Wright, L., Harkins, H., Kopriva, R., Auty, W., Malkin, L., & Myers, B. (2022). Technology-enhanced tasks to assess three-dimensional science sense-making: Possibilities and lessons learned from the ONPAR NGSS-based classroom assessment project. *Contemporary Issues in Technology and Teacher Education, 22*(2), 382-410.

Technology-Enhanced Tasks to Assess Three-Dimensional Science Sense-Making: Possibilities and Lessons Learned from the ONPAR NGSS-Based Classroom Assessment Project

Laura Wright, Heather Harkins, and Rebecca Kopriva
University of Wisconsin-Madison

William Auty
Education Measurement Consulting

Linda Malkin and Blake Myers
University of Wisconsin-Madison

The use of technology in assessment continues to evolve the field of educational measurement. This article reports on development and use of new accessible, technology-enhanced assessments designed to measure the three-dimensional science abilities of middle school students. The assessments were piloted with over 70 teachers and 8,000 students throughout the United States over a 3-year period. The adoption and implementation of technology-enhanced assessments is potentially challenging for educators, and numerous factors can influence whether new tools are successful in classroom contexts. The authors describe the assessments alongside insights from project surveys into the conditions that supported or hindered teachers' successful implementation and use of the new assessments in classroom settings. Results indicate that teachers found the assessments useful for supporting the transition to instruction based on Next Generation Science Standards and preparing students for new state science tests. Successful uptake of the materials in the classroom was supported by professional learning that anticipated teachers' content, technology, and pedagogical needs. While the assessments were overall successful, areas for potential improvement are also described, including improved reporting formats that are more teacher and student friendly.

In the late 1990s, Randy Bennett, the director of the National Assessment of Educational Progress Technology–Based Assessment Project, envisioned three phases to the reinvention of assessment through technology (Bennett, 1998). First, infrastructure building would be necessary to enable widespread use of computer-based assessments. Second, he foresaw that technology would support a transformation of question-and-response formats and scoring sophistication, enabling the field to move beyond multiple-choice items and dichotomous scoring. Third, he envisioned assessment being rooted in cognitive science, serving individual and institutional purposes, and allowing teachers and students to utilize assessment feedback to enhance learning.

In the decade following Bennett's prediction, assessment did not characteristically change (Tucker, 2009a). However, starting in 2010, the U.S. Department of Education supported several assessment consortia focused on developing large-scale summative online assessments that aimed to measure student mastery of new standards. Smarter Balanced, the Partnership for Assessment of Readiness for College and Careers, WIDA, English Language Proficiency Assessment for the 21st Century, and Dynamic Learning Maps transformed large-scale assessments from paper and pencil to online administration, making delivery more streamlined. Phase 1 of Bennett's transformation was realized, with administration of most large-scale assessments for K-12 students taking place via computer.

Another vision for technology-enhanced assessment has been to provide greater accessibility to students with disabilities and English learners (Almond et al., 2010). In the past, these students have received traditional tests with accommodations added on. In one study, researchers found that over 75 different assessment accommodation strategies were used for English learners, including dictionaries, glossaries, extra time, and test translations (Rivera & Collum, 2006). However, not all accommodations were appropriate to English learners' needs, and their use showed mixed results (Kieffer et al., 2009).

Rather than adding accommodations after an assessment has already been developed, researchers have envisioned that technology-enhanced assessments could be designed with accessibility principles at the outset to include the greatest number of students (Thurlow et al., 2006). For example, technology could offer the opportunity to embed tools within the assessment platform to reduce the need for post hoc accommodations or provide scaffolding to support student understanding (Almond et al., 2010). Research on accessibility may, in fact, dovetail with Phase 2 of innovation, reenvisioning how item stimuli is presented to students.

Numerous researchers have written about the potential of technology for Phase 2 innovation, asserting that technology in assessment has the potential to afford new opportunities that were not possible with paper and pencil tests, notably, the ability to better assess the construct of interest as well as skills and reasoning abilities (Alonzo & Ke, 2014; Gane et al., 2018; Pellegrino & Quellmalz, 2010).

Proponents argue that technology offers the chance to change what is observed in the assessment context and how it is observed, because technology-enhanced assessments have the capability of delivering novel

stimuli and gathering unique responses that are not possible in traditional formats (Gane et al., 2018; Kopriva & Wright, 2017; Tucker, 2009b). Moreover, technology-enhanced assessments provide the potential to gather evidence of student learning behaviors as well as interpret them. For example, technology-enhanced assessments can contain animations and graphics to present information to students dynamically and offer novel response types that enable students to draw, model, and carry out investigations. Computer algorithms can be generated to interpret these behaviors automatically, easing the burden of scoring for educators with less subjectivity.

Thus, technology may afford educators the opportunity to assess practices and skills that are better matched to the types of reasoning and response process that are of interest (Gorin & Mislevy, 2013). This feature is especially useful under the *Next Generation Science Standards* (NGSS, NGSS Lead States, 2013), which call for students developing three-dimensional science abilities — understanding of disciplinary core ideas and mastery of science and engineering practices and crosscutting concepts — as well as assessing these dimensions in an integrated way.

Assessing students' three-dimensional abilities has been one of the challenges facing science education since the inception of NGSS (Alonzo & Ke, 2016; Pellegrino, 2012; 2013; Pellegrino et al., 2014; Songer & Ruiz Primo, 2012). Although science education has been at the forefront of exploring how to present and interpret complex questions in assessment environments (Pellegrino & Quellmalz, 2010), reviews of pre-NGSS science assessments indicate that most high stakes science tests were unidimensional, focused on disciplinary core ideas. Further, most test formats were limited to multiple choice questions, which limited abilities to make inferences about other dimensions of NGSS (Sawchuk, 2019). Technology holds a great deal of promise for assessing students' three-dimensional abilities, because it provides a context for students to use and apply their reasoning skills in innovative ways.

To meet the challenge of NGSS assessment, several groups have embarked upon researching and designing technology-enhanced assessment materials intended to be used in classroom contexts. One such project is a collaboration involving the BEAR Center and Stanford University, which created online assessments for middle and early high school focused on two science topics, the physical behavior of matter and ecology, along with the practice of argumentation. Technology is utilized to administer the assessments through the Berkeley Assessment System Software; items are text based, and the tasks are hand-scored according to rubrics. Example tasks and scoring may be found at http://scientificargumentation.stanford.edu.

Another group, the Next Generation Science Assessment Collaborative, has created science assessment tasks, rubrics, and accompanying instructional resources for Grades 3-5 and Grades 6-8, available at https://ngss-assessment.portal.concord.org/. A variety of science content areas are covered by the assessments. Technology enables the use of models, videos, data analysis tools, as well as other tools that allow students to demonstrate understanding (Damelin & McIntyre, 2021). Upon completion of the project, the collaborative anticipates that 200

tasks will be available, along with scoring rubrics for teachers to use in classroom contexts.

A third project, High-Adventure Science, has created six classroom activities and related item sets pertaining to cutting-edge Earth Science topics alongside the practice of argumentation. Starting around 2014, the project began to explore the use of technology to automatically score student-generated arguments for two of their new high school units. The assessments provide both individual student and classroom-level feedback to students and teachers. Example materials can be found at http://has.concord.org/index.html#interactives.

The ONPAR project, the focus of this article, has also created technology enhanced assessment tasks for middle school science classroom use. The project leverages technology to design challenging, accessible assessments. The assessments are appropriate for students who struggle with text heavy assessments such as English language learners and students with disabilities in reading, as well as mainstream students (Kopriva et al., 2021). Assessments also utilize technology to offer automatic scoring and reporting. The project has completed 12 units of science materials covering life, physical and earth sciences, and a total of 75 assessment tasks. Examples of the ONPAR approach to assessment are available on the project website at http://iiassessment.wceruw.org/projects/.

These four projects range in their utilization of technology — from administering assessments via computers in online formats, presenting information and providing interactive online tools, and utilizing accessibility resources to scoring responses and providing reports automatically. These approaches illustrate the ways NGSS assessments are leveraging technology for the purposes of measurement and the various ways that educators will need to become adept in technology use in classroom contexts. Teachers will have to become comfortable administering assessments online, supporting student interaction with innovative item types, and interpreting and using assessment results that are automatically scored, rather than scored by themselves.

Even though technology-enhanced assessments hold a great deal of promise for improving measurement of student knowledge and skills, adoption and implementation of new technology is challenging for educators (Koehler & Mishra, 2009), especially at a time when implementation of new standards also requires substantial instructional shifts from educators (Alonzo & Ke, 2014; Reiser, 2013). If educational measurement is to realize Bennett's third phase of reinvention and inform teaching and learning, classroom contexts and the way assessments are used by educators need to be researched and understood.

Koehler and Mishra noted the importance of understanding the affordances and constraints of new technologies and the ways they influence teacher behaviors. Their framework, technology, pedagogy, and content knowledge (TPACK), describes how teachers combine pedagogical content knowledge (Shulman, 1987) with their understanding of educational technologies. Successful use of technologies in classrooms, they claimed, requires that teachers develop a knowledge base consisting of content, pedagogical methods, and technology. Further, technology use

is particular to each content domain and should be influenced by the pedagogical practices specific to each discipline (see Bull et al., 2019). Developing fluency with these three domains allows teachers to have a deep and flexible understanding of teaching with technology which, in turn, helps them utilize technology to advance student learning.

Koehler and Mishra (2009) noted that external factors such as time, teacher beliefs about pedagogy (see also Ertmer, 2005), and access to training may influence their success with technology adoption and implementation. They recommended that professional development be designed with these factors in mind.

Similarly, Gane et al. (2018) recommended that teachers receive support to understand and utilize technology-enhanced assessments, noting that educative supports have the potential to increase teachers' success. However, in-service teachers have indicated that they need more professional learning on classroom assessment (DeLuca & Klinger, 2009; Klinger et al., 2012), and many commercial classroom assessment packages do not provide a robust program of professional learning. Thus, when adopting a classroom assessment, especially one that is technology enhanced, it is important to consider the amount of training and support teachers need to use it successfully so that it can be fully leveraged for instructional aims.

The project described in this article researched and developed assessments with accessibility principles from the outset, aimed at all three levels of Bennett's vision for technology-enhanced testing. Assessments were fully delivered online, utilized innovative item types and scoring, and sought to provide instructionally useful assessment data. The ONPAR project researched and developed innovative multisemiotic (Kress & van Leeuwen, 2001) science assessment tasks for middle school science classroom use, utilizing visuals, action, sound, and language to communicate to and from students in the assessment environment.

Using a variety of communicative methods reflects the varied ways students learn and reason in science classrooms and addresses access needs of students who may struggle with the language load of traditional tests such as English language learners (ELLs; Kopriva, 2008; Logan-Terry & Wright, 2010). The project developed 12 units of science materials, including 75 assessment tasks to assess students in middle school science classroom contexts. When the assessments were piloted, project participants also committed to participating in a series of professional learning meetings to learn about the assessment targets, how to implement the assessments, and how to utilize scoring and reporting information for instructional purposes.

Because the assessments were novel in all aspects, including the NGSS focus, the types of items, as well as the automatic scoring and reporting, the project sought to investigate the conditions that supported or hindered teachers' successful implementation and use of ONPAR classroom assessments through a survey. In the sections that follow, we provide an overview of the project and review results from project surveys to describe the overall success of implementation, as well as the factors that supported and hindered its success.

Description of Assessment Approach

The ONPAR assessment methodology is a unique multisemiotic approach that uses a wide range of representations both to present assessment items and to open up response types (http://iiassessment.wceruw.org/projects/). The theoretical underpinnings and empirical support for the assessment methodology come from the fields of semiotics (Jewitt, 2008; Kress, 2003, 2010; Kress & van Leeuwen, 2001), cognitive science (Gee, 2007; Graf & Kinshuk, 2008; Myers, 2015; Pellegrino et al., 2001) and Evidence Centered Design (ECD; Kane, 2013; Mislevy, 2009; Mislevy, 2013).

The assessment approach has been developed and researched through a series of federally funded grants and has demonstrated success in addressing the linguistic and cultural barriers encountered by low-English proficient ELLs on assessments for large-scale, summative purposes, such as annual state accountability measures (Kopriva et al., 2016; Kopriva et al., 2021). Further, research has shown that ONPAR items can successfully measure challenging science concepts and skills of ELLs' using novel computer-interactive techniques that largely redirect the language comprehension and production loads to multisemiotic representations (Kopriva et al., 2016; Kopriva & Wright, 2017). The current project sought to extend this line of research, applying the assessment methodology to develop assessments intended for use in middle school science classrooms.

To develop the assessments for the most recent project, a systematic approach was undertaken by project staff, using ECD as a starting point. Development of each ONPAR science unit began with identification of the NGSS Performance Expectations (PEs) on which assessment tasks and items would be based. Once the PE for a specific task was identified, it was then *unpacked* (Harris et al., 2016) to fully understand the depth and scope of what demonstrable student abilities were expected. In unpacking, task developers focused on the *connections* between the dimensions. That is to say, rather focusing on DCIs, SEPs, or CCCs on their own, task developers considered the relationships between at least two dimensions at a time (i.e., DCI & CCC, DCI & SEP, and SEP & CCC).

Understanding the standards this way supports the eventual development of multidimensional assessment items that incorporate at least two of the NGSS's three dimensions. By disentangling the connections between the dimensions while maintaining a vision of them in the whole of the PE, ONPAR domain analyses focused on the components of the PE that cue assessment task contexts, screen content, and interactive elements.

In addition to NGSS PEs, task designers used the current understanding of learning progressions in science education (Alonzo & Elby, 2019; National Research Council [NRC], 2012) to determine assessment goals and evidence. Learning progressions describe how students develop successively more sophisticated ways of reasoning about science content as they obtain more experience with phenomena and representations and improve their cognitive abilities (Smith et al., 2006). Because NGSS for middle school encompasses the entire Grade 6 – 8 band, task designers used learning progression endpoints for Grade 5 and Grade 8 (NRC, 2012)

to bracket the conceptual understanding assessed on screens, in tasks, and across assessment units.

In addition to unpacking the NGSS, the assessment approach also considered test takers from the outset of development. The ONPAR approach is unique in its emphasis on the *person dimension* described by Kopriva et al. (2016) and Kopriva and Wright (2017). These researchers argued that assessments must be designed in ways to allow the widest range of test takers to demonstrate their abilities. Consequently, the *experience* of individuals taking the test factor significantly into all stages of the task development.

At the item series level, this means identifying contexts that would be familiar or accessible to most learners. At the screen level, this translates into a set of design practices that provide users with multiple ways to understand and respond to questions. ONPAR screen design practices include the following (Kopriva & Wright, 2017; Wright, 2013):

- Communicating multimodally; including static and dynamic imagery and text.
- Providing on-demand audio of questions in English and Spanish. (Other languages have been translated in past projects; in the current project, the majority of ELLs were Spanish-speaking.)
- Structuring necessary language with ELLs and struggling readers in mind.
- Supporting vocabulary *not* being assessed with graphics and animations.
- Providing text-based screen descriptions for those who prefer to read.

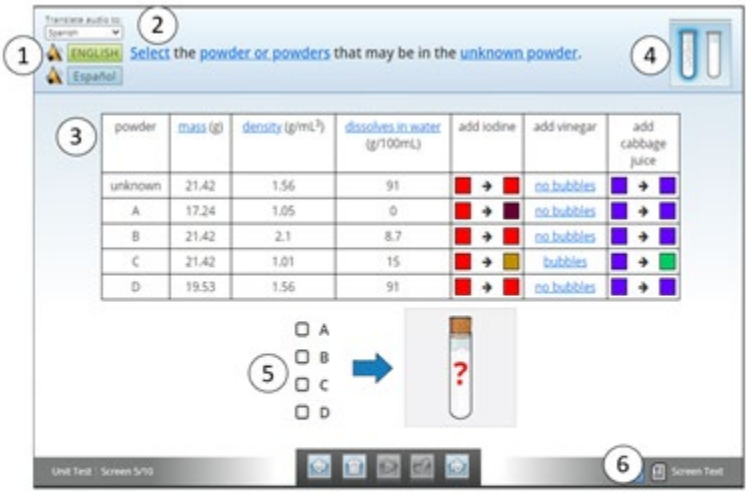
When considering the person dimension at this stage in assessment development, task designers considered how assessment content could be made accessible within the assessment environment to a range of diverse learners. Figure 1 shows a sample screen shot of an ONPAR item from a chemistry task and its accessibility features.

All ONPAR tasks address scientific phenomena; context screens are used to set the scene for the entire task and help to activate a student's schema. They also mark shifts between item series and provide students with the necessary background information to respond to subsequent screens. Item series typically include opportunities for students to demonstrate understanding of key content (DCIs); employ science and engineering practices through the application of key DCIs within the context; and apply the focal crosscutting concept(s) (CCCs) to the problem.

Items were organized to elicit evidence of student understanding or misunderstanding on screens as well as within and across item series. This structure provided an opportunity to collect evidence using multiple means of representation and added to the reliability of inferences drawn from answer patterns across screens. Typically, assessment tasks included two to three item series, for a total of six to 10 items. Additionally, items in ONPAR tasks spiraled in sophistication; introductory questions focused on simpler content and questions at the end of a task focused on more complex content and reasoning. Video 1 shows an item series from an

ONPAR chemistry task focused on changes in states of matter. The video shows a context screen and two related subsequent items that required students to model changes in states of matter as well as explain their model.

Figure 1
ONPAR Multisemiotic Item and Accessibility Features



1. Read aloud available in English and select languages.
2. Simplified English sentences present the task demand.
3. Graphics provide focal information.
4. Words and phrases hyperlinked to supportive graphics; example pop up of bubbles image.
5. Reduced language load on response options.
6. Screen text provides information in textual form for those who prefer to read.

Video 1
Particle Nature of Matter Item Series 2

https://youtu.be/OA8KKN1j3XI

Scoring of tasks was developed alongside the items to ensure that they were targeting key NGSS dimensions. ONPAR tasks have two levels of scoring, item level scoring and task level scoring, which are output into individual and classroom reports. Item level scoring occurred at the screen level (one item on a screen) and resulted in numeric scores that varied based on the complexity of the question and the degree of understanding implied by a test takers' response.

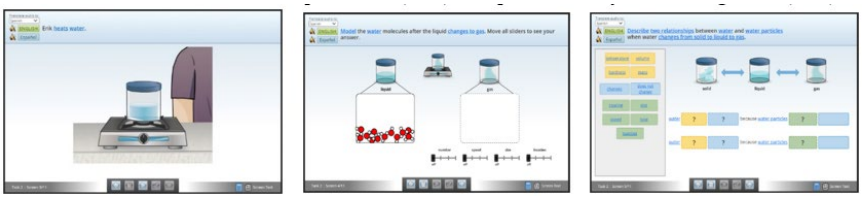
For example, a screen that targeted student understanding of conservation of matter could be scored on a scale from 0-3. A score of 0 indicated that the test taker did not demonstrate understanding, whereas a score of 3 indicated that the test taker had provided evidence of full understanding. Scores of 1 or 2 provided evidence of limited or developing understanding. Each screen had its own unique scoring rubric created by a task developer to reflect the latest research in learning progressions and was vetted by project psychometricians.

Task level scoring consolidated information for the same NGSS dimension across various screens of the assessment task. Thus, items measuring the same dimension (DCI, SEP, or CCC) were tagged with codes that were used to track specific answer patterns across screens. These codes were triggered by individual screen actions identified by task designers as meaningful. Examples include whether certain elements were included in a model or if specific phrases were used within a larger explanation.

The codes enabled scoring rules to be derived from natural language into logic and computer algorithms, allowing for automatic scoring of complex student behaviors. Task level scoring rules resulted in a set of diagnostic statements organized by NGSS dimension. Reports automatically provided information on how students performed on DCI, SEP, and CCC separately.

An example from an ONPAR task follows to illustrate how the novel items and scoring are realized in practice. The example is from a Grade 6 chemistry task, which is comprised of three item series and 11 screens. The measurement goals are to assess students' understanding of what happens to particles when thermal energy is added or removed and between different states of matter (solid, liquid, gas), as well as assess students' ability to apply cause and effect (CCC) in different contexts, construct explanations (SEP), and plan and carry out investigations (SEP).

Figure 2
ONPAR Item Series (Context Screen, Modeling Item, and Explanation Item)



The example is from the second item series, which begins with a short animation illustrating the grounding phenomenona for subsequent response screens. The first item asks students to model water particles in two states of matter. Students manipulate sliders to adjust the number, speed, size and location of particles. As they change the position of the sliders, they see changes in the particles in the model.

The next item prompts students to describe the changes in the model using a statement frame. The statement frame provides a scaffold for creating the language-based description. Color-coded answers are provided on the left side of the screen to drag into the statement frame on the right side of the screen.

The first item is worth 4 points and the second item is worth 3 points. The total points possible on each screen is an indicator of item complexity. Partial credit is based on unique scoring rules determined for each response screen. For example, on the modeling item, users earn 1 point for

each properly modeled variable. On the explanation item, users earn 3 points for two correct statements; 2 points for one correct statement and one partially correct statement; and 1 point for one partially correct statement. Answers that do not meet these criteria do not earn points.

In addition to the numeric scores, codes track answer patterns within and across items and determine the diagnostic statements that appear on the automated score reports. On the example items, answer patterns related to particle number, size, and speed are tracked and combined with answer patterns on other items that assess similar content. If enough codes are triggered, statements such as those below appear on individual reports:

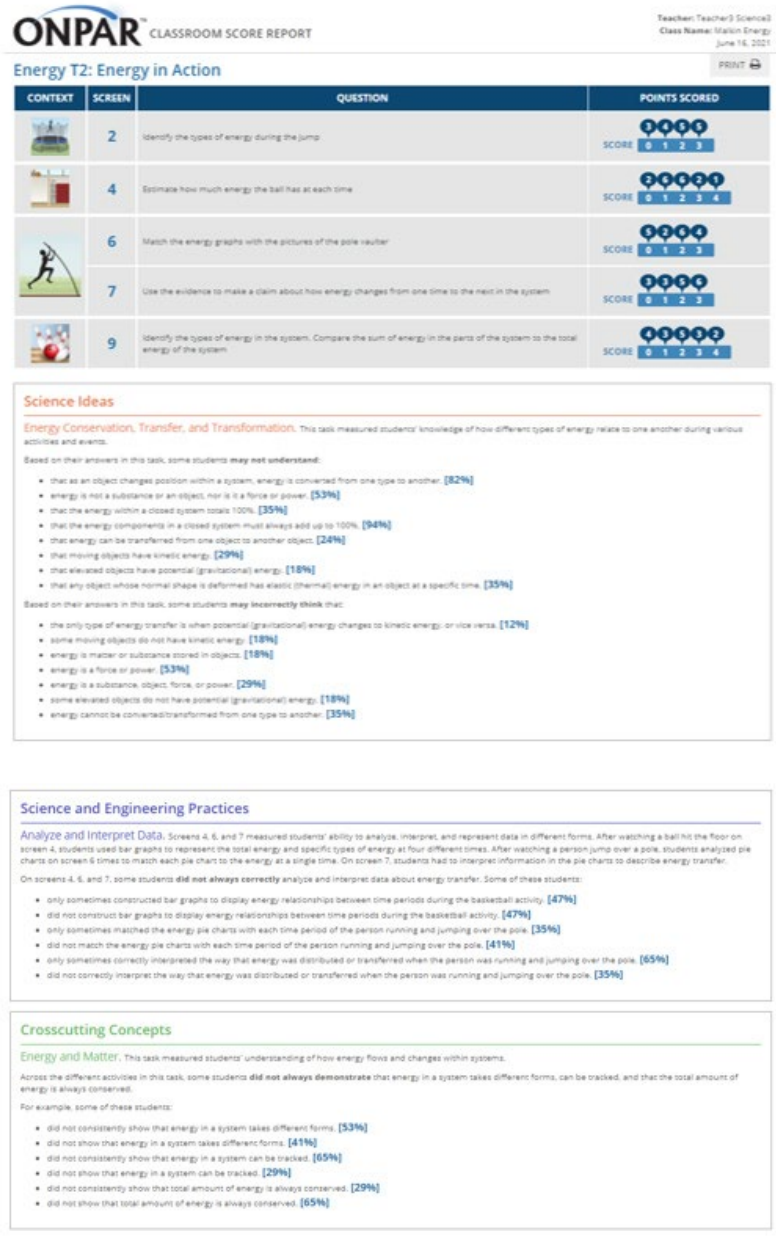
> Based on your answer, you may not understand the relationship between:
>
> - temperature change *and* changes to particles.
>
> - the mass of a substance *and* particles within a substance.

The algorithm used to trigger the statement about temperature looks for specific codes generated on Screens 4, 5, 10, and 11. The algorithm used to trigger the statement about mass looks for specific codes generated on Screens 2, 4, 5, and 11. Each diagnostic statement has its own unique algorithm. If no statements are triggered, a message describing the test taker's inferred understanding appears, such as "On the screens in this task, you correctly described how the movement and spacing of atoms and molecules affects temperature and the state of matter."

Upon completion of an ONPAR task, students and teachers received automatically generated score reports. Classroom reports were available to teachers in the online portal and provided them with an overview of how all students performed in their class. Individual reports were available to students and teachers in the online portal and provided detailed information on individual student performance.

The top portion of reports focused on numeric scoring and indicated the number of points awarded by screen. The lower portion of reports focused on diagnostic reporting of the three NGSS dimensions. Statements in this portion of the report indicated aspects of the dimensions that students may need to work on to develop fuller understanding or ability. Classroom reports were interactive so that teachers could click through to read information about specific students and view individual reports for different students in their class. Figure 3 illustrates an ONPAR classroom score report for a task assessing student understanding of energy, energy conservation and transformation.

Figure 3
ONPAR Classroom Score Report for Energy Task 2



ONPAR CLASSROOM SCORE REPORT

Teacher: Teacher3 Science3
Class Name: Malibi Energy
June 16, 2021

Energy T2: Energy in Action

PRINT

CONTEXT	SCREEN	QUESTION	POINTS SCORED
	2	Identify the types of energy during the jump	SCORE 0 1 2 3
	4	Estimate how much energy the ball has at each time	SCORE 0 1 2 3 4
	6	Match the energy graphs with the pictures of the pole vaulter	SCORE 0 1 2 3
	7	Use the evidence to make a claim about how energy changes from one time to the next in the system	SCORE 0 1 2 3
	9	Identify the types of energy in the system. Compare the sum of energy in the parts of the system to the total energy of the system	SCORE 0 1 2 3 4

Science Ideas

Energy Conservation, Transfer, and Transformation. This task measured students' knowledge of how different types of energy relate to one another during various activities and events.

Based on their answers in this task, some students **may not understand**:

- that as an object changes position within a system, energy is conversed from one type to another. [82%]
- energy is not a substance or an object, nor is it a force or power. [53%]
- that the energy within a closed system totals 100%. [35%]
- that the energy components in a closed system must always add up to 100%. [94%]
- that energy can be transferred from one object to another object. [24%]
- that moving objects have kinetic energy. [29%]
- that elevated objects have potential (gravitational) energy. [18%]
- that any object whose normal shape is deformed has elastic (thermal) energy in an object at a specific time. [35%]

Based on their answers in this task, some students **may incorrectly think** that:

- the only type of energy transfer is when potential (gravitational) energy changes to kinetic energy, or vice versa. [12%]
- some moving objects do not have kinetic energy. [18%]
- energy is matter or substance stored in objects. [18%]
- energy is a force or power. [53%]
- energy is a substance, object, force, or power. [29%]
- some elevated objects do not have potential (gravitational) energy. [18%]
- energy cannot be converted/transformed from one type to another. [35%]

Science and Engineering Practices

Analyze and Interpret Data. Screens 4, 6, and 7 measured students' ability to analyze, interpret, and represent data in different forms. After watching a ball hit the floor on screen 4, students used bar graphs to represent the total energy and specific types of energy at four different times. After watching a person jump over a pole, students analyzed pie charts on screen 6 times to match each pie chart to the energy at a single time. On screen 7, students had to interpret information in the pie charts to describe energy transfer.

On screens 4, 6, and 7, some students **did not always correctly** analyze and interpret data about energy transfer. Some of these students:

- only sometimes constructed bar graphs to display energy relationships between time periods during the basketball activity. [47%]
- did not construct bar graphs to display energy relationships between time periods during the basketball activity. [47%]
- only sometimes matched the energy pie charts with each time period of the person running and jumping over the pole. [35%]
- did not match the energy pie charts with each time period of the person running and jumping over the pole. [41%]
- only sometimes correctly interpreted the way that energy was distributed or transferred when the person was running and jumping over the pole. [65%]
- did not correctly interpret the way that energy was distributed or transferred when the person was running and jumping over the pole. [35%]

Crosscutting Concepts

Energy and Matter. This task measured students' understanding of how energy flows and changes within systems.

Across the different activities in this task, some students **did not always demonstrate** that energy in a system takes different forms, can be tracked, and that the total amount of energy is always conserved.

For example, some of these students:

- did not consistently show that energy in a system takes different forms. [53%]
- did not show that energy in a system takes different forms. [41%]
- did not consistently show that energy in a system can be tracked. [65%]
- did not show that energy in a system can be tracked. [29%]
- did not consistently show that total amount of energy is always conserved. [29%]
- did not show that total amount of energy is always conserved. [65%]

The ONPAR project aimed to fully leverage the affordances of technology to design and program the assessment materials. Technology enabled novel presentation of content by way of graphics and animations; information was conveyed to students dynamically rather than through text-heavy means. Technology also offered students multiple points of accessibility. Onscreen help such as miniature graphics and animations, translation, and read-alouds provided students numerous opportunities for on-demand support. The technology-enhanced environment also allowed the project to open response types so that students could model,

design experiments, graph, and create statements to express understanding and reasoning abilities in innovative ways. Opening response types enabled researchers to observe different behaviors that better matched the constructs of interest.

Finally, computer algorithms allowed project staff to create automatic scoring and reporting so that teachers and students received immediate feedback. This action afforded teachers and students opportunities to address learning needs in a timely fashion. Importantly, technology was harnessed to increase the types of questions students were asked, expanding what was asked in the assessment environment as well as how it was asked.

While the materials were designed to be as user friendly as possible, knowing the multiple challenges science teachers were facing in using new content standards as well as a new technological tool, the project undertook a pilot that focused not only on the psychometric properties of the assessment tasks, but also the teacher component of the implementation process. In the sections that follow is information on how the tasks were implemented in classrooms, as well as results from the survey given to investigate the factors that affected the overall success of the implementation process.

Methods

Participants

ONPAR pilots took place from fall 2017 through spring 2020. The project sought school districts in states that had adopted NGSS and were implementing the new standards. Once interested sites were identified, the project staff applied for research approval in the districts, and sought teachers who wished to volunteer for the project. All teacher participants taught science in Grades 6, 7, or 8.

Assessments were piloted with 71 middle school science teachers and approximately 8,000 students throughout the United States. Sites included public and private schools in urban, suburban, and rural districts. Other than volunteer status and middle school science teacher status, there were no other requirements for the participants in the study, such as years of teaching or teaching background. Project staff did note that because teachers were volunteers, many could be considered early adopters, who expressed that they liked technology and were not particularly anxious about using it in their classroom.

Procedures

To prepare for using the assessment materials, teachers were asked to attend three out-of-school, unit specific, professional learning sessions with ONPAR project staff via a web-conference application. Meetings were scheduled to correspond to when teachers were teaching the focal units with their students. Each virtual session was designed around a single assessment task to maintain focus on specific content, practices, and concepts assessed.

The first meeting in each unit provided an overview of key ideas and skills assessed across the entire unit, links to standards and learning progressions, and screen-by-screen discussions of the assessment goals and strategies in the first task. Teachers were asked to interact with screens as students and provide feedback during the meeting and were frequently asked to predict how their students would respond to the same screen. This approach resulted in lively discussions between teachers and researchers and prompted teachers to reflect on their own curriculum and instruction.

In some cases, a discussion of correct answers clarified science ideas for teachers and provided them with opportunities to think about how they might incorporate those ideas into their teaching. Asking teachers to interact with the screens also provided them with the needed hands-on experience to coach their students confidently on interacting with novel assessment items. The second and third meetings also centered around screen-by-screen task discussions; however, instead of starting with an overview of standards, teachers were asked to discuss relative success of implementing prior tasks with students and assessment results. These discussions provided ample opportunity for teachers to reflect on their teaching, student performance, and better supporting student growth. Teachers were also asked to bring sample student and classroom reports to the second and third online meetings so that project staff could support teachers' abilities to read and interpret the reports and think about how to use the results to inform instruction.

Online meetings lasted approximately 60 minutes each, and teachers were compensated for their out-of-school meeting and planning time for the project. Thus, for one ONPAR unit, a teacher typically spent approximately 3 hours of time in online professional learning meetings.

Once they were initially trained, teachers administered the assessment tasks to their intact middle school science classrooms. Administration of the three ONPAR assessments usually took place over a span of 4-6 weeks during a typical unit of instruction. For example, during a life science unit on the topic of ecology, teachers would teach their own NGSS-focused lesson plans to develop student understanding of concepts such as food webs. Once they felt that students were ready, they administered an assessment task on the same topic to gauge student understanding and ability with the focal content. Teachers had control over teaching prior to use of the materials as well as time of administration and pacing of the assessments.

Teachers were asked to administer two 30-minute extended assessment tasks and one 50-minute end-of-unit test during the relevant science unit, spaced appropriately within their instruction. Thus, teachers would cover material for Task 1 and then administer the task, cover the material for Task 2 and then administer the task, and give the end-of-unit test at the completion of a unit. Additional materials in the project portal were available to support teachers' use of the assessments such as short training videos that reviewed the measurement goals of each task and demonstrated novel item types and a teacher guide that provided information on measurement goals, discussion questions, and extension ideas for instruction. Tasks were also available for students to retake, complete with scoring and a second comparative report.

To prepare students to use the assessment tasks, teachers instructed students to watch a short tutorial video demonstrating the ONPAR technology-enhanced items. Students were also allowed to use practice items on the project website.

After students had watched the video and practiced with an ONPAR item, students took one of the assessment tasks. Some teachers also previewed how to interact with the assessment tasks with their students prior to administering by projecting tasks on a whiteboard and describing them but did not give students clues about answers. Teachers were asked to ensure that students individually answered tasks for the purposes of the study but were encouraged to discuss answers and use tasks creatively for instructional purposes after students had completed them.

Instruments

After teachers had completed a unit and administered all tasks, they also answered a survey about their experience with the materials for that specific unit. The survey contained approximately 18 statements that offered participants a 4-point Likert scale response: *strongly agree, agree, disagree* and *strongly disagree*. Statements were grouped into four themes that provided insight into the classroom conditions that supported or hindered their implementation of the ONPAR assessments: overall satisfaction of teachers and students, usefulness of on-demand teacher resources and reporting information, alignment and usefulness of tasks for teachers, and usefulness of tasks for students (as perceived by teachers).

Four open-ended questions were also included after each thematic section to offer teachers an opportunity to clarify responses or provide additional interpretive information. After teachers reported they had administered the last task of the unit, project staff ensured that the assessment data had been received in the project data server and then distributed an electronic link to an online survey so that teachers could answer the project survey. Teachers were told that their project stipend would be distributed once their survey responses had been received. The response rate to the survey was 100%, likely due to the incentive tied completion of project activities. A copy of the survey statements and responses are included in the appendix.

Cronbach's Alpha was used to examine the internal consistency of the survey responses administered across all years of the study and was calculated at 0.9. Additionally, Likert scale responses were quantified to identify the percentage of participants who responded to each of the four levels of agreement. Results were also examined by year to explore whether there was a change in survey results over time. No statistically significant change was found in teachers' results by year of the study, so overall results were examined across all years of the pilot.

For this analysis, project staff aggregated the overall percentage of agreement (*strongly agree* and *agree*) and disagreement (*disagree* and *strongly disagree*) to identify which statements had the strongest agreement and which had the strongest disagreement. The statements that had the strongest agreement and disagreement for each thematic section

of the survey were identified and explored more fully in the open-ended comments so we could better understand the conditions that supported and hindered successful implementation of the technology-based assessments in middle school science classroom contexts.

The results section provides an overview of survey responses with the highest rates of agreement and disagreement for each thematic section of the survey, as well as illustrative comments that help provide insights into the overall trends. The high reliability of survey results may be due to teachers meeting multiple times together in professional learning meetings and sharing experiences and perspectives on the assessments.

Results

Overall Teacher and Student Satisfaction

The first thematic section of the survey investigated teacher and student satisfaction. Overwhelmingly, teachers responded favorably to the piloting experience, with 97% strongly agreeing and agreeing that they had a positive experience; 3% disagreed that they had a positive experience, all of whom were in the first year of the pilot. Teachers commented that ONPAR was a useful tool that aligned with their curriculum and allowed for multiple representations of student understanding. Note that all teachers were volunteers in this study, which may have contributed to a high teacher satisfaction rate. In more typical implementation conditions, there could be a lower rate of teacher satisfaction.

To a lesser extent, 86% of teachers strongly agreed and agreed that students had a positive experience with the ONPAR pilot, while 13% disagreed that the pilot experience was positive for students. Teachers' open-ended comments indicated a nuanced explanation for the lower rating. Teachers noticed that students struggled with numerous aspects of the assessments, including adjusting to the new types of items and meeting a higher level of challenge in keeping with the NGSS. However, teachers did not see these struggles as entirely negative. Some teachers noted that they were a productive part of the learning process and could spur them on as a teacher. One teacher explained,

> I disagreed that students had a positive experience using ONPAR materials because this is what they would say, NOT because I think it was a non-beneficial experience for them. Many of the students became frustrated with the difficulty level of some of the screens, and then were more discouraged after seeing the negative feedback on their score reports. I realize that this is a reflection of their deficiencies in experiences thinking three-dimensionally in science — and therefore needs to be enhanced by me as the teacher — but I thought it should be noted.

Thus, teachers may have been choosing a satisfaction rate that reflected actual student comments about the new assessment. Despite this fact, even when rating student satisfaction lower than their own, teachers found the experience of using ONPAR beneficial for students as learners and considered it an opportunity for growth. Section 4 of the survey results explores perceptions of student usefulness in greater depth.

Ease of Use of On-Demand Teacher Resources and Reporting

The second section of the survey dealt with the array of on-demand resources offered to teachers in the online portal. Teachers had access to training videos, teacher guides, and classroom and individual reporting information. Overall, teachers agreed that these resources were useful. The highest rated survey item related to the task guides offered in the portal; 92% of teachers strongly agreed and agreed that the task guides were useful for planning to use the assessment tasks. Task guides were reported to be less useful for planning next steps, with only 86% of teachers strongly agreeing and agreeing that the task guides helped identify next steps for instruction.

Compared to the on-demand supports offered in the portal, ONPAR professional learning opportunities were highly useful and supportive of teachers' experience with the materials. Of all statements on the survey, the online training with ONPAR staff received the highest rate of agreement; 99% of teachers strongly agreed and agreed that the meetings with project staff helped them implement the digital tasks with their students, 98% of teachers agreed that the meetings helped them integrate the unit into their instructional plan, and 97% of teachers agreed that the meetings helped them understand how to use the score reports to inform instruction.

The high ratings suggest that the professional learning plan surrounding the new assessment was one of the keys to its overall success. Positive comments from teachers included the following:

- "The video meetings prior to usage made the application with students so much easier."
- "The meetings with the ONPAR staff always help me to understand the material."
- "The ONPAR staff is knowledgeable about student need, best practices and how to advise teachers in both areas. They know the tasks and were patient during training and implementation."
- "The online meetings were a wonderful way to share and prepare for the lessons."
- "The project meetings were instructional, and the project managers were nice and answered all my questions. They also worked with me on scheduling meetings. They were very flexible."

These comments reflect that the training addressed elements described in Koehler and Mishra's (2009) TPACK framework. Teachers had the opportunity to receive training on how the assessment dealt with NGSS content, how to utilize the technology for pedagogical purposes, as well as how to use the technology itself. Teachers also noted that time was a factor and that the project's flexible approach to arranging training to meet their schedules was appreciated.

Survey results from this thematic section indicated that the weakest ONPAR materials for teacher usability were the classroom score reports; 13% of teachers disagreed and strongly disagreed that the reports helped them identify student learning needs, and 11% of teachers disagreed or strongly disagreed that the reports were easy to understand.

While the overall agreement rate was still high, the relatively high disagreement rate suggests that improvements could be made to reporting. This result is not surprising as it was the first time the project had designed reports for use with ONPAR assessments. Previous projects focused on item and task design but had not investigated reporting formats. Teacher comments helped to clarify that the information provided in the reports was rich and useful, but the format of the reports was not user-friendly. One teacher noted, "Because the score reports are so comprehensive, I wouldn't say that they are easy to understand and interpret, but they were useful."

Several teachers suggested displaying reporting information in a spreadsheet or roster so that it would be easier to read. One teacher commented,

> Need a way to see all the student's scores in a class in one go. Teachers do not have time to click on each individual student to see scores. Seeing them all together helps us (teachers) look for trends in the scores for certain students and adjust our teaching accordingly.

Such comments were not limited to teacher experience with classroom reports. Teachers also found the individual reports "time-consuming and cumbersome to go through." One teacher noted,

The format in which the student reports were generated was not helpful. Having to go to each student individually to gain feedback on their score was not helpful. I have 160 students. It would take me hours to go through all of that data.

Importantly, teacher comments regarding scoring and reporting highlight their desire to use data and information to drive their NGSS instruction. However, the format of the scoring and reporting information reduced their ability to use data and should be reenvisioned so it is more user-friendly. While the project provided useful data and rich information to teachers, it was not packaged in a way that could be readily taken up and used by them. Unfortunately, due to budget and time constraints associated with a grant, the project was not able to undertake revisions of the reports to reflect teachers' suggestions.

Alignment and Usefulness for Teachers

The third thematic section of the survey investigated how easy it was for teachers to use ONPAR given their external circumstances, such as school and district climate, alignment with curriculum, felt need, and time constraints. Overall, teachers indicated that the materials fit with their district and school goals; 94% of teachers strongly agreed and agreed that the materials were consistent with their school climate and reforms occurring in their districts and schools.

Many of the districts and schools were in states that had adopted the NGSS, yet materials to support implementation and assessment of NGSS were lagging. Many teachers commented that ONPAR was helpful as they started to shift to the new standards; 92% of teachers strongly agreed and

agreed that the materials filled a need. Many teachers noted that ONPAR was potentially useful for preparing students for the state science test, as in the following quotation:

> I definitely think these [ONPAR] assessments align with the new state standards better. After seeing some sample questions from the 8th grade end of course test, these assessments will give my students a better sample of what is expected.

According to this thematic section of the survey, the weakest area was related to the amount of time needed to interpret scoring and reporting information. This result is consistent with comments teachers made about reporting with regard to the on-demand materials. Twenty-two percent of teachers disagreed and strongly disagreed that setting aside time to interpret the classroom score report was easy. Results were similar for the amount of time teachers needed to interpret classroom score reports to plan differentiated instruction. Because validity of assessments not only considered the assessment instrument itself but also interpretation and use, moving forward, the project staff should consider whether teachers are able to readily use the information provided by the assessment.

Perceived Usefulness for Students

The final thematic section of the survey investigated teachers' perceptions of student experience. Teachers' perception of overall student satisfaction was lower than their own satisfaction due to a variety of factors. Many teachers felt that students struggled to understand how to answer the new types of questions in the assessment tasks; 52% disagreed and strongly disagreed that students understood how to answer new item types in Year 1.

Because of this response in Year 1, the project staff expanded the survey statements to try to identify which item types were particularly difficult for students in Years 2-4. The most problematic item type was modeling with "sliders." This item type required that students play with the sliders, moving a bar to the right, to see how aspects of the model changed when the slider position changed. Other item types, such as statement frames and data interpretation, were less problematic for students.

While responses in Year 1 indicated that new item types were a particularly problematic aspects for teachers, the project was more successful in subsequent years due to specific actions taken by project staff. First, a directions video was added to the student portal to provide tailored information on the ONPAR item types. Second, project staff suggested to teachers during training meetings that they walk students through tasks if they felt that doing so would provide a better user experience for students. Teachers were encouraged to orient students to the assessment context and how to interact with items without prompting or giving answers. Some teachers took a few minutes prior to the assessment to show the tasks screen by screen on a whiteboard. This step provided an orientation for students and helped them understand how to interact with the different items.

Finally, through repeated usage many teachers became more familiar with the item types and felt more confident in their own understanding of how to use the technology-enhanced items and support students on their own. Their confidence may have led students to have greater confidence as well.

Two other aspects of student experience were also rated relatively low, the overall level of challenge of the assessment and student understanding of score reports. Teachers commented that the overall challenge of ONPAR tasks was higher than what their students were accustomed to; 18% disagreed that the level of challenge was appropriate. This result may be partially due to introduction of the NGSS, which themselves were more rigorous. We anticipate that as teachers become more accustomed to the NGSS and students have more opportunities to engage in NGSS at lower grade levels, the level of challenge on the assessments will be more appropriate for middle school students.

Regarding scoring and reporting, 43% disagreed and strongly disagreed that students understood their individual score report. Teachers noted that students often focused on the numeric scores of reports rather than the reporting information provided. This finding may be due to the overall reading level of the reports. One teacher noted,

> Although I tried to have my students view their score reports, the majority of students did not look at them. The students DID look at their overall scores and note improvements from earlier tasks to the final but were not reading the actual feedback that went with the report.

Some teachers commented that students could benefit from additional support on understanding how to read and interpret their individual score reports. While some teachers shared in meetings that they used the reports as a conferencing tool with students, this practice was not common. In the future, encouraging teachers to plan student conferences around ONPAR reports at the beginning of the school year will help orient students on how to use the reports. This practice could be useful for teachers and students alike. One teacher noted,

> I did not spend much time going over the student reports with them; therefore, I don't think they understand them very well. Further, I in part, did not go over them with them in too much detail, as I don't think they are very intuitive to the students and would require a fair amount of time to explain the reports to the students.

An additional negative comment about the student reporting section focused on the type of language in the reporting section. Some teachers noted that students felt disappointed when they received low scores or feedback about what they may not understand. The reports were designed to be diagnostic and identify areas of needed improvement; however, the language may have been interpreted negatively by students. One teacher commented that students did not want to read negative feedback, which impacted their desire to read the information provided.

The strongly disagree button for students using their score reports is because they are too long for students to persevere in reading them after

expending quite a bit of mental energy on the tasks. Since the feedback was negative in most cases, it wasn't information that students wanted to read, either.

Project staff hopes to improve this area in the future. Specifically, some teachers commented that the text on individual reports was too technical and dense for students and that it did not emphasize student strengths as well as areas of needed improvement. Developing a student-friendly score report would help ensure that the results are able to be interpreted and used by students, one of the key stakeholders of the assessments, and enable them to have greater autonomy in learning.

Discussion

Bennett's (1998) vision of the reinvention of assessment through technology is helpful to keep in mind regarding the potential impact on classrooms. If educational measurement is to benefit teaching and learning, the way technology-enhanced assessments are used in classroom contexts should be understood. Technology can affect assessment use at several different levels in classroom contexts, ranging from serving as the mode of test administration to offering new types of online items, tools, and methods of gathering responses and providing different types of reporting information to support teaching and learning. Considering these possibilities from the outset and creating on-ramps will help educators anticipate their own and their students' needs to fully leverage the technological power. Adopting and implementing new assessments and technologies is a challenging task for teachers in today's educational environment, and numerous conditions can support or hinder their successful uptake at each of these levels.

Regarding infrastructure, survey results from the ONPAR project indicate that the new technology-enhanced assessments were successful overall in NGSS-oriented science classrooms. Project expectations for implementation time and technology infrastructure were appropriate in the range of schools and districts participating in the pilot. Teachers had ready access to a variety of devices in their schools including iPads, Chromebooks, and computers, and ONPAR's online administration worked equally well on the range of devices.

Teachers and students were able to connect, login, and complete the assessments in the online environment with few to no interruptions. From an infrastructure standpoint, teachers were able to navigate this aspect of technology-enhanced assessments, and the project's professional learning opportunities were successful in orienting teachers to the online administration requirements from the outset. Teachers were accustomed to administering assessments online, and this skill appears to be part of today's teachers' professional repertoire.

Regarding the second phase of assessment reinvention, innovative items and response types and increased sophistication in scoring, ONPAR leveraged technology to present assessment stimuli in innovative ways and offer unique response mechanisms. Graphics and animations helped set up assessment scenarios, and students responded in numerous ways including modeling, graphing, conducting investigations, and explaining.

Teachers reported that they appreciated the innovative approach to assessment and that the materials met their instructional needs for supporting diverse learners in their classrooms. Initially, teachers voiced some concern over student understanding of how to answer new types of questions and the level of challenge of the NGSS-based questions. However, survey responses indicated that teachers felt with opportunity to practice, increased familiarity with innovative assessments, and time with the NGSS, their students would adapt to the new item types and assessment methodology and would become accustomed to the level of rigor expected.

The TPACK framework (Koehler & Mishra, 2009) is useful for thinking of the components of a successful program of professional learning. Through the online meetings, teachers became familiar with the target NGSS (content), and assessment methodology (technology). While ONPAR's on-demand teacher supports in the assessment portal such as the short training videos and task guides could also somewhat meet teachers' needs, the online meetings were rated the highest aspect of the project, demonstrating the benefit of supporting teachers' adoption and use of the materials with real time contact. Survey results clearly indicated that teachers felt the need to be oriented to innovative item types so that they could anticipate needs and feel confident supporting their students. Those implementing new assessments should assume that teachers may want training and opportunities to interact with innovative items prior to using them with their students. Assessments that employ innovative item types should help teachers anticipate the need to orient students to the novel interactions and not expect that students will be initially totally self-sufficient.

ONPAR also undertook efforts to inform teaching and learning through creation of automatic score reports; technology enabled the provision of sophisticated automatic scoring and reporting. Even though teachers appreciated the automatic scoring, the instructional potential of the reporting information proved to the most difficult component of the technology-enhanced assessment to fully realize (TPACK and pedagogy). Numerous teachers commented on how little time they had for reviewing results and planning subsequent instruction.

In the pilot, professional learning meetings required teachers to set aside time for thinking and reflecting on how their students performed and how that would influence their teaching. The time teachers saved in hand scoring assessments was spent in data analysis and reflection instead. Thus, when promoting automatically scored assessments, it may be beneficial to suggest that time saved in hand scoring should be spent engaged in a different assessment-related practice, such as data analysis or in group discussion with other teachers about assessment results.

Surveys did suggest that reporting formats could be improved for teachers and students. Packaging classroom scoring information in an easily interpretable format may help alleviate the amount of time it took teachers to gain insights into student performance and target areas of needed improvement. Incorporating graphic displays or putting results in a roster format may help. These refinements may support teachers' interpretation and use of assessment results for instructional planning.

Similarly, students' use of the automatic scoring and reporting presented challenges. While time was not necessarily an issue, the amount and density of text, as well as wording that was interpreted negatively, hindered students' ability and desire to read the reports. One possible solution is to create interactive reports that require students to analyze and interpret their own assessment data, think about and incorporate feedback, and reflect on their learning needs. If technology is to support reinventing assessment and enable assessment to inform learning, test data must be translated into actionable information for teachers and students.

Further work on ONPAR should refine the reporting mechanisms for both teachers and students so that technology better meets users' needs. Additional research should investigate teacher and student interpretation and use of assessment results to support ONPAR's validity argument. It is important to note, once again, that all teacher participants volunteered in the study and that this self-selected group may have responded more positively to the new technology than what a general population would. Future research should also explore how ONPAR is taken up by teachers who are not self-selected and may not necessarily be early adopters of technology.

Finally, other NGSS technology-enhanced assessments should investigate teacher use of assessment materials in classroom contexts and report on what aspects lead to their successful use. If technology is to support the ability of assessment to meaningfully inform instruction, then researchers need to pave a path for anticipating teachers' and students' needs and create best practices for meeting them.

Conclusion

The ONPAR project took up the challenge of reinventing assessment, harnessing the affordances of technology to provide accessible, challenging, NGSS-based assessment tasks to inform and support teaching and learning in middle school science classrooms. Given the newness and rigor of the NGSS and the level of innovation of the assessment, implementation could have been challenging. Supporting teachers during the time they were transitioning to this new educational tool was important to ensuring its success. The targeted professional learning helped meet the needs teachers faced in day-to-day instruction. As technology-enhanced assessments become a regular part of classroom contexts, the ways teachers and students interact with them and make use of them is an important consideration. If assessment is ultimately to inform teaching and learning, researchers must provide insight into how these tools are successfully adopted and used for pedagogical purposes.

Author Note

The contents of this article were developed under grant S368A150019 from the Department of Education. However, those contents do not necessarily represent the policy of the Department of Education and you should not assume endorsement by the Federal Government.

References

Almond, P., Winter, P., Cameto, R., Russell, M., Sato, E., Clarke-Midura, J., Torres, C., Haertel, G., Dolan, R., Beddow, P., & Lazarus, S. (2010). Technology-enabled and universally designed assessment: Considering access in measuring the achievement of students with disabilities—A foundation for research. *Journal of Technology, Learning, and Assessment, 10*(5). https://ejournals.bc.edu/index.php/jtla/article/view/1605/1453

Alonzo, A. C., & Elby, A. (2019). Beyond empirical adequacy: Learning progressions as models and their value for teachers. *Cognition and Instruction*, *37*(1), 1-37. https://doi.org/10.1080/07370008.2018.1539735

Alonzo, A.C., & Ke, L. (2016). Taking stock: Existing resources for assessing a new vision of science learning. *Measurement: Interdisciplinary research and perspectives 14*(4), 119-152. https://doi.org/10.1080/15366367.2016.1251279

Bennett, R. E. (1998). Reinventing assessment: Speculations on the future of large-scale educational testing. Educational Testing Service. https://www.ets.org/Media/Research/pdf/PICREINVENT.pdf

Bull, G., Hodges, C., Mouza, C., Kinshuk, Grant M., Achambault, L., Borup, J., Ferdig, R.E., & Schmidt-Crawford, D.A. (2019). Conceptual *dilution*. *Contemporary Issues in Technology and Teacher Education, 19*(2). https://citejournal.org/volume-19/issue-2-19/editorial/editorial-conceptual-dilution

Damelin, D., & McIntyre, C. (2021). Technology-enhanced assessments for NGSS classrooms. *@Concord, 25*(1). https://concord.org/newsletter/2021-spring/technology-enhanced-assessments-for-ngss-classrooms/

DeLuca, C., & Klinger, D. A. (2010). Assessment literacy development: Identifying gaps in teacher candidates' learning. *Assessment in Education: Principles, Policy & Practice, 17*, 419–438. https://doi.org/10.1080/0969594X.2010.516643.

Ertmer, P.A. (2005). Teacher pedagogical beliefs: The final frontier in our quest for technology integration? *Educational Technology, Research and Development, 53*(4), 25-39. https://doi.org/10.1007/BF02504683.

Gane B.D., Zaidi S.Z., & Pellegrino J.W. (2018). Measuring what matters: Using technology to assess multidimensional learning. *European Journal of Education. 53*, 176–187. https://doi.org/10.1111/ejed.12269.

Gee, J. P. (2007). *What video games have to teach us about learning and literacy* (2nd ed.). Palgrave Macmillan.

Gorin, J. S., & Mislevy, R. J. (2013). *Inherent measurement challenges in the Next Generation Science Standards for both formative and summative assessment.* https://www.ets.org/Media/Research/pdf/gorin-mislevy.pdf

Graf, S., & Kinshuck (2008). Technologies linking learning, cognition, and instruction. In J. M. Spector, M. D. Merrill, J. van Merriënboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed.; pp. 305-316). Taylor and Francis.

Harris, C. J., Krajcik, J. S., Pellegrino, J. W., & McElhaney, K.W. (2016). *Constructing assessment tasks that blend disciplinary core Ideas, crosscutting concepts, and science practices for classroom formative applications*. SRI International. https://www.sri.com/wp-content/uploads/pdf/constructing_assessment_tasks_2016.pdf

Jewitt, C. (2008). Multimodality and literacy in school classrooms. *AERA Review of Research in Education, 32*, 241-267. https://doi.org/10.3102/0091732X07310586.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. https://doi.org/10.1111/jedm.12000.

Kieffer, M.J., Lesaux, N.K., Rivera, M., & Francis, D.J. (2009). Accommodations for English language learners taking large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research, 79*(3), 1168-1201. https://doi.org/10.3102/0034654309332490

Klinger, D. A., Volante, L., & DeLuca, C. (2012). Building teacher capacity within the evolving assessment culture in Canadian Education. *Policy Futures in Education, Special Issue: Developing Sustainable Assessment Cultures in School Learning Organisations, 10*(4), 447- 460. https://doi.org/10.2304/pfie.2012.10.4.447

Koehler, M.J., & Mishra, P. (2009). What is technological pedagogical content knowledge? *Contemporary Issues in Technology and Teacher Education, 9*(1). https://citejournal.org/volume-9/issue-1-09/general/what-is-technological-pedagogicalcontent-knowledge/

Kopriva, R. J. (2008). *Improving testing for English language learners: A comprehensive approach to designing, building, implementing, and interpreting better academic assessments*. Routledge.

Kopriva, R. J., Thurlow, M. L., Perie, M., Lazarus, S. S., & Clark, A. (2016). Test takers and the validity of score interpretations. *Educational Psychologist, 51*(1), 108–128. https://doi.org/10.1080/00461520.2016.1158111

Kopriva, R.J., & Wright, L. (2017). Score processes in assessing academic content of non-native speakers. In J. Pellegrino & K. Ercikan (Eds.), *Validation of score meaning in the next generation of assessments* (pp. 100-112). Routledge.

Kopriva, R.J., Wright, L.J., Triscari, R., & Shafer-Willner, L. (2021). Examining a multisemiotic approach to measuring challenging content for English learners and others: Results from the ONPAR elementary and

middle school science study. *World Journal of Education Research*, *8*(1), 1-25. https://doi.org/10.22158/wjer.v8n1p1

Kress. G. (2003). *Literacy in the new media age*. Routledge.

Kress, G. (2010). *Multimodality. A social semiotic approach to communication*. Routledge Falmer.

Kress, G., & van Leeuwen, T. (2001). *Multimodal discourse: The modes and media of contemporary communication*. Oxford University Press.

Logan-Terry, A., & Wright, L. J. (2010). Making thinking visible: An analysis of English language learners' interactions with access-based science assessment items. *AccELLerate!*, *2*(4), 11–14. https://ncela.ed.gov/accellerate/summer2010

Mislevy, R., Liu, M., Cho, Y., Fulkerson, D., Nichols, P., Zalles, D., Fried, R., Haertel, G., Cheng, B., DeBarger, A., Villalba, S., Mitman Colker, A., Haynie, K., & Hamel, L. (2009). *A design pattern for observational investigation assessment tasks* (Large-Scale Assessment Technical Report 2). SRI International. https://ecd.sri.com/downloads/ECD_TR2_DesignPattern_for_ObservationalInvestFL.pdf

Mislevy, R. J. (2013). Evidence-centered design for simulation-based assessment. *Military Medicine*, *178*(10), 107-114. https://doi.org/10.7205/MILMED-D-13-00213

Myers, B. (2015). The cognitive basis of ONPAR assessment: A white paper. http://iiassessment.wceruw.org/research/whitePapers/CogSci-ONPAR-bm.pdf

National Research Council. (2012). *A Framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. The National Academies Press. https://doi.org/10.17226/13165

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. The National Academies Press. https://doi.org/10.17226/18290

Pellegrino, J. (2012). Assessment of science learning: Living in interesting times. *Journal of Research in Science Teaching, 49*(6), 831-841. https://doi.org/10.1002/tea.21032

Pellegrino, J. (2013). Proficiency in science: Assessment challenges and opportunities. *Science, 340*(6130), 320-323. https://doi.org/10.1126/science.1232065

Pellegrino, J., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. National Academy Press. https://doi.org/10.17226/10019

Pellegrino, J., & Quellmalz, E. S. (2010). Perspectives on the integration of technology and assessment. *Journal of Research on Technology in Education, 43*(2), 119-134. https://doi.org/10.1080/15391523.2010.10782565

Pellegrino, J., Wilson, M., Koenig, J.A., & Beatty, A.S. (2014). *Developing assessments for the Next Generation Science Standards*. National Academies Press. https://doi.org/10.17226/18409

Reiser, B. J. (2013). *What professional development strategies are needed for successful implementation of the Next Generation Science Standards*? https://www.ets.org/Media/Research/pdf/reiser.pdf

Rivera, C., & Collum, E. (Eds.). (2006). *State assessment policy and practice for English language learners: A national perspective*. Lawrence Erlbaum and Associates.

Sawchuk, S. (2019). Science curriculum reviews are out and results aren't great. *EdWeek*. https://www.edweek.org/teaching-learning/science-curriculum-reviews-are-out-and-results-arent-great/2019/02

Shulman, L.S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1-22. https://doi.org/10.17763/haer.57.1.j463w79r56455411

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic-molecular theory. *Measurement: Interdisciplinary Research and Perspectives, 4*(1-2), 1–98. https://doi.org/10.1080/15366367.2006.9678570

Songer, N., & Ruiz Primo, M.A. (2012). Assessment and science education: Our essential new priority? *Journal of Research in Science Teaching, 49*(6), 683-690. https://doi.org/10.1002/tea.21033

Thurlow, M.L., Thompson, S.J., & Lazarus, S.S. (2006). Considerations for the administration of tests to special needs students: Accommodations, modifications, and more. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (653–673). Lawrence Erlbaum.

Tucker, B. (2009a). *Beyond the bubble: Technology and the future of student assessment*. Education Sector. https://www.immagic.com/eLibrary/ARCHIVES/GENERAL/EDSCTRUS/E090213T.pdf

Tucker, B. (2009b). The next generation of testing. *Educational Leadership, 67*(3), 48-53. https://www.ascd.org/el/articles/the-next-generation-of-testing

Wright, L.J. (2013). *Multimodality and measurement: Promise for assessing English learners and students who struggle with the language demands of tests*. http://iiassessment.wceruw.org/research/

Appendix
Survey Statements and Results

Question	Strongly Agree	Agree	% Agree	Disagree	Strongly disagree	% Disagree	N/A	% N/A
1. My overall experience using the ONPAR materials was positive.	94	57	**97%**	5	0	**3%**	-	-
2. My students' overall experience using the ONPAR materials was positive.	36	98	**86%**	21	1	**14%**	-	-
3. The ONPAR training videos prepared me to use the formative tasks in my classroom.	81	49	**83%**	5	0	**3%**	21	**13%**
4. The task guides helped me understand what concepts and skills would be assess so I could plan for when to use the tasks during instruction.	83	61	**92%**	1	1	**2%**	10	**6%**
5. The classroom score reports were easy to understand and interpret.	59	79	**89%**	11	4	**10%**	3	**2%**
6. The individual score reports were easy to understand and interpret.	53	78	**84%**	12	4	**11%**	9	**6%**
7. The score reports helped me identify student learning needs.	58	74	**84%**	19	2	**13%**	3	**2%**
8. The Task Guides helped me identify next	56	78	**86%**	6	2	**5%**	14	**9%**

steps for instruction for my students.								
9. The ONPAR tasks aligned with the curriculum materials I currently use.	62	82	**93%**	12	0	**8%**	-	**-**
10. It was easy to set aside enough time to familiarize myself with the ONPAR materials so I could use them with my students.	70	74	**92%**	11	1	**8%**	-	**-**
11. The project meetings with ONPAR staff helped prepare me to use the tasks and materials in my classroom.	110	43	**98%**	1	2	**2%**	-	**-**
12. It was easy to set aside enough instructional time to administer the tasks to my students.	69	77	**93%**	7	3	**6%**	-	**-**
13. I was able to set aside enough time to interpret the assessment results and plan instruction with information from the reports.	38	83	**78%**	31	3	**22%**	-	**-**
14. Overall, these assessment materials are consistent with my school climate, and reforms occurring in my district and school.	76	70	**94%**	10	0	**6%**	-	**-**

15. These materials fill a need I have in my classroom.	73	70	**92%**	12	1	**9%**	-	-
16. Students understood and used their individualized score report.	11	77	**56%**	55	13	**43%**	-	-
17. The level of challenge of the tasks was appropriate for my students' knowledge, skills, and abilities.	29	99	**82%**	28	0	**18%**	-	-
18. The ONPAR tasks enhanced instruction for my students.	54	89	**92%**	11	2	**8%**	-	-