dimensional claims that integrate disciplinary cores ideas, science and engineering practices and crosscutting concepts (NGSS, 2013).
3. The structures of a water molecule and a crystal of LiCl(s) are represented above. A student prepares a 1.0 \( M \) solution by dissolving 4.2 g of LiCl(s) in enough water to make 100 mL of solution.

(a) In the space provided below, show the interactions of the components of LiCl(aq) by making a drawing that represents the different particles present in the solution. Base the particles in your drawing on the particles shown in the representations above. Include only one formula unit of LiCl and no more than 10 molecules of water. Your drawing must include the following details.

- identity of ions (symbol and charge)
- the arrangement and proper orientation of the particles in the solution

(b) The student passes a direct current through the solution and observes that chlorine gas is produced at the anode. Identify the chemical species produced at the cathode and justify your answer using the information given in the table below.

<table>
<thead>
<tr>
<th>Half-reaction</th>
<th>Standard Reduction Potential at 25°C (V)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{I}^-(aq) + e^- \rightarrow \text{I}(s) )</td>
<td>(-3.65)</td>
</tr>
<tr>
<td>( 2\text{H}_2\text{O(l)} + 2e^- \rightarrow \text{H}_2(g) + 2\text{OH}^-(aq) )</td>
<td>(-0.83)</td>
</tr>
</tbody>
</table>

Figure 19.5 Example Chemistry AP task (College Board, 2014).

Table 19.1 Essential Knowledge, Science Practices, and Learning Objectives Measured by the Example Chemistry AP Task (College Board, 2014)

<table>
<thead>
<tr>
<th>Essential Knowledge</th>
<th>Science Practices</th>
<th>Learning Objectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.A.3 Solutions are homogenous mixtures in which the physical properties are dependent on the concentration of the solute and the strengths of all interactions among the particles of the solutes and solvent.</td>
<td>1.1 The student can create representations and models of natural or man-made phenomena and systems in the domain.</td>
<td>2.8 The student can draw and/or interpret representations of solutions that show the interactions between the solute and solvent.</td>
</tr>
<tr>
<td>3.C.3 Electrochemistry shows the interconversion between chemical and electrical energy in galvanic and electrolytic cells.</td>
<td>5.1 The student can analyze data to identify patterns or relationships.</td>
<td>3.13 The student can analyze data regarding galvanic or electrolytic cells to identify properties of the underlying redox reactions.</td>
</tr>
</tbody>
</table>
Summary

The choice of which dimension (or dimensions) of competency to assess affects the form and substance of the claims that one can make, and therefore one must be careful to consider and specify appropriate evidence for the given claim. In the current STEM education milieu, multidimensional claims are the desired targets, and therefore the type of evidence sought must be multidimensional as well. Just as these claims and evidence are multidimensional, the measurement models used to interpret the multidimensional evidence must also be able to model this multidimensional data.

Validation and Reporting: Measurement Models for STEM Assessment

Once assessment developers have clearly articulated the particular assessment purpose and intended inferences, and followed a principled approach to task design, the next step in elaborating a validity argument for the given assessment is to collect empirical data of various types (e.g., Jorion et al., 2015; Pellegrino, DiBello, & Goldman, 2016). The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council of Measurement in Education, 2014) specify a variety of forms of evidence that can be used to establish the validity of an assessment given its intended interpretive use (see also Pellegrino et al., 2016). One of the most important forms of evidence regarding how the tasks function, individually and collectively, is actual student performance data. Making sense of those data relative to what was hypothesized and the intended reporting process requires application of various qualitative and quantitative interpretive procedures. The application of psychometric measurement models serves as a core element of the STEM assessment development, validation and reporting process (Wilson, 2005). Such models define the relationship between student scores on individual tasks and the construct, providing the basis for a meaningful interpretation of the scores across all tasks. More specifically, measurement models aggregate student scores across tasks in such a way that they become interpretable in terms of the construct. The choice of the measurement model defines what one can infer about the construct (i.e., competence in integrated STEM). As a consequence, one must carefully consider the intended interpretation and use of the resulting scores before choosing a measurement model. The most commonly used measurement models for integrated STEM assessment are Item Response Theory models and Diagnostic Classification Models.

Item Response Theory

Item Response Theory (IRT) locates students and assessment tasks on the same latent trait (i.e., construct) continuum. Doing so allows for deriving conclusions about: (a) the likely performance of particular students on particular tasks; (b) the progression of a student in developing competence; and (c) the distribution of a sample of students in terms of their progression in developing competence in a domain (e.g., STEM). This is under the assumption that the tasks can be arranged in a way that reflects increasing competence in the domain. IRT also provides a means to examine the extent to which a set of assessment tasks are representative of a developmental perspective about the construct (see the earlier section “Theoretical Framework for Assessment: Reasoning from Evidence”). As a consequence, IRT is often used in the process of developing and validating STEM assessments to examine the extent to which the tasks utilized define a single latent continuum reflecting development of the STEM competency in question (e.g., Jorion et al., 2015; Neumann, Schecker, & Theyßen, 2019), as well as for judging student variability with respect to those STEM competencies (e.g., Hadenfeldt, Neumann, Bernholt, Liu, & Parchmann, 2016; Harwell et al., 2015).

Item Response Theory links students’ performance on a set of assessment tasks to their competence through a probabilistic model. This model describes the probability of a student performing
Successfully on a task as a function of the students’ ability (e.g., competence in STEM) and item characteristics (e.g., task difficulty). Four typical IRT models include: (a) the Rasch Model, which, along with (b) the one-parameter logistic (1PL), includes one parameter for each item representing its difficulty; (c) the two-parameter logistic (2PL) model, which includes an additional parameter representing the item’s discrimination (i.e., how well the item differentiates among students of differing ability levels); and (d) the three-parameter logistic (3PL) model, which adds a parameter to account for successful performance on forced-choice items due to the opportunity to guess the correct answer (Baker, 2001). IRT models highlight the role of the items in the assessment process. They acknowledge that although items often naturally differ in their difficulty, this variability is actually crucial for assessing different levels of competency among students (Wilson, 2005).

Both the Rasch and the 1PL models arrange persons and items (tasks) on a common continuum, with a “person ability space” on one side and a corresponding “task difficulty space” on the other side. The graphical representation mapping the distribution of students based on their ability onto the distribution of tasks based on their difficulty is commonly known as a Wright Map, an example of which is shown in Figure 19.6. The continuum is such that a person at a given point (e.g., just above 0 on the logit scale) is estimated to have a 50 percent chance of correctly answering a task also located at that given point (e.g., question 5). Tasks that range on the easier side of the continuum (lower down on the map) would be solved correctly by that same person with an increasingly higher probability, and tasks on the more difficult side (further up the map) would be solved correctly by that same person with an increasingly lower probability.

The Rasch model, like the 1PL model, differs from other IRT models in that it assumes that the items all have the same discrimination ability (i.e. all items contribute to the student ability continuum in the same way) and only differ in terms of their difficulty (i.e. their position on the latent continuum). In some cases, however, items or groups of items may differ in the extent to which they contribute to differentiating among students. In this case, we can use the 2PL IRT model, which is statistically equivalent to a confirmatory factor analysis (Wirth & Edwards, 2007). It is commonly used in order to identify items or item groups that exhibit discrimination or factor loadings that differ from the other items. The third model, the 3PL model, is simply a 2PL model with an additional term to adjust the probability of solving an item correctly. This additional term is also called the guessing parameter, as it allows for accounting for the possibility to guess the right answer in forced-choice items. The main use of the 3PL model is to account for the guessing probability or to examine the extent to which items were subject to guessing (e.g., Jorion et al., 2015).

Sometimes if whole groups of items have a similar discrimination that differs from the discrimination of the other items, this indicates that the item groups are assessing different constructs. For such cases, IRT offers multidimensional versions of the described models. These models provide one ability estimate for each dimension, plus an estimation of the correlations between the dimensions—in addition to item difficulty parameters for each item relative to each dimension. This is essentially the same as a series of independent one-dimensional models; however, the multidimensional models provide unbiased estimates of the correlations and student ability estimates because they are accounting for the correlations. Therefore, a multidimensional model is recommended over the use of multiple one-dimensional models (Wu, Adams, & Wilson, 2007).

It is important to note, however, that there are two fundamentally different types of multidimensionality. The first type, which is also known as between-item dimensionality, refers to assessments in which we have different sets of tasks measuring different constructs. For example, one set of tasks requires students to solve problems in chemistry, and one set of tasks requires them to solve mathematical problems. Sometimes, however, we are in a situation where individual tasks are measuring different aspects of a construct. This is the case, for example, when students are required to integrate their knowledge about mathematics and physics in order to solve an engineering problem, such as the egg challenge. Knowledge about mathematics, physics and engineering problem-solving skills
would all be dimensions of their competence that students might draw on in such an assessment of their competence in engineering. This is also known as *within-item dimensionality*. Interestingly, students do not necessarily have to be competent with respect to each dimension in order to perform successfully on the described problem. It is possible that students exclusively draw on their problem-solving skills to design a solution (e.g., by trial and error), or that they draw on their physics

---

**Figure 19.6**  Example of a Wright Map showing the distributions of student (left side) and item (right side) performance on a concept inventory. The scale is in logit units.
knowledge about the forces acting on an egg upon hitting the ground and mathematics knowledge in how these forces will be spread based on the shape of the egg. That is, students’ abilities in the different dimensions can to some extent compensate for each other. Such multidimensionality and the respective models are termed compensatory models. The multidimensional versions of Rasch, the 2PL and the 3PL model are all of this type.

**Diagnostic Classification Modeling**

In some assessments of students’ competence, the key dimensions are non-compensatory. Imagine, for example, a task that requires students to develop and use a model of the food chain in a given ecosystem. Obviously, in order to successfully perform this task, students need to integrate their modeling skills with their knowledge about the constituents of the given ecosystem. Neither outstanding modeling skills nor extensive knowledge about the constituents of this ecosystem alone will help students solve the task. This situation requires a different approach to modeling student performance, known as Diagnostic Classification Modeling (DCM). This approach includes a family of models (for an overview, see Rupp & Templin, 2008). Amongst other uses, DCMs have been used to investigate students’ responses to an assessment based on a cognitive model that delineates the specific KSAs required to solve each task of a larger task set where the tasks call upon different but overlapping assemblages. Models used in this way have also been referred to as cognitive diagnosis models or cognitive diagnostic models (Rupp & Templin, 2008).

As discussed earlier, integrated STEM competence is characterized by the capacity to integrate different knowledge, skills and abilities to identify and solve problems typical for the domain across a wide(r) range of contexts (Weinert, 2001). In assessment, we are trying to assess if or to what extent students have these KSAs and can bring them together to solve problems in the domain. Sometimes we are interested in which KSAs students do and do not possess; DCMs allow one to answer this question.

Diagnostic classification models are related to IRT models but differ in important assumptions about what governs student success on a task. In IRT models of the type discussed in the previous section, one assumes that performance on a task is predicted through a continuous latent ability (i.e., a single construct). In DCMs, students’ performance is predicted by one or more dichotomous latent variables (e.g., the presence or non-presence of one or more KSAs). DCMs allow for a direct interpretation of the results in terms of the KSAs that an individual student possesses. The grain size at which these KSAs are assessed is determined (a priori) by the assessment developer and can be a key part of the assessment design process as described earlier. The heart of a DCM is the so-called Q-matrix which specifies which of the range of n (elements of) knowledge, skills or abilities are required for each of the m items in the assessment. The information carried by the Q-matrix is often also referred to as the loading structure to highlight the similarity between DCM and (confirmatory) factor analyses (CFAs). In DCMs, each KSA is assumed to be required for successful task performance, and therefore, if a student does not possess a specific KSA required by the task, he or she should not perform well on the task. Thus, DCMs are often considered non-compensatory measurement models (Rupp & Templin, 2008).

One use of DCMs is in the analysis of performance on concept inventories (e.g., Jorion et al., 2015; Kuo, Chen, & de la Torre, 2018). Jorion and colleagues (2015) highlight the potential of DCMs to examine individual students’ conceptual profiles; that is, which of the concepts tested by concept inventory students have mastered and which ones they have not. DCMs could also be utilized to detect if students hold a particular misconception or not (e.g., Bradshaw & Templin, 2014). Most existing DCMs can detect if students have mastered or not mastered a particular concept, or if students hold a particular misconception, but not both. However, an incorrect response to an item may be due to a student lacking mastery of a concept, possessing some misconceptions, or both.
(Kuo et al., 2018, p. 180). To address this issue, Kuo et al. (2018) recently proposed a DCM that can be used to simultaneously assess mastery of a concept (or any kind of skill, for that matter) and the existence of a misconception.

In summary, while the primary purpose of IRT models is to provide a single score representing students’ ability on a latent continuum, the primary purpose of a DCM approach is a more fine-grained diagnostic analysis (Rupp & Templin, 2008). As discussed in the following, the interpretive use context for applying a DCM to assessment data differs from the context of applying IRT.

STEM Assessment Along a Continuum From Classrooms to Large-Scale Interpretive Uses

Assessments can serve a range of different purposes that fall along a continuum from a fine-grained, criterion-based diagnosis to a single-value, norm-based evaluation of student learning. Teachers assess learners’ competence to understand where their students are in the learning process, and to plan the next step accordingly. However, teachers also use assessments in order to grade student learning, so that the learners and others will know the extent to which learners have developed competence in STEM. Sometimes assessments are designed to assess the extent to which learners have the competencies to be sufficiently prepared for future learning; for example, when universities administer entry or placement tests in order to understand whether learners bring the prerequisites to successfully complete studies in medicine (Hissbach, Klusmann, & Hampe, 2011). Finally, assessment results are used to evaluate the performance of schools, districts or state or national education systems (e.g. Organisation for Economic Co-operation and Development [OECD], 2007). The one end of this continuum is marked by assessments most proximal to instruction and use in the classroom assessments. The other end of the continuum is marked by assessments rather distal to instruction, like (large-scale) assessments for accountability or normative comparison purposes (Ruiz-Primo, Shavelson, Hamilton, & Klein, 2001). Both come with their very own specific characteristics, and we review each in turn.

Classroom-Based Assessments

Assessments proximal to instruction typically aim to provide formative information to the teacher; that is, students’ current competence and potential next steps in the instructional process. Classroom assessments are designed or selected by teachers and given as a part of instruction (NRC, 2014). They are given during or closely following an instructional activity or unit. Through classroom-based assessments the teacher aims to obtain information about students’ learning. This information can and should be derived from multiple sources, including student verbal contributions, artifacts that students produce during learning or quizzes that the teacher administers at the end of a lesson or shorter instructional unit. Classroom assessments can be designed primarily to guide instruction (formative purposes) or to support decisions made beyond the classroom (summative purposes). Assessments for formative purposes occur during the unit, assessments for summative purposes are administered at the end of the unit. However, ideally even assessments for summative purposes such as end-of-year tests can also serve formative purposes in that they provide students and teachers with feedback about what KSAs students have mastered and what KSAs students are still struggling with and should continue to work on (or the whole class should work on if there are KSAs the whole class has failed to master). In order to provide this kind of detailed feedback to students and teachers, fine-grained information is needed beyond a single overall measure of the extent to which students have developed competence in STEM or a STEM field.

As discussed in the previous section, assessment designed to provide detailed information about students’ competence in STEM at the level of the mastery of individual KSAs are well suited to
Assessing Integrated STEM Competencies

the application of Cognitive Diagnostic Models. This is particularly important in the assessment of integrated STEM competence, where students are typically presented different tasks each requiring a range of KSAs. A total score across all tasks (obtained, for example, through IRT analysis) would provide only general information on where students reside on the latent continuum delineating students’ development of competence or integrated competence in STEM. Information on specific ideas or skills students are struggling with would be hidden. For example, a student may be somewhat proficient in constructing explanations but lack understanding of the ideas of energy dissipation and conservation. This student will likely fail to use this idea to construct an explanation of why things stop when they are not continuously supplied with energy. CMDs can be used to provide this level of detailed information. However, it is important to note that in classroom-based assessments, we can often draw only on information from a small number of tasks (sometimes only one task, if we are counting learning activities as a task). Drawing inferences from a small number of tasks requires maximum caution in terms of the conclusions that can be drawn from students’ performance on these tasks. In our example here, a student may have understood energy dissipation and conservation but have misread the task or may simply not have been motivated that day to provide an elaborate response. That is, ideally, even or better: especially in terms of more fine-grained formative assessment aiming to obtain evidence about students’ mastery of a range of specific KSAs, we need a range of tasks or, more generally, a range of observations, ideally from different sources of evidence (i.e. different types of assessment). The challenge here is to integrate information from different sources. However, DCMs can also help with this. DCMs can, for example, be utilized to analyze performance on a series of different learning activities or, in fact, any kind of task administered to students, as long as these tasks can be categorized in terms of the requirement that successful performance on the task requires the KSA in question.

In summary, designing a diagnostic assessment and interpreting it via DCM is the ideal approach when the aim is to obtain detailed information about the mastery or non-mastery of specific KSAs, as is often the case in classroom-based assessments. In addition to individual student diagnostic information, a DCM-based approach can also be useful for learning about different response strategies of different groups of respondents for the same set of tasks, about different response strategies of single individuals across different tasks and about different response strategies of single individuals within the same task, if data can be collected that provide sufficient information about these cases (Rupp & Templin, 2008, p. 236; see also Pellegrino et al., 2001). The most important aspect of classroom-based assessment is to draw on multiple sources of evidence for reliable and valid inferences, since classroom-based assessments usually involve small numbers of tasks.

Large-Scale Assessments

Large-scale assessments are designed or selected by districts, states or nations to assess student achievement for various purposes including evaluation of the quality of schools and of state or national education systems. They typically test students at a particular age or grade on broad educational goals such as knowledge-in-use in a domain-like science (e.g. NGSS Lead States, 2013). The tasks used in these assessments cover a broad range of different topics across the domain (e.g., high school biology). The question is to which extent students have developed the envisioned competence in the domain. The challenge is to achieve sufficient reliability and validity in the assessment of such a more broadly defined construct given the intended interpretive use. Often large-scale assessments aim to assess the extent to which students are prepared for future learning (e.g., success in higher education). An example is the College Board’s AP Chemistry Test, where scores on that test are used by colleges to judge whether students are ready to take courses beyond typical introductory college chemistry. Given such interpretive uses, not only must the assessment represent the construct domain, it also should have high reliability and minimal measurement error so that we are (a) distinguishing well
between students and (b) are sufficiently confident that students who meet some scoring benchmark indeed have the knowledge-in-use envisioned as required for the judgement in question (i.e., access to or placement in a higher education program). Ideally, in such testing contexts we are less interested in what specific knowledge students have and more interested in whether students have met a certain benchmark (e.g., solved 60 percent of the questions; Hissbach et al., 2011) or how they compare to other students (e.g., when we are comparing different educational programs). In essence, we aim to obtain a single measure for each student (a) that is a function of students' competence in a domain, and (b) that reliably differentiates between students.

Given such interpretive goals, IRT models are commonly used for large-scale assessments. Such models allow test construction and administration such that different students answer different sets of items drawn from a large pool of domain-relevant tasks that have been previously calibrated in terms of difficulty relative to student ability (for an international overview over large-scale assessments of standards, see Bernholt et al., 2012). In such cases the focus is more on the extent to which a large group of students have met a particular standard than on obtaining information about how well an individual student is doing. If, for example, the Rasch Model has been demonstrated to apply to a population of students and a universe of items, the model will also apply to any sample of students or items (Bond & Fox, 2007). As a consequence, it does not make a difference which students have been administered which items as long as there is a sufficient overlap between items across sets and students having worked on the same item sets. Thus, even though students have been administered different tasks, we obtain estimates for students' ability (e.g., competence) that are all located on the same scale and can thus be easily compared. Assessments like the NAEP and PISA science assessments use such test design, analysis and reporting procedures. In addition, IRT models allow for including additional information in or through the so-called background model. A common method in large-scale assessments is to include descriptive analyses about what students abilities may depend on (see, for example, Rost, Senkbeil, Walter, Carstensen, & Prenzel, 2005). Sometimes in large-scale assessments we are interested in more differentiated analyses such as students’ ability in terms of different areas or aspects of the domain. For example, in the assessment of students’ competence in experimentation, we may be interested in the extent to which students can plan, carry out or analyze experiments. In this case, we can apply IRT multidimensional models, which will yield information about student competence in different yet related areas of the construct (Neumann et al., 2019).

In summary, large-scale assessments are designed to support inferences about the extent to which students have met benchmarks with respect to pre-defined educational standards that usually focus on one content domain. They typically aim to do so for large samples or entire populations of students, but sometimes they also try to reliably identify whether single students have met certain benchmarks (i.e., in entry, placement or licensure testing). IRT-based approaches provide ways to accommodate requirements of large-scale assessment such as covering a broad range of content or reliably determining students’ level of competence relative to other students assessed. Through standard setting procedures, it is also possible to describe the performance continuum in terms of levels of proficiency, which allow for identifying the level of proficiency a student or group of students has achieved in terms of their progression in developing competence in the STEM disciplines.

**Conclusion**

In STEM education there are many different types of assessments that can be designed for different interpretive purposes and contexts of use. No one approach to assessment can evaluate everything that is important in achieving integrated STEM competence. What is common across all types and approaches is that assessment involves a process of reasoning from evidence that must follow a principled approach to design and interpretation to obtain valid and reliable inferences about competencies that are not directly observable. We have highlighted how the purpose of an assessment, the
intended inferences from scores, the design of assessment tasks and the validation and psychometric measurement approaches must all be in alignment. In closing we would also note that innovations in technology are enabling the design of new types of assessments where students can virtually perform complex tasks and provide complex forms of evidence about why they might think the way they do, what they know and how they are learning (Gane et al., 2018). When applied to the assessment of integrated STEM competencies, there are numerous possibilities for innovation at both the classroom and large-scale interpretive use contexts.

In addition to the design and validation of specific assessments, coherent assessment systems will need to be created in order to meet the information needs of the various stakeholders in STEM education. Such systems operate across levels from the classroom to schools, districts, states and nations. Pellegrino (2014) suggests five features of a coordinated assessment system: (a) emphasis on assessment of higher-order skills more than on rote memorization; (b) high-fidelity assessment of critical abilities; (c) standards that are internationally benchmarked; (d) assessments that are instructionally sensitive and educationally valuable; and (e) assessments that show strong evidence of being valid, reliable and fair. It is our hope that the principles discussed in this chapter can be applied to inform the creation of such coordinated assessment systems for integrated STEM.

There has been a long-standing divide between the STEM teaching and learning community and the educational assessment community regarding what we assess, how we assess and what the results mean relative to gauging the progress of teaching and learning (Pellegrino, 2017). In this chapter, we have argued that in order for the new vision of integrated STEM education to be successful, all parts of the education system must align with the same goals of the reform. Considering the strong role that assessment plays in classrooms around the world, there is an opportunity for assessment and integrated STEM learning experts to work together to design assessments that help translate this vision for reform into real, authentic learning opportunities for students.

Notes
1. We define practices in the same manner as the Framework for K-12 Science Education (NRC, 2012). A practice refers to authentic “doing” in the discipline, i.e., applying deep disciplinary knowledge by engaging in the same practices that disciplinary experts engage in. Examples of practices include modeling in science, design in engineering, and reasoning abstractly and quantitively in mathematics.
2. In the case of a multidimensional construct, multiple construct maps need to be specified, one for each dimension (NRC, 2014).
3. “Deep” knowledge is relative to the experience of the learner. What is deep disciplinary knowledge for a student who is taking a chemistry class in high school might not be for a chemistry undergraduate student.
4. In the egg challenge, one has to design a device that prevents an egg that is being dropped from a given height from bursting.
5. The most prominent difference being that in factor analysis models, the latent variables are continuous.

References


Assessing Integrated STEM Competencies


