Table of Contents

The Wisconsin Center for Education Research (WCER) in collaboration with Detroit Public Schools, St. Martin Parish Public Schools, Educational Measurement Consulting, The Learning Agency, and Activate Learning LLC propose the *Teaming Up for Equity in Science: Supporting NGSS Three-dimensional Learning and Achievement through Actionable Assessment* project. This project addresses **Absolute Priority 1** (AP1: Demonstrates a Rationale), **Absolute Priority 3** (AP3: Field-Initiated Innovation- STEM) and **Competitive Preference 1** (CP1: Promote Equity and Adequacy in Educational Opportunity and Outcomes). The project will implement an entrepreneurial, evidence-based, field-initiated innovative assessment called ONPAR that will improve achievement and attainment for middle school science students, especially English Learners (ELs) and students who struggle with reading. The project will also rigorously evaluate the innovation with a randomized controlled trial that meets What Works Clearinghouse standards without reservation.

ONPAR is a web-based, technology-enhanced classroom assessment system that reconsiders the way language is used in testing by increasing the presence of multisemiotic representations (Kress & van Leeuwen, 2001) such as visuals, graphics, animations, and onscreen assists to ask questions and gather input. Validation studies indicate that this method effectively conveys the meaning of challenging constructs that often require a heavy language load, thereby providing more equitable access for students disadvantaged by traditional, language-heavy assessments ([redacted], 2017; [redacted] et al., 2021). Thus, ONPAR can offer better information for about 40% of students so that instructional, and programming decisions can better support high-need students.

The most recent ONPAR project, completed in 2019, resulted in a full range of online middle school science classroom assessments aligned to the Next Generation Science Standards

(NGSS) (NGSS Lead States, 2013), automatic scoring and reporting, and teacher professional development for classroom implementation. ONPAR is unique in that it aims to measure students' science abilities *and* prompt students and teachers to plan next steps for learning. Thus, when implemented with fidelity, the full ONPAR suite functions as an intervention.

The proposed project will examine how teachers and students leverage ONPAR assessments and reporting data to inform teaching and learning. We will also conduct a randomized controlled trial to assess ONPAR's effectiveness. The four goals of this project are: (1) refine and validate ONPAR materials to meet the identifiable needs of users, (2) increase teacher efficacy for NGSS assessment and data-driven instruction, (3) improve outcomes for high-need middle school science learners, particularly ELs and students who struggle with reading, and (4) build capacity for sustaining and expanding use of the ONPAR system.

A. Significance

Since the reauthorization of the Elementary and Secondary Education Act under No Child Left Behind, large-scale assessments have formed the backbone of accountability for K–12 schools. Though many scholars have articulated drawbacks to an overemphasis on large-scale testing (e.g., Darling-Hammond, 2007), systematic assessment has helped shed light on educational inequality. Recent work calls for the development of thoughtful systems of assessment that provide both annual accountability and ongoing monitoring in classrooms (Shepard et al., 2018) because they can deliver timely data to teachers and students, inform day-to-day instructional decisions, and point to the assistance students need to enhance learning (Noyce & Hickey, 2011; Shepard et al., 2020). Research shows that the most effective teachers use student assessment data on an ongoing basis to inform plans for student growth (Pellegrino et al., 2014), yet most teachers struggle to make inferences from classroom evidence to inform instructional decisions (Pellegrino et al., 2014; Morrison, 2008). Shepard et al. (2020) write that assessments are "most useful when they provide specific substantive insights about student thinking—where student understandings are on firm ground as well as where they are stuck, and more importantly what alternative conceptions might be in the way of making progress" (p. 6).

If they are to make data part of an ongoing cycle of instructional improvement (Hamilton et al., 2009), **teachers must have access to effective formative assessment materials designed to provide meaningful data *while they teach*** (Pellegrino et al., 2014; Morrison, 2008) and need to understand how to interpret data to make instructional decisions. In addition, to amplify the potential of assessment for learning, **students must have access to data and be taught how to use it to set learning goals** (Hamilton, et al., 2009). Offering students resources and opportunities to make interpretive decisions from assessment data will enable them to take charge of their own learning. **The innovative ONPAR intervention is a new strategy that holds promise for both teachers and students to accomplish these aims**.

Prior Research. Five large projects have contributed to ONPAR's research base. The first three (PR #S368A060007, PR #S368A070001, and PR #S368A090029) examined the viability of a multisemiotic approach to assessment for ELs in science and mathematics, and validated the standardized layout of screens and accessibility features. Psychometric research indicated that low English proficient ELs and students with disabilities in reading scored significantly better on ONPAR than traditional tests, and, holding content ability constant, scored as well as their non-EL peers. Mainstream students scored similarly on both the traditional and ONPAR tests, indicating that the ONPAR approach did not give ELs or those with reading disabilities an undue advantage ([redacted], 2017; [redacted] et al., 2021).

The other two projects focused on classroom-based assessment in mathematics and science. One (PR #R305A140117) found interactions between language proficiency or reading ability and science ability were statistically significant. The other (PR #S368A150019) produced 12 units of middle school science assessment materials and researched their technical quality. User feedback was positive and indicated that there was a high need for accessible assessment materials aligned with NGSS. Survey results indicated that the materials were feasible for classroom use, technology infrastructure was adequate, tasks were appropriately challenging, teacher materials were adequate, and training was highly effective ([redacted] et al., 2022). Psychometric results indicated that the tasks were valid and reliable for intact classrooms of students ([redacted], 2020). The middle school science materials from this project are currently licensed for use in middle schools by Activate Learning, with a user base of approximately 55,000 students in the United States.

Rigorous evaluation of the ONPAR system is needed to determine whether it leads to improved learning for high-need students in middle school classrooms. Our previous grant allowed us to develop a comprehensive classroom system and implementation materials. However, results indicated that there was a significant teacher effect, which was conflated with school and district effects. The proposed randomized controlled trial will collect the data needed to investigate and explain the main and secondary effects. In addition, the previous study developed and piloted training for implementation of the materials, but the fidelity of implementation was not systematically studied.

B. Project Design

(1) Conceptual Framework

The concept underlying this project is that the ONPAR intervention is a tool for training teachers to effectively use instructionally-embedded NGSS assessments in classrooms and harness data for instruction. The assessments validly and reliably measure student understanding of conceptually complex NGSS material *and* provide useful information to teachers and students to guide instruction and learning ([redacted], et al., 2022). By using ONPAR, teachers develop skills to support the learning of high-need students, while students develop greater autonomy and ownership. Providing accessible NGSS assessments also enables fairer assessment for high-need students and helps improve outcomes. **Appendix G** includes a logic model.

Theory of Change: Research Foundations

Contemporary views of science proficiency in the United States under the NGSS (NGSS Lead States, 2013) call for students to develop their science abilities in three dimensions: disciplinary core ideas; science and engineering practices; and crosscutting concepts (National Research Council, 2012). Assessing students' abilities has been one of the challenges facing science education since the inception of the 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), pre-NGSS science assessments were unidimensional and focused on disciplinary core ideas, and most test formats were limited to multiple-choice questions (Sawchuk, 2019).

Numerous researchers have noted that use of technology in assessment affords new opportunities that were not possible with paper and pencil tests, notably the ability to better assess the construct of interest, skills, and reasoning abilities (Alonzo & Ke, 2016; Gane et al., 2018; Pellegrino & Quellmalz, 2010). Technology-enhanced assessment offers the chance to change *what* is observed in the assessment context, and *how* it is observed, because computers have the capability to deliver novel stimuli and gather unique responses that are not possible in traditional formats (Gane et al., 2018; [redacted], 2017; Tucker, 2009). Moreover, technology-enhanced assessments have potential to gather as well as interpret evidence of student learning behaviors. Assessments can present information dynamically using animations and graphics and offer novel response types that enable students to draw, model, and conduct investigations. Computer algorithms can be generated to interpret these behaviors with less subjectivity, easing the burden of scoring for educators. Thus, technology can efficiently assess practices and skills that are better matched to the types of reasoning and response processes that are of interest (Gorin & Mislevy, 2013). This is especially useful under the NGSS. However, adoption and implementation of new technology is challenging for educators (Koehler & Mishra, 2009). Gane and colleagues (2018) recommend that teachers receive support to help utilize technology-enhanced assessments. This project will provide much needed NGSS assessment tools (**AP3**) and instructional support to improve the outcomes of high-need students (**CP1)**.

(2) Goals and Objectives

The project's four goals will lead to improvement of the ONPAR system, successful implementation, and rigorous research and dissemination (see **Table 1** below).

Goal 1: Refine and validate ONPAR materials to meet the identifiable needs of users. By the end of Y1, all ONPAR score reports will be updated based on past user feedback and validated with panels of science educators (objective 1.1). Reviews will be made to student reports and administrative reports, both of which were newly developed after the completion of the 2019 project. In addition, by the start of the 2023-24 school year, all ONPAR units will be programmed using the Competencies and Academic Standards Exchange (CASE) tagging, a new digital system for documenting and referencing learning standards and competencies (objective 1.2). This will make it easier for districts, schools, and teachers to identify the standards ONPAR materials address. Finally, by the end of Y1, the project will conduct Web Content Accessibility Guidelines (WCAG) review and incorporate at least two additional WCAG accessibility features (objective 1.3). WCAG provides technical specifications to improve the accessibility of digital applications across all devices for people with a wide range of disabilities. These enhancements will ensure that ONPAR materials meet industry standards for educational technology.

Goal 2: Increase teacher efficacy for NGSS assessment and data driven instruction (**AP3**). Project staff will work with two partner school districts that currently license ONPAR, Detroit Public Schools and St. Martin Parish. By the end of Y3, 80 middle school teachers in Grades 7 and 8 will be recruited for the project; 40 teachers in the treatment group will receive ongoing training (objective 2.1). Treatment teachers will attend nine hours of online training with project staff per year. In the first meeting, teachers will be introduced to ONPAR assessments, discuss NGSS 3D assessment, and use backward planning for assessments. Teachers will be asked to plan for implementation of assessments during the relevant unit of instruction. Recommendations for assessment pacing will be provided, though teachers will be given the freedom to determine when to administer an ONPAR assessment task. Teachers will also attend follow-up sessions where they will discuss score reports. Project staff will provide training on how to read ONPAR reports, how to utilize reporting feedback to plan instruction, and how to support students' independent use of reports. Project staff will be in touch regularly with participants to support retention. Teachers will receive cash incentives upon completion of each unit in accordance with their district salary ranges. We have found this strategy works well to retain teachers. During Y2 and 3, periodic meetings will be held with administrators to update them on teacher and student participation (objective 2.2); they will be given access to and trained to use an administrative portal to monitor ONPAR data. During Y2 and 3, 80 teachers will take at least one survey per completed unit (objective 2.3) and at least 10 participants will be interviewed (objective 2.4) to evaluate their experience related to NGSS assessment and instruction. Based on our previous project, we expect 90% of teachers to report an increase in support for NGSS assessment and instruction after using ONPAR.

Goal 3: Improve the outcomes for high-need middle school science learners, particularly ELs and students who struggle with reading (**CP1**). The project will conduct a randomized controlled trial to measure intervention effectiveness. During Y2 and 3, a total of 40 treatment teachers will implement the ONPAR system with 4,500 students (objective 3.1). Each ONPAR unit has four to five formative assessments to be instructionally embedded when the teacher deems feedback is useful for instruction. Following each assessment, teachers and students receive automatic score reports. Treatment teachers will train students to interpret and use the score reports and create action plans. At the end of the unit, treatment teachers will administer the ONPAR end-of-unit assessment and a traditional science test (see Appendix J for samples); in the comparison condition, comparison teachers will administer traditional formative assessments during unit instruction as well as the traditional science test. A total of 9,000 students will take traditional science tests. Additional data from schools and districts will be gathered including demographic information on students, state assessment scores for science, math, and reading, EL status and proficiency level, and special education status (objective 3.2). The external evaluator will analyze survey and assessment data (objective 3.3).

Goal 4: Build capacity for sustaining and expanding the ONPAR assessment system. ONPAR is licensed through Activate Learning LLC. Through this project, WCER and Activate

Learning will work together to enable expansion of the product. During Y1, a survey of districts licensing the product will be conducted (objective 4.1) to understand their NGSS instructional needs and digital infrastructures. Currently ONPAR is licensed in 75 LEAs. During Y2-4, a sustainability plan will be crafted. To disseminate information about ONPAR, the team will participate in at least one conference each year of the project (objective 4.2). Project staff will also work with WCER communications to publish articles and other written sources of information about ONPAR throughout the project (objective 4.3).

Table 1: **Goals, Objectives, and Outcomes**

Goals	**Objectives**	**Outcomes**
1. Refine and validate ONPAR materials to meet the identifiable needs of users	1.1 By the end of Y1, ONPAR score reports will be updated based on prior user feedback and validated with panels of science educators 1.2 By the start of the 2023-24 school year, all ONPAR units will be tagged using the CASE system 1.3 By the end of Y1, at least 2 additional accessibility features will be incorporated into the ONPAR suite of tools	· 72 ONPAR digital score reports are updated and revised · Two validation panels are held to review score reports · 12 units of assessments are reviewed and programmed for CASE tagging · 12 units of assessments are programmed for additional WCAG compliant accessibility features

2. Increase teacher efficacy for NGSS assessment and data-driven instruction	2.1 By the end of Y3, 80 teachers will be recruited for the project; 40 teachers in the treatment group will receive ongoing training 2.2 During Y2 and 3, periodic meetings will be held with district and school administrators 2.3 During Y2 and 3, 80 teachers will take one survey per completed unit 2.3 During Y2 and 3, at least 10 teachers will be interviewed	· 80 teachers are recruited for study · 40 teachers are assigned to treatment and 40 teachers are assigned to control conditions · 54 PD meetings are held with treatment teachers · 8 meetings are held with district and school leaders · 80 teachers are surveyed and 10 teachers are interviewed · 90% of treatment teachers report an increase in support for NGSS assessment and instruction

3. Improve the outcomes for high-need middle school science learners, particularly ELs and students who struggle with reading	3.1 During Y2 and 3, a total of 40 treatment teachers will implement the ONPAR system 3.2 During Y3 and 4, district and state data will be collected from 2 school districts. 3.3 During Y3 and 4, project data will be analyzed.	· 4,500 students taken ONPAR assessments (GPRA1,2) · 9,000 students take traditional science tests (GPRA1,2) · Data on 9000 students is collected (GPRA1,2) · students show a statistically significant (non-directional t-test, $p<.05$) difference in mean scores on the traditional science test and state assessments (GPRA3,4,5)
4. Build capacity for sustaining and expanding the use of the ONPAR assessment system	4.1 During Y1, a needs analysis and technology capabilities survey of districts will be conducted 4.2 Participate in one conference in each year of the project 4.3 Publish articles and other written sources of information about ONPAR throughout the project.	· 75 surveys to current district users are sent; data is collected and analyzed · 14 proposals submitted to conferences · Four journal articles are written and submitted and 84 social media posts are made.

(3) Addressing Needs of Target Population

Teacher Needs: Assessing students' NGSS abilities is one of the challenges facing science education (Alonzo & Ke, 2016; Pellegrino, 2012; 2013; Pellegrino et al., 2014; Songer & Ruiz Primo, 2012). ONPAR fully supports NGSS learning through its training, implementation materials, assessments, and reporting. Research shows that teachers were highly satisfied with ONPAR materials and that it met their need for NGSS assessment ([redacted] et al., 2022). ONPAR also addresses a need that teachers have for insight into student performance. All too often teachers are left with assessment scores but relatively little insight into what students found difficult. ONPAR's unique reporting mechanisms provide ongoing insights to teachers so that they can better use instructional time to meet student learning needs.

Student Needs: ONPAR was developed to respond to the need for fair, accessible assessment strategies for today's diverse students ([redacted], 2008). The assessments are rooted in multisemiotic communication; text-heavy traditional test items are re-envisioned with graphics, animations, and novel item types, lessening the burden students face with a linguistic-heavy assessment. Research indicates that ELs and students with disabilities in reading perform better on ONPAR than when using traditional tests measuring the same constructs, yet ONPAR works equally well for general education students ([redacted] et al., 2016; [redacted] et al., 2021).

ONPAR also addresses students' need for clear reporting information. Rather than treating assessment as an endpoint, ONPAR helps students see assessment as a catalyst in the learning cycle. Teaching students to examine their own data and set learning goals can improve outcomes (Hamilton et al., 2009). ONPAR's reports move students to reflect on bigger ideas, set learning goals, and collaborate with their teachers, ultimately leading to students to become more self-motivated learners. **ONPAR meets the needs of teachers and students alike by providing innovative NGSS assessments that support teaching and learning**. The two partner school districts, Detroit and St. Martin Parish, are seeking NGSS assessments that meet diverse students' needs. Specifically, the average reading proficiency of middle school students in Detroit is 13% and 12% of students are designated as ELs. While 72% of middle school students are proficient in reading in St. Martin Parish, 54% students are in poverty and are deemed at risk. ONPAR is an ideal tool for ensuring fairer assessment for these high-need students.

C. Personnel

Principal Investigator (PI) [redacted]**, WCER**, will assume primary responsibility of the project at WCER (.45 FTE). She will oversee refinement of ONPAR materials; data collection; collaboration with advisors; and capacity and expansion efforts. [redacted] brings 20 years of experience in education research and grant management. She is a sociolinguist whose research focuses on how students express understanding through discourse in complex learning environments. She has worked in science education for over 15 years, conducting research on how students make sense of phenomena through spoken and written language. She has served as Co-PI and PI on recent ONPAR grants including the one that established it as a commercial product. [redacted] contributes expertise in assessment development, EL education, multimedia analysis, and educational linguistics.

Co-PI [redacted] **WCER** (.725 FTE), will oversee day-to-day aspects including material revision, staff management, and teacher training. [redacted] spent nearly 20 years as a science educator and administrator before becoming a middle school science teacher. She has designed and delivered a wide variety of STEM programs for schools and the public, including a district-wide effort to train elementary teachers in inquiry science. She also managed a 4-year NSF-funded project designed to bring urban and suburban students together for high-impact inquiry science experiences. [redacted] has worked on ONPAR since 2016, designing professional learning and training teachers to use the ONPAR system. [redacted] holds a BS in Meteorology, an MA in Urban Education Studies, and a certificate in online learning. She is certified to teach Grades 7–12 general science.

Research Methodologist, [redacted], University of North Carolina-Greensboro (.06 FTE), will oversee research design and psychometric aspects of the project. [redacted] holds a PhD in quantitative psychology and an MS in statistics. Since 2005, he has taught multiple advanced statistics courses including courses on hierarchical linear modeling, multivariate statistics, and structural equation models, and is a leading expert on diagnostic classification models. His current research focuses on modeling approaches to scoring exams that identify skills and/or attributes that an examinee has mastered or has not mastered.

[redacted] **Education Measurement Consulting** (.08 FTE), will serve as the evaluator for the project. [redacted] was the administrator of assessment programs at the district and state level in Oregon for over 20 years, and has worked on all aspects of assessment design, test administration, scoring, and reporting results. At the Oregon Department of Education, he had supervisory responsibility for all aspects of the state's assessment and accountability systems, including Oregon's innovative adaptive online assessment, assessments of students with disabilities and ELs, and performance assessments of writing and mathematics. [redacted] currently provides psychometric services for state assessment and accountability systems. He holds a PhD in educational psychology.

WCER project personnel will also consist of three positions to be hired for the project: a full-time **science outreach specialist** with expertise in science professional development; a full-time **researcher** with background in qualitative and quantitative research methods; and one **graduate student** with experience in education. UW–Madison is an equal opportunity employer and will ensure compliance with federal and state laws protecting against discrimination in the hiring process. In addition, UW has adopted policies that both emphasize existing protections and supplement them with protections against discrimination that are not available under either federal or state law. Federal and state laws provide separate prohibitions against discrimination that is based on race, color, creed, religion, sex, national origin or ancestry, age, or disability. State law additionally prohibits discrimination that is based on sexual orientation, arrest or conviction record, marital status, pregnancy, parental status, military status, or veteran status. The project will seek applicants representing diverse backgrounds.

WCER Technical Services Department (TS) will refine digital materials and maintain the testing portal and database of results and support maintenance and technical support for the online ONPAR system and data management housed on a SQL server.

The project will also use a **Technical Advisory Committee** (TAC) of experts with deep experience in the disciplines that inform the conceptual and operational underpinnings of this project. Twice per year, members will meet and participate in online and/or face-to-face meetings, with additional communication as needed. The TAC will include: [redacted] (associate professor of science education at Michigan State University) whose research focuses on learning progressions, formative assessment in science, and how assessment materials interact with teacher instruction to support learning; [redacted]**,** (PI on prior ONPAR projects), whose research focuses on innovative assessment; and [redacted] (professor at University of Illinois-Chicago) whose research focuses on cognitive science, learning, and NGSS assessment. Three Activate Learning staff will serve in-kind on the project: [redacted], director of curriculum and instruction, [redacted], vice president of sales, and [redacted]

[redacted], chief product officer. The project will also consult with an educational product consultant who specializes in marketing educational technology tools.

D. Management Plan

The work will be grouped into four key areas: (1) project management; (2) teacher outreach and support; (3) research; and (4) technical. **Table 2** outlines staff roles and responsibilities; a timeline appears in Appendix J. The **project management team** will be comprised of the PI ([redacted]), Co-PI ([redacted]), and methodologist ([redacted]). This team will meet every 2 weeks to discuss project goals, milestones related to the project timeline, and budget. The PI will take overall responsibility for budget, meetings, reviewing goals, objectives, and timeline. The team will also craft the sustainability plan, meeting quarterly with Activate Learning and educational consultant. The **teacher outreach and support team** will consist of one Co-PI ([redacted]), the science education researcher, and a graduate student. They will be in charge of recruitment, training, and monitoring participant progress and will meet weekly to discuss progress and resolve any issues. Co-PI [redacted] has extensive experience working with middle school teachers and led recruitment and training in the past ONPAR grant. The **research team** will be comprised of the PI ([redacted]), Co-PI ([redacted]), methodologist ([redacted]), evaluator ([redacted]), and the UW researcher. This team will meet every 2 weeks during the pilot period to review data quality. During Years 3 and 4, the entire project team will meet weekly to review data and analyses and work on academic papers and presentations. The **technical team** will be comprised of a manager from WCER TS and one Co-PI ([redacted]). This team will meet weekly during Year 1. The TS manager will oversee a staff of two software developers. As is the industry standard, the TS manager assigns IT development in 2-week sprints and allocates staff time using a scrum manager. The TS team meets with their manager on a daily basis. The TS manager will meet with project staff weekly in Y1. Additionally, the director of TS and the PI will meet once a month to ensure that progress is made and that the project is within budget. In Y2–4, during the intervention and analyses, the TS manager and Co-PI ([redacted]) will meet every 2 weeks to review data captures and resolve any technical issues. The ONPAR staff has successfully conducted an assessment pilot and analyses with this structure in the past.

***Table 2*: Staff Roles and Responsibilities**

Team	Lead	Other Staff	Responsibilities
Management	PI [redacted]	Co-PI [redacted] Methodologist [redacted] Ed Consultant Activate Learning	Manage timeline Manage budget Create sustainability plan
Teacher Outreach and Support	Co-PI [redacted]	Science Outreach Specialist Graduate student	Coordinate with school districts Recruit and train teachers Monitor participant progress

Research	PI [redacted]	Methodologist [redacted] Evaluator [redacted] Researcher Co-PI [redacted] Graduate student	Review and/or develop research instruments and identify traditional tests Assign participants to treatment and control conditions Monitor data quality Analyze data Disseminate findings
Technical	Co-PI [redacted]	Technical services manager PI [redacted] Technical services director	Perform technical work on ONPAR reports Implement additional WCAG accessibility features Create and implement CASE tags Monitor technical services budget

E. Evaluation

The project evaluation seeks to answer three research questions (RQs): **RQ1**. Does the use of ONPAR assessments increase performance on non-ONPAR proximal and distal NGSS measures? **RQ2**. Do ONPAR end-of-unit assessments measure high- need students' target NGSS in such a way that non-construct relevant variables (e.g., language proficiency or reading ability) are effectively minimized? **RQ3**. How effective and useful is the ONPAR intervention for supporting teachers' NGSS-based assessment and teaching practices?

(1) Evidence the Project Meets What Works Clearinghouse Standards

Overview Years 1–4: To investigate the proposed RQs, the evaluation plan will include an impact evaluation and a process evaluation. Evaluation data will include survey and interview data from teachers; administrative data, such as enrollment and demographic data; and state assessment data (distal measures). Classroom assessment outcomes, ONPAR and traditional (proximal measure), will also be collected. In **Year 1**, the evaluator will conduct a process evaluation on the validation process for reviewing final ONPAR materials and finalize plans for conducting the randomized controlled trial (RCT). The impact evaluation will consist of an RCT that meets WWC standards without reservations (version 4.1). The RCT will evaluate six of the 12 ONPAR science assessment units used alongside Activate Learning's IQWST science units. The project will study ONPAR usage with three of four IQWST science units typically used in a school year to account for possible slippage in pacing at sites due to external factors such as weather, testing, etc. In **Year 2**, the impact evaluation will focus on outcomes from three units with Grade 8 students. In **Year 3**, the evaluation will examine outcomes from six units; three units with Grade 7 students and the same three units studied in Y2 with Grade 8 students. Final analyses will be conducted in **Year 4**.

The **sample** will include Grade 7 and 8 science teachers and their students (see Analytic Plan below). The evaluator will use **stratified random sampling** to assign participants to treatment and control groups; within a strata, half will be randomly assigned to treatment and the remaining half to control. Prior to assigning participants, the evaluator will identify school demographic factors associated with educational outcomes, such as percentage of students receiving free or reduced meals, percentage of student racial/ethnic background, percentage of ELs at the school, geographic setting (rural, urban), and performance on state assessments. The evaluator will create **strata** of teachers with similar characteristics from which they will randomly assign teachers to treatment and control groups, which will help reduce selection bias. Project staff will then be in direct contact with teachers to explain the requirements for participation. If concerns arise, project staff will contact school principals to troubleshoot. The project will only take teachers who are new to using ONPAR assessments; any teacher who has previously used ONPAR will be excluded. To ensure minimal contamination between treatment and control groups, separate portals will be created for the two groups. The **treatment portal** will contain all training materials as well as the array of ONPAR assessments and materials. The **control portal** will only contain access to the traditional science test.

Prior to the project, **treatment group** teachers will answer a survey to provide information on the level of support they receive for NGSS-based instruction and assessment and an estimate of the amount of time they spent on planning and grading assessments last year. Demographic information will be collected such as educational background, years of teaching, familiarity with NGSS, and experience with IQWST curriculum materials. Teachers will also take the Michigan Assessment Consortium Assessment Literacy Self-Assessment for Educators to gauge their assessment literacy.

Treatment teachers will administer a total of four ONPAR assessments during one IQWST unit, three during instruction and one at the end of the unit. To protect against bias in timing of assessment administration, teachers will be asked to give the end-of-unit assessment after completing a certain IQWST lesson near the end of the unit. The final assessment will include the ONPAR end-of-unit assessment as well as a commercial traditional science assessment (proximal measure); the commercial test will be the same for all teachers teaching a given unit (e.g., Concord Consortium assessment for Chemistry unit or Cognia assessment for

Plate Tectonics unit; see Appendix J for sample traditional tests). The traditional assessment will be selected on the basis of measuring the same NGSS as ONPAR but with greater reliance on text. Selection of the traditional assessment will be reviewed by the TAC. At the end of a unit, **treatment teachers** will also answer a survey to gather information about the level of support they received for NGSS-based instruction and assessment during this unit, amount of time for planning and grading, administration information to gauge the fidelity of implementation, as well as feedback about their experience and perceived student experience.

Prior to the project, **control group** teachers will answer the same beginning-of-year survey as the treatment teachers about the level of support they receive for NGSS-based instruction and assessment, and an estimate of the amount of time they spent on planning and grading assessments last year. Demographic information will be collected, such as educational background, years of teaching, and familiarity with IQWST curricular materials. As with treatment teachers, control teachers will be asked to take the Michigan Assessment Consortium Assessment Literacy Self-Assessment for Educators to gauge their assessment literacy.

In the **control condition**, teachers will also use IQWST instructional units and will be allowed to use any traditional formative assessment strategies they would normally use. Participants will log the frequency of formative assessment on a project template, recording the IQWST lesson on which the assessment was based, the NGSS it aimed to measure, and grading procedures. They will also be asked to provide a sample of the assessment to the project team. At the end of an IQWST instructional unit, teachers will administer the same traditional science test as treatment teachers (see description above). As mentioned above, to protect against bias in timing of assessment administration, teachers will be asked to administer the end-of-unit assessment after completing a certain IQWST lesson near the end of the unit.

Control group teachers will be surveyed after each unit to gather information on the level of support they receive for NGSS-based instruction and assessment, amount of time they spend planning and grading assessments, as well as their use of formative assessment materials during the unit of instruction. This cycle will continue through each of the IQWST units.

Table 3: **Groups for Analysis**

Year 1	**Year 2**
Grade 8 treatment (1-year experience)	Grade 8 treatment (2-year experience)
Grade 8 control	Grade 8 control
n/a	Grade 7 treatment
	Grade 7 control

(2) Feedback and Periodic Assessment

Formative evaluation will provide timely performance feedback on progress toward the project's intended outcomes. Data from document review, surveys, and interviews will inform continuous improvement. Periodic feedback will be provided to the project team through quarterly data memos, annual reports, and monthly meetings with the evaluator. In Y1, the evaluator will review plans for validation panels and for project implementation. In Y2 and 3, during the implementation, the evaluator will meet every two weeks with the project team to ensure that project milestones are being met so that goals and objectives can be achieved. The evaluator will review relevant outcome data collected as well as survey and interview data, and discuss potential changes that could improve program implementation and impact. Changes will be implemented between Y2 and 3.

Performance data on all project outcome measures will be updated and detailed in quarterly memos and annual reports, delivered 30 days before federal reporting dates. Annual reports will provide a summary and narrative description of the project's progress toward each goal and objective. Each annual report will be delivered prior federal deadlines and will provide accurate and high-quality data for the annual performance report. Finally, the impact evaluation will provide a rigorous examination of the program's impact on student outcomes.

(3) Key Project Components and Outcomes

RCT: To answer RQs 1 and 2 above, the study will use a randomized controlled trial experimental design. Data will come from participating school districts as well as directly from the project SQL server and will be at the student level. District data will include demographic data (gender, race/ethnicity, socioeconomic status, EL status and test scores, disability status, state English language arts, mathematics, and science test data; state science tests are given in Grade 8 and serve as the distal measure). The dataset will allow students to be linked to their respective teachers. Additionally, data from ONPAR assessments and the traditional science assessments will be collected. Traditional tests serve as the proximal measure in the analysis.

The **treatment sample** in Year 2 will include 20 Grade 8 teachers and an estimated 75 students per teacher, giving a total of 1,500 treatment students. The **comparison sample** in Year 2 also will include 20 Grade 8 teachers and an estimated 1,500 students. This estimate is 20% higher than needed to account for attrition. A power analysis conducted in PowerUp! (Dong & Maynard, 2013) using 16 teachers per group found that this sample within Year 2 will be sufficient to detect small to medium effects of approximately .35 standard deviations. For this analysis it is assumed that α=.05, using a two-tailed test, where power is set at 0.80. Additionally, an ICC of .15 is assumed with R^2 for Level 1 covariates set to .25 and an R^2 for Level 2 covariates set to .25. The **treatment** and **control samples** in Y3 will double as the project works at two grade levels. Thus, Y3 samples will be 20 Grade 7 and 20 Grade 8 treatment teachers and 20 Grade 7 and 20 Grade 8 control teachers. Assuming that each teacher has 75 students, this results in 1,500 Grade 7 and 1,500 Grade 8 treatment students and 1,500 Grade 7 and 1,500 Grade 8 control students. The additional sample will be sufficient to detect smaller effect sizes. The study will feature improved control of construct irrelevant variance through randomization and improved measures of English proficiency and reading level. We expect to find interactions with "small" to "moderate" effect sizes indicating that ONPAR tests are unbiased with respect to English proficiency and reading level.

For RQ1 and RQ2, we will use a hierarchical linear model (HLM, Raudenbush & Bryk, 2002), with students nested within classrooms. Traditional regression makes assumptions about the independence of errors among all observations. When using nested data (e.g., students that are nested within a given teacher), the assumption of independent errors is often violated. HLM accounts for this dependency by specifying models at two levels. The first level (e.g., student level) models the variation between students within teacher, whereas as the second level (e.g., teacher level) models the variation between teachers. By partitioning the variation in this way, we can account for dependencies that result from multiple students sharing the same teacher that cannot be explained by other covariates.

To answer the **first two research questions** using HLM, ONPAR end-of-unit assessment will be used as the dependent variable. Students are nested in classrooms and the impact analysis will test for differences between treatment and control conditions while controlling for additional covariates such as race/ethnicity, socioeconomic status, and gender. **For RQ2**, English Ability and traditional science tests will be included as independent variables in addition to a variable indicating treatment/control conditions. In this case, interaction terms will also be included to test for whether or not English ability and reading ability moderates the relationship between the traditional test scores and ONPAR end of unit assessments. Note that this analysis will be conducted three times: Grade 8 in Year 2, Grade 8 in Year 3, and Grade 7 in Year 3. These results will be compared to explore whether the effects are similar across grades. The specific models are provided in **Appendix J**. We expect students will show a statistically significant (non-directional t-test, $p<.05$) difference in mean scores on the traditional science test (proximal measure) and state assessments (distal measures).

We will answer **RQ3** using a repeated measures design using the teacher beginning-of-year survey and post unit surveys. In the repeated measures design, the pre and post surveys will be the single within factor and treatment will be a between factor. We will control for additional covariates such as race/ethnicity, socioeconomic status, and gender, and will explore changes across units as a follow-up exploratory analysis specific to only the treatment. We will complete a repeated analysis using only the treatment across the four survey observations, with follow-up post hoc tests for differences between the repeated surveys. Again, these analyses will be completed three times; for Grade 8 in Year 2, Grade 8 in Year 3 and Grade 7 in Year 3. The results will be compared to explore for differences between Grade 7 and 8. We expect 90% of treatment teachers to report an increase in support for NGSS assessment and instruction.

References

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.

[redacted]

Darling-Hammond, L. (2007). Race, inequality, and educational accountability: The irony of "No Child Left Behind. *Race, Ethnicity and Education*, *10*(3), 245–260.

Dong, N. & Maynard, R. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental studies. *Journal of Research on Educational Effectiveness*, *(6)*24: 24–67.

Gane B. D., Zaidi S. Z., & Pellegrino J. W. (2018). Measuring what matters: Using technology to assess multidimensional learning. *European Journal of Education. (53)*2,176–187. https://doi.org/10.1111/ejed.12269.

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.

Hamilton, L., Halverson, R., Jackson, S., Mandinach, E., Supovitz, J., & Wayman, J. (2009). *Using student achievement data to support instructional decision making* (NCEE 2009-4067). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. http://ies.ed.gov/ncee/wwc/publications/practiceguides/.

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/.

[redacted]

[redacted]

[redacted]

[redacted]

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

Morrison, J. (2008/9). Why teachers must be data experts. *Educational Leadership, 66*(4). https://www.ascd.org/el/articles/why-teachers-must-be-data-experts

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.

Noyce, P.E., and Hickey, D.T. (Eds.). (2011). *New Frontiers in Formative Assessment*. Harvard Education Press.

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. & 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.

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.

Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (2018). Using learning and motivation theories to coherently link formative assessment, grading practices, and large-scale assessment. *Educational Measurement: Issues and Practice*, *37*(1), 21–34.

Shepard, L. A., Diaz-Bilello, E., Penuel, W. R., & Marion, S. F. (2020). *Classroom assessment principles to support teaching and learning.* Center for Assessment, Design, Research and Evaluation, University of Colorado Boulder. https://www.colorado.edu/cadre/sites/default/files/attached-files/classroom_assessment_principles_to_support_teaching_and_learning_-_final_0.pdf

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

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

[redacted]