Received: 1 March 2024 | Revised: 1 July 2024 | Accepted: 30 July 2024

DOI: 10.1002/sce.21905

RESEARCH ARTICLE

Science Education WILEY

Development of a questionnaire on teachers' beliefs, preparedness, and instructional practices for teaching NGSS science with multilingual learners

Scott E. Grapin[1] | Courtney Plumley[2] | Eric Banilower[2] | Alycia J. Sterenberg Mahon[3] | Laura Craven[2] | Kristen Malzahn[2] | Joan Pasley[2] | Abigail Schwenger[4] | Alison Haas[4] | Okhee Lee[4]

[1]University of Miami, Coral Gables, Florida, USA

[2]Horizon Research, Inc., Chapel Hill, North Carolina, USA

[3]Western Michigan University, Kalamazoo, Michigan, USA

[4]New York University, New York, New York, USA

Correspondence

Scott E. Grapin, Department of Teaching and Learning, School of Education and Human Development, University of Miami, Merrick Bldg 319-E, Coral Gables, FL 33146, USA.
Email: sgrapin@miami.edu

Abstract

The limited availability of research instruments that reflect the vision of the Next Generation Science Standards (NGSS) restricts the field's understanding of whether and how teachers are making instructional shifts called for by the standards. The need for such instruments is particularly urgent with teachers of multilingual learners (MLs), who are called on to make shifts in how they think about and enact instruction related to both science and language. The purpose of this study was to develop and gather validity evidence for a questionnaire that measures elementary teachers' beliefs, preparedness, and instructional practices for teaching NGSS science with MLs. We report on the development of the questionnaire over three phases that elicited multiple sources of validity evidence: (a) domain specification and expert review, (b) item writing and cognitive interviews, and (c) piloting and final item selection. Data included feedback from experts in science and language domains, cognitive interviews with 48 teachers, and a pilot with 310 teachers. Results indicated that the questionnaire differentiates among teachers with different levels of the underlying constructs and also that teachers' scores relate to their characteristics (e.g., familiarity with the NGSS). We highlight two implications for emerging research on NGSS-based instrumentation: (a) the difficulty of communicating with teachers about science and language instructional shifts while teachers are still developing their understanding of such shifts and (b) the potential of emerging NGSS-based instruments to inform professional development. We close with future directions for our research project specifically and the field of science education broadly.

KEYWORDS

elementary science, instrumentation, multilingual learners, teacher learning

The last decade has been characterized by efforts across the globe to reform science education for all students (Harris et al., 2019). In United States K-12 education specifically, A Framework for K-12 Science Education (National Research Council, 2012) and the Next Generation Science Standards (NGSS; Next Generation Science Standards Lead States, 2013) laid the foundation for contemporary approaches to science instruction in which students engage in three-dimensional learning that blends science and engineering practices (SEPs), crosscutting concepts (CCCs), and disciplinary core ideas (DCIs) to make sense of phenomena and build understanding coherently over time. Grounded in these contemporary approaches, research efforts in the science education community have coalesced around developing NGSS-designed curricula (e.g., Campbell & Lee, 2021; Reiser et al., 2021) and professional development (PD) programs that support teachers in implementing the curricula (e.g., Short & Hirsh, 2020). Together, curricula and PD programs comprise interventions that are currently being implemented in the education system at a large scale (e.g., Krajcik et al., 2023; Lowell & McNeill, 2023; Schneider et al., 2022). However, one challenge to understanding the impacts of such interventions is that NGSS-based instruments (e.g., questionnaires to measure teachers' preparedness, observation protocols to measure the quality of instruction) are still in their early stages of development and validation (e.g., Chen & Terada, 2021; Fulmer et al., 2021; Hayes et al., 2016).

The demand for instruments to measure the impacts of interventions is even more urgent in light of persistent inequities facing historically marginalized students in science education. Multilingual learners (MLs) are one such group that has been denied equitable opportunities to participate in science learning (Grapin et al., 2023; National Academies of Sciences, Engineering, and Medicine, 2018). In United States K-12 education specifically, one in five students reports speaking a language other than English at home (National Center for Education Statistics, 2024), and this fast-growing population of MLs includes students classified by their schools as English learners (ELs). While the NGSS present challenges for MLs, particularly given the language-intensive nature of SEPs (e.g., arguing from evidence; Hakuta et al., 2013; Lee et al., 2013), the standards hold promise for working toward equity. For example, as MLs engage in SEPs to make sense of phenomena in their classroom communities, they participate in meaningful interactions that cultivate their assets, including their everyday language and nonlinguistic modalities of communication (e.g., González-Howard & McNeill, 2016; Grapin, 2019; Lee, 2021; National Academies of Sciences, Engineering, and Medicine, 2018; Suárez, 2020).

Realizing the promise of the NGSS for MLs will require that teachers make shifts in how they think about and enact science instruction. Teachers need to think about science instruction in ways that reflect contemporary approaches (*beliefs*), feel prepared to engage in such instruction (*preparedness*), and enact such instruction in their classrooms (*instructional practices*). For example, teachers need to believe that MLs are capable of contributing ideas in the science classroom even with emerging English proficiency, feel prepared to engage MLs in SEPs (e.g., arguing from evidence), and enact instructional practices that support students in doing so (e.g., providing scaffolds to support MLs' engagement in argumentation). Paying attention to teachers' beliefs, preparedness, and instructional practices for teaching science with MLs is particularly crucial in United States elementary classrooms, where science instruction has not been prioritized (e.g., Banilower et al., 2018; National Academies of Sciences, Engineering, and Medicine, 2022) while many MLs are concentrated in these grades (e.g., National Center for Education Statistics, 2024). However, the field of science education currently lacks instruments to measure such constructs.

The purpose of this study was to develop and gather validity evidence for a questionnaire on elementary teachers' beliefs, preparedness, and instructional practices for teaching NGSS science with MLs. The study is significant in multiple respects. First, our questionnaire could be used or adapted in impact studies of NGSS-based interventions, particularly in linguistically diverse science learning contexts. Second, lessons learned from the development and validation process, as documented in detail in this study, could inform the development and validation of other instruments for studying the impacts of NGSS-based interventions and reform-oriented educational interventions broadly. Finally, in addition to being essential for rigorously studying interventions, instrumentation work can contribute to operationalizing constructs in ways that are meaningful to teachers and inform the design of PD experiences.

In this article, first, we review the emerging literature on NGSS-based instrumentation, with a focus on instruments that measure teacher constructs. Then, we describe the research context in which our instrumentation work was carried out. Next, we describe the questionnaire development and validation process. Finally, we discuss implications of the instrument and the process of developing and validating it as well as future research directions.

1 | LITERATURE REVIEW

Although the NGSS were released over a decade ago, instruments for measuring the standards' impacts are only beginning to emerge, in part because the last decade has seen a focus on curriculum design (e.g., Campbell & Lee, 2021; Reiser et al., 2021) and, more recently, curriculum-based PD (e.g., Lee et al., 2023; Lowell & McNeill, 2023). Emerging instruments elicit information about multiple aspects of contemporary science instruction, including teachers' knowledge, beliefs, and practices (e.g., Fulmer et al., 2021; Hayes et al., 2016; Nollmeyer & Bangert, 2017), classroom instruction (e.g., Chen & Terada, 2021; Martínez et al., 2022), and students' experiences (e.g., Campbell, Lee, Longhurst, et al., 2021; Penuel et al., 2023). These instruments employ a variety of data collection methods, including questionnaires (e.g., Fulmer et al., 2021), observation protocols (e.g., Chen & Terada, 2021), and electronic portfolios (e.g., Martínez et al., 2022). While each method has affordances and limitations (Desimone, 2009), questionnaires are particularly useful for eliciting information about aspects of teacher learning that are not directly observable (e.g., beliefs and perceptions of preparedness) as well as for collecting data from large samples of teachers about what they report doing in their instruction over a period of time (e.g., frequency of instructional practices).

Given the focus of our questionnaire on teachers' beliefs, preparedness, and instructional practices, we review the emerging literature on questionnaires that employ Likert-style items to address teacher constructs. Moreover, given our focus on MLs, we review studies across science education and language education. For each study, we provide key details regarding constructs, items, and sources of validity evidence.

1.1 | Studies in science education

Banilower and colleagues' research (Banilower et al., 2013, 2018) with the National Survey of Science and Mathematics Education (NSSME) has occurred over multiple iterations. The most recent iteration of the survey (Banilower et al., 2018) measured, among other constructs, science teachers' beliefs, preparedness, and instructional practices related to reform-oriented science. The survey underwent rigorous development and validation that involved collecting multiple sources of validity evidence, including evidence based on test content (e.g., domain specification and expert review), evidence based on response process (e.g., cognitive interviews with teachers), and evidence based on internal structure (e.g., factor analysis). After administering the survey to a national probability sample of science, math, and computer science teachers in grades K-12 in the 50 states and the District of Columbia, results were analyzed by equity factors, such as the percentage of students in each teacher's class from racial and ethnic groups historically underrepresented in STEM. As our review of the literature makes evident, Banilower and colleagues' research with the NSSME and other instruments (e.g., Smith et al., 2014) has been a crucial resource for instrumentation efforts in the field of science education. Our instrumentation work, as reported in this manuscript, builds on Banilower and colleagues' (2018) most recent iteration of the NSSME while extending to address constructs related to teaching MLs in science.

Hayes et al. (2016) designed and validated a survey to measure science teachers' instructional practices, including inquiry-based practices based on previous science standards as well as practices related to SEPs in the NGSS. The 31-item survey, which was tested with 397 science teachers in Grades 3–10, used or adapted items from previously developed instruments, particularly the NSSME (Banilower et al., 2013). Hayes et al. reported multiple sources of validity evidence, including evidence based on test content, response process, and internal structure. By bridging traditional and NGSS-based instructional practices, Hayes et al.'s work represents an important step toward developing instruments useful for measuring teachers' instructional practices in the context of the NGSS, specifically related to SEPs.

Other instruments have targeted constructs related to teachers' knowledge and beliefs about language in science instruction. Fulmer et al. (2021) developed and validated a 15-item questionnaire to measure teachers' knowledge of language as an epistemic tool (e.g., knowing that language is essential to learning science, knowing that language includes multimodal representations). Their instrument development and validation work involved 158 in-service and preservice teachers and was motivated by the need to gauge teachers' readiness to foster "epistemically rich classroom environments that match the vision of the NGSS" (Fulmer et al., 2021, p. 461). Similar to the studies above, Fulmer et al. reported multiple sources of validity evidence, including evidence based on test content, response process, and internal structure. Fulmer et al.'s instrument provides a foundation for future research aimed at "studying effects of professional development—on teachers' knowledge, then to classroom practices, and then to student outcomes" (p. 461).

Most recently, Lowell and McNeill (2023) conducted a longitudinal investigation of the teaching and self-efficacy beliefs of 322 in-service middle school teachers who participated in NGSS curriculum-based PD. Lowell and McNeill's survey consisted of 15 items adapted from previous studies (e.g., Reiser et al., 2017), including the NSSME (Banilower et al., 2013). Although most survey items addressed teachers' beliefs about science instruction broadly (e.g., asking whether it is better for science instruction to focus on science ideas in depth rather than cover more science ideas), some items addressed issues related to language specifically. For example, an item asked whether "at the beginning of instruction on a science idea, students should be provided with definitions for new scientific vocabulary that will be used" (Lowell & McNeill, 2023, p. 1468). Lowell and McNeill did not report validity evidence, arguing instead that "evidence for validity includes that the instrument has been used in past studies to measure the same construct being measured in this study" (p. 1467).

1.2 | Studies in language education

There is a long history of investigating teachers' beliefs about language and language learners. This focus on beliefs (and related constructs, such as 'attitudes' and 'ideologies') has been motivated by concerns that teachers hold beliefs about language derived from their personal and professional experiences that are likely to influence their instructional practices with MLs (e.g., Borg, 2018). Two prominent instruments are the Language Attitudes of Teachers Scale (LATS; Byrnes & Kiger, 1994) and the Beliefs About Language Learning Inventory (BALLI; Horwitz, 1985). Over time, these instruments have been adapted by researchers to study teachers' attitudes and beliefs about language and language learners in STEM subjects and other content areas.

Huerta et al. (2019) adapted LATS (Byrnes & Kiger, 1994), along with existing surveys in science education (e.g., Lee, 2004), to measure in-service teachers' attitudes toward ELs in science instruction. For example, their survey asked teachers whether they believed "incorporating ELs' culture and background [would] help them learn during science instruction" and whether "teaching a set of vocabulary words intensively across several days [would] help ELs learn during science instruction" (p. 6). Huerta et al. administered the 13-item survey to 553 in-service teachers and reported validity evidence based on test content and internal structure. While the survey did not target teachers' instructional practices, a key motivation for Huerta et al.'s study was the potential for attitudes to "influence teacher practice" (p. 2).

Similarly, Accurso (2020) adapted BALLI (Horwitz, 1985) to measure changes in preservice teachers' beliefs about language and language learning in content areas (math, science, social studies, language arts). Whereas Huerta et al.'s (2019) survey above was specific to science and focused on teaching ELs, Accurso's (2020) survey was not specific to any content area and focused on teaching language broadly. For example, Accurso's survey asked teachers whether they believed "students need to be explicitly taught the English they need for learning different subjects" (p. 11). Accurso administered the 24-item survey to 18 preservice teachers before and after their participation in a teacher education course focused on teaching language across content areas. Similar to Lowell and McNeill (2023) in science education, Accurso did not report validity evidence but instead emphasized the use of the survey items in prior research. Interestingly, Accurso also examined how teachers enacted their beliefs as instructional practices, specifically, how they provided feedback on a sample math explanation, although teachers' instructional practices were not a construct targeted by the instrument.

1.3 | Summary

Our review of the literature across science education and language education indicates contributions of prior research as well as areas ripe for further development. In science education, studies on instrumentation in the context of the NGSS are only beginning to emerge. The small number of studies that exist target a variety of constructs, including teachers' beliefs, knowledge, preparedness, and instructional practices. However, existing instruments (e.g., Hayes et al., 2016) tend to focus on a single construct (e.g., teachers' instructional practices but not their beliefs) and only some aspects of NGSS-based instruction (e.g., instruction related to SEPs but not other dimensions of science learning, such as DCIs and CCCs). While instruments have begun to address issues related to language specifically (e.g., Fulmer et al., 2021; Lowell & McNeill, 2023) and equity broadly (e.g., Banilower et al., 2018) in contemporary science instruction, none explicitly address MLs, a fast-growing student population for whom the NGSS present both opportunities and challenges (Lee et al., 2013).

In language education, researchers have adapted instruments to measure teachers' attitudes and beliefs about language and language learners across content areas, including STEM subjects. However, some of these instruments focus on content areas broadly (e.g., Accurso, 2020), while others focus on science instruction specifically but do not necessarily reflect contemporary approaches based on the NGSS (e.g., Huerta et al., 2019). Also, these instruments tend to target teachers' beliefs or attitudes that are likely to influence their instructional practices but do not target instructional practices themselves or teachers' preparedness to implement them. This, in turn, restricts the field's efforts to develop "a complex portrait of content teaching for MLs" that considers multiple constructs and their interrelations (Viesca et al., 2019, p. 304).

Finally, across science and language education, our review indicates substantial variation in the amount and sources of validity evidence reported in the studies. Given these contributions and limitations of previous instrumentation work, we sought to develop and gather robust validity evidence for a questionnaire that measures teachers' beliefs, preparedness, and instructional practices for teaching NGSS science with MLs.

2 | RESEARCH CONTEXT OF THE STUDY

This study was carried out in the context of a larger research project focused on conducting a multiyear quasi-experimental field trial of an NGSS-designed curriculum in a linguistically diverse urban school district. The Science And Integrated Language (SAIL) curriculum is a yearlong fifth-grade science curriculum for all students with a focus on MLs (Lee et al., 2019). The SAIL curriculum consists of four units that address all 16 performance expectations in the NGSS for fifth grade. The first unit in the curriculum was awarded a Badge of Distinction by Achieve, Inc. (https://www.nextgenscience.org/resources/grade-5-sail-garbage-unit), indicating the highest quality in NGSS instructional materials design, and was also featured in the consensus report *English Learners in STEM Subjects* (National Academies of Sciences, Engineering, and Medicine, 2018). The curriculum is accompanied by a curriculum-based PD program that prepares teachers to implement the curriculum and develops teachers'

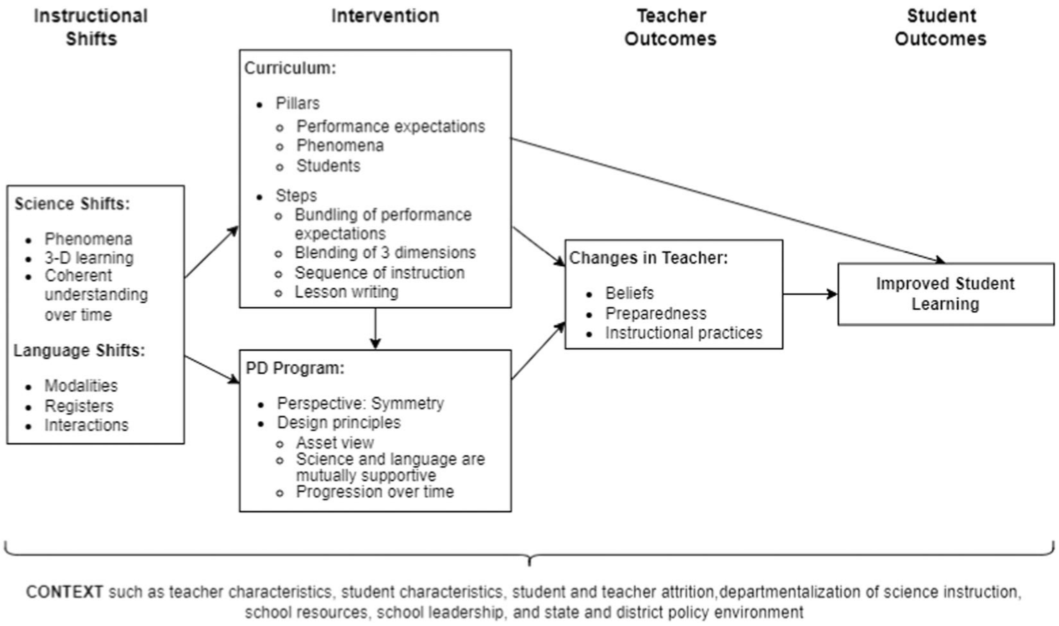
Instructional Shifts
Intervention
Teacher Outcomes
Student Outcomes
Science Shifts:
• Phenomena
• 3-D learning
• Coherent understanding over time
Language Shifts:
• Modalities
• Registers
• Interactions
Curriculum:
• Pillars
○ Performance expectations
○ Phenomena
○ Students
• Steps
○ Bundling of performance expectations
○ Blending of 3 dimensions
○ Sequence of instruction
○ Lesson writing
PD Program:
• Perspective: Symmetry
• Design principles
○ Asset view
○ Science and language are mutually supportive
○ Progression over time
Changes in Teacher:
• Beliefs
• Preparedness
• Instructional practices
Improved Student Learning
CONTEXT such as teacher characteristics, student characteristics, student and teacher attrition,departmentalization of science instruction, school resources, school leadership, and state and district policy environment

FIGURE 1 Theory of change.

understanding of both the conceptual and practical aspects of teaching NGSS science with MLs (Lee et al., 2023). Together, the curriculum and PD program comprise our full intervention.

A primary goal of the larger research project is to study the intervention's impacts on both teachers and students. Figure 1 displays the project's theory of change. Beginning from the left side of the figure, the project is conceptually grounded in science and language instructional shifts (see Lee et al., 2019, and the next section). These shifts guided the development of the intervention, including the curriculum (Haas et al., 2021) and the curriculum-based PD program (Lee et al., 2023). As teachers engage with the curriculum and participate in the PD program, it is expected they will develop beliefs, preparedness, and instructional practices for teaching NGSS science with MLs (see 'Teacher Outcomes' in Figure 1). Changes in teachers' beliefs, preparedness, and instructional practices, along with students' direct engagement with the curriculum, are expected to result in improved student learning (see 'Student Outcomes' in Figure 1). The theory of change highlights multiple contextual factors that are likely to mediate the impacts of the intervention, particularly in the context of a large urban school district (see 'Context' at the bottom of Figure 1).

To study the intervention's impacts on teachers and students, our research team has been developing a suite of instruments, including a teacher questionnaire, classroom observation protocol, and student science assessment. In this manuscript, we report on the development and validation of the teacher questionnaire, which targets three constructs from the project's theory of change: (a) teacher beliefs, (b) teacher preparedness, and (c) teacher instructional practices (see 'Teacher Outcomes' in Figure 1).

Instructional practices have been a major focus across fields, including science education (e.g., Windschitl et al., 2012), language education (e.g., Peercy et al., 2023), and education research broadly (e.g., Loewenberg Ball & Forzani, 2009), because of the potential of such practices to directly impact student learning. *Beliefs* have also received a great deal of attention across fields, including science education (e.g., Jones & Park, 2023), language education (e.g., Borg, 2018), and education research broadly (e.g., Pajares, 1992), because of their potential to influence teachers' instructional practices. *Preparedness* has become a central focus of policy on teacher learning across science education (e.g., National Academies of Sciences, Engineering, and Medicine, 2021), language education (e.g., National Academies of Sciences, Engineering, and Medicine, 2018), and education research broadly (e.g., National Academies of Sciences, Engineering, and Medicine, 2020), particularly in light of concerns regarding teachers' preparedness for elementary science instruction and instruction with MLs. In the context of intervention work, a focus on preparedness can also help unpack why teachers enact (or not) instructional practices. For example, a teacher may believe an instructional practice is important but not feel prepared to enact it. As a result, the teacher may not enact the practice frequently or at all. In contrast, another teacher may believe the practice is important and feel prepared to enact it but not do so frequently for other reasons (e.g., contextual factors). By targeting multiple constructs, our questionnaire could offer a more comprehensive and nuanced picture of the intervention's impacts than would be possible by targeting any construct alone.

Each construct (beliefs, preparedness, and instructional practices) is operationalized in two domains that reflect the intervention's focus on NGSS science and MLs: (a) teaching NGSS science (hereafter the 'science domain') and (b) teaching MLs in science (hereafter the 'language domain'). Thus, the questionnaire consists of six scales total (i.e., three constructs x two domains):

1. Beliefs about Teaching NGSS Science (hereafter 'Sci Beliefs')
2. Beliefs about Teaching MLs in Science (hereafter 'Lang Beliefs')
3. Preparedness for Teaching NGSS Science (hereafter 'Sci Prep')
4. Preparedness for Teaching MLs in Science (hereafter 'Lang Prep')
5. Instructional Practices for Teaching NGSS Science (hereafter 'Sci Practices')
6. Instructional Practices for Teaching MLs in Science (hereafter 'Lang Practices')

The questionnaire serves multiple purposes for our research project. First, it enables us to examine the intervention's impact on teachers, for example, how their beliefs, preparedness, and instructional practices in science and language domains change (or not) over time. Second, the questionnaire could help unpack the intervention's impact on students, for example, whether improvements (or not) in student learning are mediated by teachers' reported instructional practices.

3 | QUESTIONNAIRE DEVELOPMENT AND VALIDATION

Our development and validation work is grounded in a perspective on validity as "the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores" (Messick, 1989a, p. 6; see also Messick, 1989b). From this perspective, validity is a unitary concept consisting of multiple facets that contribute to an "integrative evaluative judgement" (Messick, 1989a, p. 10). In contrast with perspectives on validity that emphasize distinct types (e.g., content validity, criterion validity), this perspective frames construct validity as the comprehensive goal of validation work that subsumes multiple facets within a "unified faceted framework" (Messick, 1989a, p. 10). In other words, "all validation is construct validation" (Messick, 1989a, p. 8).

This unitary yet multifaceted perspective on validity undergirds the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), which are grounded in "validity [as] a unitary concept" (p. 14). Specifically, the Standards articulate multiple sources of validity evidence that "support the intended interpretation of test scores for specific uses" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, p. 21): evidence based on test content, evidence based on response process, evidence based on internal structure, evidence based on relations to other variables, and evidence based on consequences.

We developed our questionnaire over three phases that elicited multiple sources of validity evidence to support the use of the instrument for its purpose, specifically, to measure elementary teachers' beliefs, preparedness, and instructional practices for teaching NGSS science with MLs. The first phase, *domain specification and expert review*, elicited evidence that the specifications adequately represented each domain (i.e., evidence based on test content). The second phase, *item writing and cognitive interviews*, elicited evidence that teachers interpreted the items as intended (i.e., evidence based on response process). The third phase, *piloting and final item selection*, elicited evidence that each scale provided a reliable measure that differentiated among teachers with different levels of the underlying construct (i.e., evidence based on internal structure). Exploratory analyses of pilot data also elicited evidence of the relation between teachers' beliefs, preparedness, and instructional practices, as measured by the scales, and teacher characteristics, such as their K-12 teaching experience and familiarity with the NGSS (i.e., evidence based on relations to other variables). Thus, similar to Fulmer et al. (2021), our initial development and validation work focused on sources of validity evidence internal to the questionnaire, while future research could build on this work to gather evidence based on additional sources (e.g., relation between scores on the questionnaire and external variables; see 'Future Research Directions').

3.1 | Phase 1: Domain specification and expert review

Domain specification typically involves reviewing standards documents and relevant literature as well as soliciting input from domain experts (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). We drew from multiple and varied sources to define the science and language domains of our questionnaire. Domain specifications for science and language provided the foundation for operationalizing teachers' beliefs, preparedness, and instructional practices within each domain.

First, we reviewed consensus documents on contemporary science education (e.g., *A Framework for K-12 Science Education* by National Research Council, 2012) and teaching MLs in science (e.g., *English Learners in STEM Subjects* by National Academies of Sciences, Engineering, and Medicine, 2018). Second, we drew on relevant literature. Specifically, we conducted two focused literature reviews: one on contemporary science education with MLs (Buxton & Lee, 2023) and another on teacher learning to implement the NGSS with all students and MLs in particular (Lee et al., 2023). Finally, we drew heavily from our own research programs that have made contributions to science and language domains in terms of both conceptualization and operationalization. In terms of conceptualization, we have developed an extensive research program focused on investigating multiple facets of contemporary science education with MLs, including teacher learning (e.g., Lee et al., 2019, 2023). In terms of operationalization, as described earlier, we have engaged in extensive instrumentation work focused on measuring teacher constructs related to reform-oriented science instruction (e.g., Banilower et al., 2013, 2018).

Our review of these multiple and varied sources indicated three areas within each domain. For the science domain, the three areas are (a) phenomena, (b) three-dimensional learning, and (c) coherence. For the language domain, the three areas are (a) modalities, (b) registers, and (c) interactions, which are situated within two framing ideas (asset-oriented view of MLs and classroom environment that promotes productive discourse). As these areas reflect consensus documents in science education (e.g., National Research Council, 2012) and science education with MLs (e.g., National Academies of Sciences, Engineering, and Medicine, 2018), we provide a brief overview of each area, including references to relevant literature. Given that the NGSS call for significant instructional shifts that have spurred parallel instructional shifts in science teaching with MLs (Lee et al., 2019), we describe each area in terms of a shift from traditional to contemporary approaches. For each shift, we offer examples of what it might look like for teachers to enact the shift in their instruction.

3.1.1 | Science domain: Phenomena, three-dimensional learning, and coherence

Phenomena refer to observable events in the natural world. Whereas traditional approaches in science education focused on canonical science knowledge and used phenomena as a hook or extension, contemporary approaches emphasize anchoring instruction in phenomena that students are compelled to figure out and that are sustained over the course of a unit (e.g., Krajcik & Czerniak, 2018). Moreover, contemporary approaches emphasize making phenomena meaningful and relevant to students' lives (e.g., Lee et al., 2019). For example, teachers use phenomena as the basis for instruction and make connections between phenomena and students' backgrounds and interests.

Three-dimensional learning encompasses three dimensions of science learning: SEPs, DCIs, and CCCs (National Research Council, 2012). Whereas traditional approaches in science education emphasized both science inquiry and science content but not always their integration, contemporary approaches emphasize blending SEPs, DCIs, and CCCs to explain phenomena (e.g., Harris et al., 2019). For example, teachers provide scaffolding as students engage in three-dimensional learning and identify evidence of this learning in students' work.

Coherence refers to how instruction is connected over time, including over a unit, over a year, and over multiple years. Whereas traditional approaches in science education consisted of individual activities or investigations that were not always connected, contemporary approaches emphasize organizing learning experiences that are driven by students' questions about phenomena and that support students in developing their science understanding coherently over time (e.g., Reiser et al., 2021). For example, teachers situate lessons in terms of what students have already figured out and what they still need to figure out to make sense of a phenomenon. Teachers also help students develop increasingly sophisticated understanding of DCIs and more independent use of SEPs and CCCs over time.

3.1.2 | Language domain: Modalities, registers, and interactions

Modalities refer to the multiple and varied channels through which communication occurs (Kress et al., 2014), including oral, written, physical, pictorial, gestural, symbolic, and computational modalities. Whereas traditional approaches to teaching MLs in science tended to privilege language and use nonlinguistic modalities as temporary scaffolds (see Grapin, 2019, for a critique), contemporary approaches emphasize leveraging the affordances of modalities for engaging in SEPs (e.g., developing and using models) and representing science ideas (e.g., Pierson et al., 2021; Suárez & Otero, 2023). For example, teachers provide opportunities for students to communicate science ideas using multiple modalities and guide students to consider how different multimodal representations (e.g., diagram vs. computer simulation) can be strategically leveraged to represent different aspects of a science idea.

Registers refer to ways of using language for particular contexts or purposes (Valdés et al., 2014). Whereas traditional approaches to teaching MLs in science tended to treat the specialized science register as a precursor or prerequisite to learning science (e.g., by preteaching science vocabulary at the beginning of a lesson), contemporary approaches emphasize building on the assets that MLs bring to the science classroom, including their everyday language (e.g., Brown & Ryoo, 2008; Lee et al., 2019), and supporting them to develop the specialized science register as a *product* of learning science (National Academies of Sciences, Engineering, and Medicine, 2018). For example, teachers encourage students to use their everyday language to make sense of phenomena. Teachers also provide scaffolds at the word level (e.g., introducing vocabulary in context), sentence level (e.g., recasting students' contributions to discussions), and discourse level (e.g., providing graphic organizers for argument writing) to help students communicate science ideas with greater precision.

Interactions refer to the settings and participants involved in communication (Bunch, 2014), for example, one-to-one interactions (e.g., one student talking to a partner) and one-to-many interactions (e.g., one student talking to the whole class). Whereas traditional approaches to teaching MLs in science engaged students in interactions that often lacked a clear purpose, contemporary approaches emphasize engaging MLs in a range of meaningful, goal-directed interactions in a community of practice (e.g., Gibbons, 2015) and guiding them to adapt their language to meet the communicative demands of those interactions (e.g., more explicit language use in one-to-many interactions that lack a shared frame of reference; National Academies of Sciences, Engineering, and Medicine, 2018). For example, teachers use various participation structures in the science classroom and encourage students to use language explicitly as they communicate with more "distant" audiences.

In addition to the three areas of modalities, registers, and interactions, our review surfaced two framing ideas in the language domain. These framing ideas, which figured prominently in the *English Learners in STEM Subjects* consensus report (National Academies of Sciences, Engineering, and Medicine, 2018), create the conditions for enacting the instructional shifts described above related to modalities, registers, and interactions. The first idea involves adopting an asset-oriented view of MLs as bringing knowledge and experiences that can be leveraged during science instruction (see, in particular, Chapter 1 of National Academies of Sciences, Engineering, and Medicine, 2018). For example, teachers need to believe that MLs are capable of making valuable contributions to science discussions even with emerging English proficiency. The second idea involves creating a classroom environment that engages students in productive discourse with others (see, in particular, Chapter 4 of National Academies of Sciences, Engineering, and Medicine, 2018). For example, teachers explore students' ideas even if they are not complete and encourage students to expand on their thinking and that of their peers. These two framing ideas (which we referred to as 'MLs as learners' and 'Positive classroom environment' in the language domain specifications) interrelate since "teachers are crucial to creating classroom environments that can leverage the assets that ELs bring" (National Academies of Sciences, Engineering, and Medicine, 2018, p. 89).

3.1.3 | Expert review

After developing draft specifications for the science and language domains, we solicited feedback on the specifications from four domain experts across science education and language education. The two domain experts in science education, respectively, had expertise in NGSS-based instrument development and teacher learning. The two domain experts in language education, respectively, had expertise in NGSS science instruction and assessment with MLs and language and literacy instruction with MLs in the context of content standards broadly. All four experts were advisory board members for the larger project.

Given the experts' complementary areas of specialization, we shared the draft science specifications with the two experts in science education and the draft language specifications with the two experts in language education. Three questions were provided to guide their reviews: (a) Is there anything missing from the specifications (e.g., important elements of the domain that are not represented)?, (b) Are there elements that do not accurately represent current thinking about the domain and either require revision or do not belong?, and (c) Is there anything that requires further clarification (e.g., ideas you do not understand)? Some experts provided feedback with narrative responses to the questions, while others made comments and revisions directly on the specification documents.

The domain experts were largely in agreement that the specifications accurately represented current thinking in their respective domains. They also highlighted the need for clarification or elaboration in certain areas. In particular, both experts in science education noted that the focus on coherence could have been represented more saliently. While coherent instruction was addressed in the draft specifications, it was embedded within each of the three dimensions of science learning (e.g., in the section on SEPs, 'Students use and reflect on SEPs in the context of different phenomena to build their ability to use SEPs more independently over time') and was therefore not as salient. In response to this feedback, we reorganized the science domain to address coherence as its own section. Other comments focused on clarifying the meaning of specific words or phrases. For example, one of the experts asked for clarification regarding what it meant for students to use SEPs and CCCs "more independently over time," which we clarified, based on team discussion, as "becoming more adept and sophisticated with using SEPs and CCCs and with less teacher guidance."

The experts in language education also highlighted areas that required clarification or elaboration. For example, one expert indicated that, while the specifications related to modalities focused on what *students* were engaged in doing (e.g., using multiple modalities to represent their thinking), the specifications could have been more detailed about what exactly the *teacher* does to support students' proficiency with modalities. Thus, we revised the specifications to describe what the teacher does (e.g., 'Highlight how different modalities represent the same science idea'). Other comments focused on the "Positive classroom environment" section. In sharing the draft specifications, we had noted for reviewers that we were debating whether this should remain its own section or be embedded throughout the domain, as a positive classroom environment is relevant to all three areas of the language domain (i.e., a positive classroom environment is important for promoting modalities, registers, and interactions) and also overlaps with the science domain (e.g., a positive classroom environment is important for engaging students in SEPs, such as arguing from evidence). One domain expert supported keeping this section separate to maintain its salience, as a positive classroom environment is "a critical piece of creating a language-rich science classroom." Thus, we kept "Positive classroom environment" as its own section while adding an annotation indicating its relevance across other areas.

Table 1 displays the revised domain specifications in one area from the science domain (i.e., phenomena) and one area from the language domain (i.e., modalities) following expert review. These specifications, which include elements (indicated by numbers) as well as sub-elements (indicated by letters), became the basis for writing items to operationalize teachers' beliefs, preparedness, and instructional practices within each domain (described in the next section). Consistent with the iterative nature of our development process, drafting items based on the specifications led to clarification and elaboration of the specifications themselves.

TABLE 1 Sample domain specifications for science domain (phenomena) and language domain (modalities).

Phenomena: observable events in the natural world	1. Instruction should be driven by phenomena that provide opportunities for students to engage in three-dimensional learning (blending SEPs, DCIs, and CCCs). a. Students engage with phenomena that are investigable in the classroom. b. Students engage with phenomena in ways that are developmentally appropriate. c. Students engage with phenomena that are (or are made) relevant (e.g., interests, experiences, cultures).
Modalities: multiple and varied channels through which communication occurs	1. Provide purposeful opportunities for students to use different modalities (i.e., oral, written, physical, pictorial, gestural, symbolic) to represent and communicate their thinking and understanding. a. Use tasks/activities that allow for multiple modalities. b. Provide specific tools (e.g., computer simulations, physical models, drawing tools, graphic organizers) that allow students to represent and communicate ideas in multiple modalities. c. Give options for students to use different modalities to represent and communicate their thinking and understanding. 2. Model the use of different modalities to represent and communicate thinking and understanding. 3. Support students' proficiency in using multiple modalities. a. Help students use modalities strategically (e.g., consider affordances and limitations of each modality). b. Highlight how different modalities represent the same science idea.

3.2 | Phase 2: Item writing and cognitive interviews

We used the two sets of domain specifications (science and language) to generate items related to each of the three constructs targeted by the questionnaire (beliefs, preparedness, and instructional practices). Once a sizable number of items were generated, the researchers began cognitive interviews, which led to multiple rounds of revising existing items, writing new items, and further interviewing.

3.2.1 | Item writing

Two teams of researchers (one for the science domain and one for the language domain, with four researchers on one team and five on the other) drafted items aligned with the domain elements. The researchers wrote items individually and then met for "item camps" during which they reviewed and edited items for their domain. A tracking system was used to keep record of previous versions of items and make note of words or phrases that were either problematic or should be used consistently throughout the questionnaire.

The question stems and response options that guided item writing for different constructs appear in Table 2. Although in the larger research project we use the broader and more asset-oriented term "multilingual learner" (see González-Howard & Suárez, 2021, and Grapin, 2021, on terminology), in the question stems and items, we use "English learner," as this term is adopted in United States federal policy (U.S. Department of Education, 2015) and therefore widely used in the education system. Based on our prior survey development and validation work

TABLE 2 Question stems and response options for different constructs.

Construct(s)	Question stem	Response options
Sci Beliefs	Practical constraints aside, to what extent do you agree with each of the following statements about science instruction?	• Strongly disagree • Disagree • Slightly disagree • Slightly agree • Agree • Strongly agree
Lang Beliefs	Practical constraints aside, to what extent do you agree with each of the following statements about English learners (ELs) and science instruction?	• Strongly Disagree • Disagree • Slightly disagree • Slightly agree • Agree • Strongly agree
Sci Prep Lang Prep	How well prepared do you feel to do each of the following in your science instruction?	• Not at all • Somewhat • Fairly well • Very well
Sci Practices Lang Practices	Last school year, how often did you do each of the following in your science instruction?	• Never • Rarely—a few times a year • Sometimes—once or twice a month • Often—once or twice a week • All or almost all lessons

Note: The time period specified for Sci Practices and Lang Practices ("Last school year") changes based on when the questionnaire is administered. This phrasing was used in the cognitive interviews, which were conducted during the summer months.

(e.g., Banilower et al., 2018), we used Likert scales with different response options and numbers of options for each construct (six options for beliefs, four options for preparedness, five options for instructional practices).

During item writing, we realized that not all elements of the domain specifications were relevant to or feasible for each construct. For example, elements of the science domain that focused on coherence over time (e.g., engaging students in making connections across DCIs over time) did not lend themselves to instructional practices items in which teachers were asked about the frequency with which they used each practice (e.g., once or twice a month, once or twice a week). Moreover, for the beliefs construct, some items were likely to elicit overwhelming agreement or disagreement among teachers (e.g., asking teachers whether they believed in establishing a classroom culture that encourages all students to participate) and thus were written only for the preparedness construct (e.g., asking teachers whether they felt *prepared* to establish such a classroom culture).

3.2.2 | Cognitive interviews

Cognitive interviews are useful for examining whether items are interpreted by respondents as intended (Desimone & Le Floch, 2004). Ensuring teachers' meaningful interpretation of the items was particularly challenging in our study that involved teachers at the elementary level, who typically have less preparation for and experience with science teaching compared to their secondary counterparts (Banilower et al., 2018). A total of 48 teachers who were teaching science in Grades 4, 5, or 6 (i.e., the grades most relevant to teachers in the fifth-grade intervention) were recruited through the mailing list of Horizon Research, Inc. to participate in cognitive interviews. During the interviews, teachers were asked to read each item aloud and describe how they would answer if they encountered the item on a survey. The researchers used follow-up prompts and questions to gauge teachers' overall interpretation of the items (e.g., 'Why did you choose that response?' and 'Can you give examples of how you might do that in your own classroom?') and to identify specific areas of difficulty or confusion (e.g., 'Please identify any words/terms that you did not understand' and 'Describe anything that made this item confusing or difficult to answer.').

The researchers recorded teachers' responses to the questions using an audio recorder as well as written notes. Interviews lasted about 45 min during which teachers and researchers discussed as many of the items as they could (often 20–30 items). Each of the items was addressed in at least three interviews to get multiple perspectives. Once a critical mass of items had three interviews, the researchers came together for another "item camp" in which they shared how the teachers responded to each item. Items that the teachers interpreted as intended were deemed ready for the pilot. Items that the teachers did not interpret as intended were either revised or dropped (see below for details of teachers' interpretations). The revised items became the focus of subsequent rounds of interviewing. There were four rounds of interviews about items in the science domain and three rounds of interviews about items in the language domain. In total, the 48 teachers participated in 82 interviews (46 science, 34 language, 2 with items from both domains) about 724 items (452 science, 272 language).

The cognitive interviews revealed words and phrases that teachers interpreted in varying ways. Below, we provide examples from the three areas in the science domain (phenomena, three-dimensional learning, and coherence) and the three areas in the language domain (modalities, registers, and interactions). For clarity and consistency of reporting, the examples are contextualized within the instructional practices construct unless otherwise specified, though the issues are also relevant to beliefs and preparedness. Table 3 shows original and revised items for each area, including examples from all three constructs (beliefs, preparedness, and instructional practices). We dedicate substantial space to describing teachers' interpretations, which can inform the development of other instruments that similarly target contemporary approaches to science instruction while teachers are still developing their understanding (and associated terminology for describing) such approaches.

3.2.2.1 | *Phenomena*

In the interviews, teachers interpreted "phenomena" in a variety of ways, including "field trips," "hands-on learning," and "learning that is engaging." After trying out multiple other phrasings (e.g., 'naturally occurring events'), we decided on "science-related events or processes," which most teachers interpreted as intended. To further clarify the meaning of this phrase, we accompanied it with examples. Initially, the examples consisted of "weather patterns, plant growth, [and] species extinction." However, we revised the examples to convey events or processes happen*ing* (e.g., 'plants growing' instead of 'plant growth'), which avoided confusion with DCIs (e.g., life science idea of plant growth) and more closely reflected the kinds of phenomena used in contemporary science curricula.

3.2.2.2 | *Three-dimensional learning*

Before the interviews, we had anticipated that teachers from different states and districts would have varying levels of familiarity with the NGSS and thus would not necessarily be familiar with the three dimensions. Whereas "core science ideas" (used to describe DCIs) were generally interpreted by teachers as intended, the other two dimensions (SEPs and CCCs) proved more challenging. For SEPs, we initially used "practices of science" without any examples. However, teachers tended to associate this phrase with more traditional approaches to science instruction that emphasized "experiments" and "hands-on activities." Thus, we accompanied "practices of science" with examples of SEPs from the NGSS, including SEPs that would be more familiar to teachers (e.g., planning an investigation) as well as those that would be less familiar (e.g., developing models, arguing from evidence). For CCCs, we tried multiple phrases, including "overarching science concept," "lens," and "intellectual tool." However, teachers indicated that "overarching science concept" and "lens" were too broad, while they tended to associate "intellectual tool" with physical tools, such as graphic organizers. Ultimately, we settled on using "approach" followed by descriptions of

TABLE 3 Original items and revised items based on cognitive interviews.

Original item	Revised item
Science domain	
Phenomena (preparedness construct)	
Help students connect phenomena to their own experiences	Help students connect the science-related event or process that they are studying (for example: weather changing, plants growing, a species going extinct) to their own experiences
Three-dimensional learning (instructional practices construct)	
Make a connection between students' prior experiences and knowledge and the practices of science they are using to answer a scientific question	Make connections between students' prior experiences and knowledge and the practices of science they are using to explore a scientific question, such as planning investigations, developing models, arguing with evidence
Coherence (beliefs construct)	
Students learn science best when instruction allows them more independence over time in selecting an appropriate intellectual tool (for example: cause and effect, patterns, or structure and function) for exploring and communicating about a science-related event or process such as evaporation, plant growth, and erosion.	As students move up in grade level, they should be given more choice in how to approach a scientific question (for example: looking for patterns, defining and analyzing systems, examining cause and effect).
Language domain	
Modalities (beliefs construct)	
Teachers should model how different formats (for example: oral, written, pictorial, physical, gestural) can be used to represent science ideas.	Teachers should show students how to represent their thinking about science ideas in different ways (for example: oral, written, pictorial, physical, gestural).
Registers (preparedness construct)	
Develop students' understanding of science ideas without introducing the precise scientific language	Develop students' understanding of science ideas before introducing science vocabulary
Interactions (instructional practices construct)	
Help students choose a way to represent their thinking about a science idea based upon their audience and purpose	Have students consider who they are communicating with (for example: peers, teachers, community members) when deciding how to represent their thinking about a science idea, such as orally, in writing, pictorially, physically, gesturally

how students might use CCCs from the NGSS (e.g., 'an approach such as looking for patterns, defining and analyzing systems, or examining cause and effect').

3.2.2.3 | *Coherence*

Items addressing coherence were generally understood by teachers when the items (a) related to DCIs, the dimension of science learning with which teachers were most familiar (see above), or (b) did not specify a dimension. For example, asking teachers how often they "have students make connections across core science ideas" (i.e., focus on DCIs) and "make connections between what students are currently learning and what they previously learned" (i.e., no dimension specified) did not cause confusion. However, items addressing coherence in relation to

SEPs and CCCs proved more challenging, particularly when the focus was coherence over longer timeframes (e.g., over a year). For example, we attempted 23 iterations of a beliefs item addressing whether teachers should help build students' ability to use CCCs more independently over a year as they make sense of phenomena within and across science disciplines. Ultimately, as none of the items were interpreted as intended, all were either dropped or repurposed to address a domain element related to CCCs in the three-dimensional learning area. Thus, challenges related to coherence compounded challenges described above related to teachers' lack of familiarity with certain dimensions of science learning (e.g., teachers were less familiar with CCCs and thus had difficulty interpreting coherent instruction with CCCs).

3.2.2.4 | *Modalities*

Teachers generally understood “different ways of representing ideas” for “modalities” and were able to provide classroom examples for the different modalities mentioned (i.e., oral, written, pictorial, physical, and gestural). However, in items that targeted how often teachers “model different ways of representing ideas,” some teachers were confused about who was doing the modeling (e.g., students vs. teachers) and what it might look like in the classroom. Thus, we replaced “model” with “show students,” which also avoided confusion with the SEP of developing and using models. One particularly challenging idea for teachers related to helping students consider the strengths and limitations of different representations. In general, teachers tended to see value in multiple ways of representing ideas because “all students learn differently” rather than because of the affordances of representations themselves (e.g., pictorial representations help convey spatial relations among components in a system). One teacher described, “Obviously with science, you have to hit each student based on how they learn best, whether that's auditory, whether it's kinesthetic.” Ultimately, all of the items addressing strengths and limitations were dropped. However, items addressing how often teachers “point out similarities and differences between different representations” were interpreted by teachers as intended and thus retained.

3.2.2.5 | *Registers*

We initially used the phrase “precise language” to indicate the specialized science register at the word level (e.g., science vocabulary), sentence level (e.g., syntax such as ‘…because…’ for communicating cause and effect), and discourse level (e.g., argument structure). However, teachers interpreted this phrase narrowly to indicate science vocabulary (i.e., word level). For example, when asked to define “precise language,” teachers described “words you'll only need to know if you're a scientist” and “the exact term, rather than ‘thingamajig.’” We addressed this issue in two ways. First, when the item was intended to target the word level specifically, we replaced “precise language” with “science vocabulary.” For example, an item was revised to ask teachers how often they “have students explore a science idea before introducing new science vocabulary.” Second, items that were intended to go beyond the word level were revised to emphasize the purpose of language use at other levels. For example, an item was revised to emphasize the purpose of forming an argument at the discourse level: “Provide students with scaffolds, such as graphic organizers, to help them form scientific arguments (claims supported by evidence and reasoning).”

3.2.2.6 | *Interactions*

We initially used “audience” to describe how students in the science classroom participate in interactions with different people (e.g., one-to-one, one-to-many). For example, an item asked teachers how often they “help students choose a way to represent their thinking about a science idea based on their audience and purpose.” However, teachers' interpretations of “audience” varied, with some teachers thinking that it referred only to people outside of the classroom. One teacher lamented, “We don't have a science fair.” Thus, teachers tended to rate their use of this instructional practice as minimal because their students did not often have opportunities to communicate about science with people beyond the classroom. Thus, we changed “audience” to “who [students] are communicating with” and provided accompanying examples. A final version of the item above asked teachers how often they “have students consider who they are communicating with (for example: peers, teachers, community members) when deciding how to represent their thinking about a science idea.”

3.3 | Phase 3: Piloting and final item selection

Once cognitive interviews indicated that teachers were consistently interpreting items as intended, we piloted 156 of the items across the six scales (i.e., beliefs, preparedness, and instructional practices in science and language domains). K-6 teachers of science were recruited through the mailing lists and social media accounts of Horizon Research, Inc. and the National Science Teaching Association. Whereas the cognitive interviews were conducted with teachers of science in Grades 4-6, for the pilot, we expanded to K-6 teachers of science to ensure the sample size was large enough to run the necessary analyses. Teachers were offered a $30 honorarium for completing the pilot questionnaire.

Interested teachers filled out an online registration survey answering questions about their teaching contexts. Over 1000 responses were received. The registration data were cleaned, and cases were removed due to respondents not currently being K-6 teachers of science, respondents not being located in the United States, or responses being identified as bot generated. (Responses identified as bot generated included nonsensical answers to open-ended questions, such as questions requesting the respondent's mailing address.) The remaining 529 teachers were emailed the online pilot questionnaire, which they were given 2 weeks to complete. The questionnaire contained 156 items and was expected to take around 30 min to complete. Because of the number of items and expected response time, the items were randomized within each construct (i.e., beliefs, preparedness, and instructional practices) to reduce bias due to survey fatigue. The teachers also each received a unique link to the questionnaire that allowed them to pick up where they left off if they were not able to complete the questionnaire in one sitting.

A total of 358 responses were received. The data were cleaned, and an additional 48 responses were removed for not completing the questionnaire or for completing it in an unrealistic amount of time (i.e., fewer than 8 min). The median completion time was 29 min, with two-thirds of the teachers completing the questionnaire in fewer than 32 min. Because respondents could complete the questionnaire over multiple sittings and the questionnaire software captured only start and end times, completion times should be interpreted with caution; however, the median completion time is consistent with response time estimates for this type of item (e.g., Couper and Kreuter, 2013). Negatively worded items (e.g., ‘ELs are less able to contribute ideas because of their emerging English proficiency’) were reverse coded, and the data were reviewed for the quality of the responses. Cases that showed excessive use of straight-lining (i.e., selecting the same response option for every item) or other patterned responses were flagged. Cases were also flagged that had contradictory answers to similar items (e.g., ‘ELs should be mainstreamed with non-ELs rather than pulled out for science instruction’ and ‘ELs should be pulled out for science instruction rather than mainstreamed with non-ELs’). A confirmatory factor analysis was conducted with and without flagged cases and resulted in the same conclusions. Therefore, the results reported below include the full sample.

The final sample consisted of 310 K-6 teachers of science. Table 4 displays the characteristics of the pilot questionnaire participants. Most of the teachers in our sample taught science in Grades 3–6, while a smaller percentage taught science in Grades K-2. Percentages for this characteristic exceed 100 because some participants reported teaching multiple grades. Most of the teachers (88%) taught science in face-to-face classes. Nearly three quarters of the teachers (71%) reported being at least fairly familiar with the NGSS. In terms of community and school type, our sample was generally consistent with the science teacher population nationwide (Banilower et al., 2018), with the majority of participants (87%) teaching in public schools (93% nationally) and about half (51%) in suburban communities (55% nationally). However, teachers in our sample had more K-12 teaching experience than teachers nationwide, with more than a third (37%) having taught for 21 years or more (20% nationally). Also,

TABLE 4 Characteristics of pilot questionnaire participants.

	Percentage (*N* = 310)
Grade(s) taught	
K	16%
1	15%
2	20%
3	27%
4	30%
5	42%
6	38%
Class format	
Face to face	88%
Hybrid	10%
Online	2%
Familiarity with the NGSS	
Not at all	7%
Somewhat	22%
Fairly	28%
Very	43%
Years of K-12 teaching experience	
0–5	6%
6–10	16%
11–20	41%
21 or more	37%
Frequency with which science was taught	
All or most days, every week	54%
Every week, but not every day	34%
Some weeks, but not every week	12%
Community type	
Rural	21%
Suburban	51%
Urban	28%
School type	
Public	87%
Private	13%

Note: Percentages for "Grade(s) taught" exceed 100 because some participants reported teaching multiple grades.
Abbreviation: NGSS, Next Generation Science Standards.

teachers in our sample taught science more frequently, with more than half (54%) teaching science all or most days every week during the school year (18% nationally).

3.3.1 | Confirmatory factor analysis

A confirmatory factor analysis was conducted with Mplus version 8.1 (Muthén & Muthén, 1998) using weighted least squares estimation, which was appropriate given the ordinal nature of the data (Li, 2016). Model-data fit, item factor loadings, and interfactor correlations were used to evaluate the models. Model-data fit indices consisted of the comparative fit index (CFI), Tucker-Lewis index (TLI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR; Brown, 2015; Hu & Bentler, 1999; Kline, 2023). Cronbach's alpha was also calculated for each scale.

The initial confirmatory factor analysis loaded all 156 items onto their respective factors: 43 items onto Sci Beliefs, 30 items onto Lang Beliefs, 22 items onto Sci Prep, 15 items onto Lang Prep, 28 items onto Sci Practices, and 18 items onto Lang Practices. Items were removed from the model based on consideration of factor loading, adequate coverage of the domain, and ultimate response burden. For example, items with factor loadings below 0.3 were removed unless an item was needed to ensure adequate coverage of the domain. Items were also removed if they pertained to domain elements that were overly represented in the set of items to keep response burden reasonable. A total of 80 items were removed from the model.

A final confirmatory factor analysis was conducted to assess the six-factor model after items were removed. The median factor loading for this model was 0.67, with loadings ranging from 0.26 to 0.85 (Only one loading was below 0.30.). The factor loadings for the final model can be found in Appendix A). Table 5 compares the model-data fit statistics between the initial model and final model in light of evaluation standards from the measurement literature (Brown, 2015; Hu & Bentler, 1999; Kline, 2023). Modifications to reduce the scales to a reasonable number of items resulted in modest improvements and did not negatively impact fit. CFI and TLI improved from about 0.90 to 0.95 (with 1.0 indicating perfect fit on these indices). RMSEA remained at 0.03, and SRMR improved from 0.08 to 0.07 (with 0.0 indicating perfect fit on these indices). Taken together, the fit statistics for the final model indicate an overall good fit.

The interfactor correlations in the final model are displayed in Table 6. Correlations among the six factors are moderate to strong. In particular, strong correlations *within* each domain (science and language) between the preparedness and instructional practices factors are not surprising given that teachers are unlikely to use instructional practices that they do not feel prepared to use. In addition, correlations *between* domains on the preparedness and instructional practices factors are very high, which is consistent with the intertwined and mutually supportive nature of teaching NGSS science and teaching MLs in science (e.g., National Academies of Sciences, Engineering, and Medicine, 2018). Although these factors could be combined into a single preparedness factor and a single instructional practices factor, keeping them separate can be beneficial. First, it allows other researchers to select sets of items that are best aligned with their research aims. Second, it highlights the importance for practitioners and policymakers of explicitly addressing preparedness and instructional practices for teaching MLs in science.

The final number of items and Cronbach's alpha for each scale are displayed in Table 7. In total, the questionnaire consists of 76 items across the six scales. An alpha of 0.60–0.80 is generally considered evidence of moderate reliability, and an alpha over 0.80 is considered evidence of strong reliability (Tavakol & Dennick, 2011). Five of the six scales have an alpha that exceeds 0.80, while the remaining scale (Lang Beliefs) has an alpha of 0.75. The complete questionnaire can be found in Appendix B.

TABLE 5 Comparison of model-data fit statistics between the initial and final model.

Fit statistic	Initial model	Final model	Goodness-of-fit recommendations[a]
CFI	0.90	0.95	>0.95
TLI	0.90	0.95	>0.95
RMSEA	0.03	0.03	<0.05
SRMR	0.08	0.07	<0.08

Abbreviations: CFI, comparative fit index; RMSEA, root mean square error of approximation; SRMR, standardized root mean square residual; TLI, Tucker-Lewis index.

[a]Brown (2015); Hu and Bentler (1999); Kline (2023).

TABLE 6 Interfactor correlations in the final model.

	Sci beliefs	Lang beliefs	Sci prep	Lang prep	Sci practices	Lang practices
Sci Beliefs						
Lang Beliefs	0.59					
Sci Prep	038	0.34				
Lang Prep	0.40	0.34	0.98			
Sci Practices	0.59	0.33	0.75	0.74		
Lang Practices	0.46	0.35	0.74	0.77	0.90	

Note: All correlations are statistically significant ($p < 0.001$).

TABLE 7 Final number of items on each scale and reliability estimates.

	Number of items	Cronbach's alpha
Sci Beliefs	13	0.85
Lang Beliefs	16	0.75
Sci Prep	10	0.90
Lang Prep	12	0.90
Sci Practices	11	0.87
Lang Practices	14	0.86

3.3.2 | Cluster analysis

An exploratory k-means cluster analysis was conducted to examine whether the final scales could differentiate among different groups of teachers and how these groupings relate to teacher characteristics (see Table 4), thus providing validity evidence based on relations of teachers' scores to other variables. Composite scores were calculated for each scale by summing the responses to the associated items (after reverse coding negatively worded items) and dividing by the total points possible. For the composites to be on a 100-point scale, the lowest response option on each scale was set to 0, and the other responses were adjusted accordingly. Thus, a teacher who marked the lowest response option on every item for a scale would receive a composite score of 0. This ensures that a score

TABLE 8 Mean composite scores for each cluster.

	Cluster 1 (*n* = 76)	Cluster 2 (*n* = 116)	Cluster 3 (*n* = 118)
Sci Beliefs	73.83	76.48	78.30
Lang Beliefs	74.15	77.47	82.69
Sci Prep	54.79	75.29	91.50
Lang Prep	57.95	78.37	92.99
Sci Practices	60.70	68.58	81.50
Lang Practices	61.31	70.21	84.92

Note: There is a significant difference between each pair of clusters on each composite ($p < 0.05$), except for Clusters 2 and 3 on the Sci Beliefs composite ($p > 0.05$).

TABLE 9 Percentage of teachers in each cluster by teacher characteristics.

	Cluster 1 (*n* = 76)	Cluster 2 (*n* = 116)	Cluster 3 (*n* = 118)
Years of K-12 teaching experience			
0–5	12%	5%	2%
6–10	16%	16%	16%
11–20	45%	37%	43%
21 or more	27%	42%	39%
Frequency with which science was taught			
All or most days, every week	39%	53%	64%
Every week, but not every day	47%	32%	28%
Some weeks, but not every week	14%	15%	8%
Familiarity with the NGSS			
Not at all	10%	5%	6%
Somewhat	30%	26%	13%
Fairly	35%	27%	25%
Very	25%	42%	56%

Abbreviation: NGSS, Next Generation Science Standards.

of 50 is the true midpoint. The denominator for each composite was determined by computing the maximum possible sum of responses for items on each scale and dividing by 100.

We explored multiple cluster solutions. Results indicated a three-group solution best fit the data, considering cluster size and apparent differences in composite scores. Mean composite scores for each cluster are shown in Table 8. A one-way ANOVA confirmed that there are statistically significant differences in scores among the three clusters for each composite. Tukey's HSD test for multiple comparisons found statistically significant differences between each pair of clusters on all six composites, except for Clusters 2 and 3 on the Sci Beliefs composite. As shown in Table 8, composite scores increase from Cluster 1 to Cluster 2 to Cluster 3, with the largest differences appearing to be on the preparedness and practices composites.

Cluster 1 represents teachers with relatively low reported preparedness and use of instructional practices in science and language domains, while Cluster 3 represents teachers with high reported preparedness in both domains and fairly high use of instructional practices. Cluster 2 falls in the middle, with moderate reported preparedness and use of instructional practices. All three clusters have similar beliefs, though there appear to be increases across the clusters within each domain.

We examined differences in teacher characteristics by cluster. Table 9 shows the percentage of teachers in each cluster by three characteristics: (a) years of K-12 teaching experience, (b) frequency with which science was taught, and (c) familiarity with the NGSS. For all three characteristics, as the variable increased (i.e., more years of teaching experience, more frequent science teaching, and more familiarity with the NGSS), teachers' scores on the scales tended to increase. First, teachers in Cluster 1 tended to have less experience than teachers in Clusters 2 and 3. For example, 27% of teachers in Cluster 1 had 21 or more years of teaching experience compared to 42% of teachers in Cluster 2 and 39% of teachers in Cluster 3. Second, the frequency with which science was taught increases across the clusters. For example, 39% of teachers in Cluster 1 taught science all or most days every week, compared to 53% of teachers in Cluster 2 and 64% of teachers in Cluster 3. Finally, familiarity with the NGSS increases across the clusters. For example, 25% of teachers in Cluster 1 reported being very familiar with the NGSS, compared to 42% of teachers in Cluster 2 and 56% of teachers in Cluster 3.

4 | DISCUSSION

As instruments for measuring the impacts of the NGSS are sorely needed in the science education community, this study reported on the initial development and validation of a questionnaire for measuring teachers' beliefs, preparedness, and instructional practices for teaching NGSS science with MLs. Results from the first phase, domain specification and expert review, indicated three areas in the science domain (phenomena, three-dimensional learning, and coherence) and the language domain (registers, modalities, and interactions), respectively. Results from the second phase, item writing and cognitive interviews, indicated that iterative testing and revision of items facilitated teachers' meaningful interpretation. Results from the third phase, piloting and final item selection, indicated that the questionnaire provides reliable measures that differentiate among teachers with different levels of the underlying constructs and also that teachers' scores on the questionnaire relate to their characteristics. In this section, we describe implications of our development and validation process and future research directions that build on the present study.

4.1 | Implications of development and validation process

One implication of our development and validation process is the difficulty of communicating with teachers about contemporary approaches to instruction while teachers are still developing their understanding of such approaches and their differences from more traditional approaches. This difficulty was salient during cognitive interviews in the item writing phase. In some instances, teachers showed varying interpretations of terminology associated with contemporary science instruction based on the NGSS (e.g., 'phenomena,' 'crosscutting concept') and contemporary science instruction with MLs (e.g., 'modalities,' 'registers'). As a result, items in our questionnaire do not always reflect the terminology used in policy and research to characterize instructional shifts (e.g., we use 'science-related events or processes,' accompanied by multiple carefully selected examples, to communicate about 'phenomena'). In other instances, teachers still did not interpret items as intended even after multiple rounds of testing and revision. As a result, certain domain sub-elements lack coverage (e.g., sub-element about the strengths and limitations of modalities in the language domain).

This difficulty of communicating with teachers about contemporary approaches to instruction was somewhat expected given that (a) the NGSS call for significant instructional shifts (Next Generation Science Standards Lead States, 2013); (b) the status of NGSS implementation varies widely across states, districts, schools, and teachers (e.g., Neill & Paulson, 2023); and (c) contemporary approaches to science instruction with MLs are relatively new (e.g., National Academies of Sciences, Engineering, and Medicine, 2018). Thus, teachers may have been interpreting the items through the frame of reference (e.g., Spillane et al., 2002) of more traditional approaches to science instruction (e.g., interpreting phenomena as 'hands-on learning') and teaching science with MLs (e.g., interpreting registers narrowly in terms of vocabulary). Another contributing factor could be the cognitive demands on respondents when items become long with clarifying examples for multiple words or phrases.

Given this difficulty, revisions based on cognitive interviews were crucial to ensuring items were interpreted by teachers as intended (Desimone & Le Floch, 2004), particularly in our project that involves teachers participating in our intervention as well as comparison teachers. However, these revisions raise potential issues. One issue is that the questionnaire may have less face validity in the eyes of interest holders (e.g., researchers, teachers) who are already well versed in science and language instructional shifts and associated terminology. Another issue is that, as contemporary approaches to science instruction with MLs become more widely implemented in the education system (Lee & Grapin, 2024), new challenges in interpreting the items could arise (e.g., teachers may wonder whether 'science-related events or processes' refer to what has become the familiar notion of 'phenomena'). Still another issue is that the questionnaire cannot be used to collect information on domain sub-elements that currently lack coverage, even as items based on these sub-elements could become interpretable by teachers as they participate in NGSS-based PD and implement NGSS-designed curricula in their classrooms. While these issues may be particularly pronounced in the context of the NGSS, which call for significant instructional shifts that are communicated through specialized terminology (e.g., Lee et al., 2023), similar issues are likely to arise in other reform-oriented instrument development efforts, given that language is often "the chief medium that policymakers have for representing their ideas about revising local practice" (Spillane et al., 2002, p. 407). Together, these issues underscore the need to evaluate the evidence supporting the questionnaire's interpretation and use on an ongoing basis (Kane, 2013).

Another implication is the potential of emerging NGSS-based instruments to inform the design of PD experiences for teachers. This implication is timely as NGSS-based PD programs and instruments are currently major areas of focus in the field and are being developed in tandem (e.g., Lowell & McNeill, 2023). In our study, the primary purpose of the cluster analysis was to provide validity evidence based on the relation between teachers' scores on the scales and teacher characteristics. However, this analysis could also offer substantive implications for informing PD by making visible different "profiles" of teachers in relation to the constructs measured. This would begin to answer Lowell and McNeill's (2023) call for PD that embeds "more differentiated or targeted support" (p. 1478) for teachers in response to their experiences with NGSS instruction.

Specifically, the clusters identified through our analysis could help differentiate PD experiences in ways that reflect the specific goals of PD and respond to teachers' needs. For example, Cluster 1 represents teachers with relatively low reported preparedness and use of instructional practices in science and language domains. Even as teachers' familiarity with the NGSS was lowest in this cluster, one in four teachers reported being very familiar with the NGSS. Thus, differentiating PD experiences based solely on general familiarity with the NGSS might have fallen short of addressing the needs of these teachers, who likely require more targeted support related to constructs measured by the questionnaire. In addition, the largest differences across clusters appeared to be on the preparedness and instructional practices composites, suggesting that targeted supports might be particularly high leverage in these areas. As different uses of an instrument require different validity evidence (Kane, 2013), whether and how these types of analyses can meaningfully inform teachers' PD experiences requires further empirical inquiry (see 'Future Research Directions'). As described next, the relation between NGSS-based PD programs and NGSS-based instrumentation is synergistic, with instruments offering the potential to inform PD and PD facilitating the collection of further validity evidence for instruments.

4.2 | Future research directions

This study suggests future research directions that build on our initial development and validation work. One future direction involves administering the questionnaire to the same group of teachers at multiple time points. In the context of our research project that focuses on the yearlong SAIL curriculum, we administer the questionnaire at the beginning and end of each school year, which enables us to collect validity evidence that the instrument can detect changes in teachers' beliefs, preparedness, and instructional practices as a result of their participation in curriculum-based PD and implementation of the curriculum in their classrooms. Moreover, as the intervention will occur over a 2-year period, with the second year involving both teachers new to the intervention and those who are continuing participation into the second year, we will examine the instrument's sensitivity to teachers' years of participation.

A second future direction involves collecting further validity evidence based on relations to other variables. For example, studies could investigate whether embedding targeted supports for teachers in PD impacts their beliefs, preparedness, and/or instructional practices in the targeted areas. Studies could also examine how scores on the questionnaire relate to scores on instruments that target similar constructs using other data collection methods. For example, studies could examine the relation between teachers' *reported* instructional practices, as measured by our questionnaire, and teachers' *observed* instructional practices, as measured by classroom observation protocols. While NGSS-based observation protocols are beginning to emerge (e.g., Chen & Terada, 2021), they tend to focus on specific aspects of NGSS instruction (e.g., SEPs) and do not address science instructional shifts comprehensively or teaching MLs in science. In the context of our research project, we are developing and validating a classroom observation protocol to measure the quality of instruction in science and language domains (Lee et al., 2024). The protocol produces numerical ratings that will enable us to examine the relation between teachers' reported and observed instructional practices quantitatively.

A third future direction involves collecting and analyzing qualitative data related to constructs measured by the questionnaire. For example, in-depth interviews with teachers about their beliefs could serve to support or refute interpretations based on the questionnaire. The collection and analysis of qualitative data are particularly important given that extensive research in language education indicates the fluid and context-dependent nature of teachers' beliefs (e.g., Bacon, 2020) as well as the affordances of qualitative analyses for complicating interpretations of such beliefs based on survey instruments alone (e.g., Anderson et al., 2022).

A fourth future direction, particularly for researchers who may use or adapt our questionnaire for their own projects, is to investigate its use in grade levels beyond K-6 (i.e., the grade levels taught by teachers in our pilot). Currently, NGSS-based interventions are being carried out in middle school (e.g., Harris et al., 2022; Lowell & McNeill, 2023) and high school (e.g., Schneider et al., 2022). Given that our domain definition addressed science and language instructional shifts broadly (e.g., National Academies of Sciences, Engineering, and Medicine, 2018; National Research Council, 2012) and was not limited to science teaching at the elementary level, it is likely that many items are relevant across K-12 grades. Still, extending beyond K-6 grades will require collecting multiple sources of validity evidence (e.g., evidence that middle and high school teachers interpret the items as intended). This research could facilitate the collection of data in a range of contexts with larger and more diverse samples of teachers that represent the science teacher population nationwide.

A fifth and final research direction involves the use of the questionnaire for pedagogical purposes. For example, science methods instructors could use our questionnaire to engage preservice teachers in reflecting on their beliefs at the beginning of a course, thus providing useful diagnostic information, or at both the beginning and end, thus promoting teacher reflection on how their beliefs changed in response to course content. The questionnaire could also be used to engage preservice teachers in reflecting on their preparedness for teaching MLs in science and how often they use instructional practices (e.g., in student teaching). This research would capitalize on an affordance of instrumentation work that it requires operationalizing high-level ideas about instruction, which teachers may perceive as too theoretical to implement, into more concrete beliefs and practices. As with the other research directions articulated above, extending to new teacher populations (e.g., preservice teachers) and purposes (e.g., pedagogical) will require the ongoing collection of validity evidence and continued scrutiny of the questionnaire's interpretation and use.

5 | CONCLUSION

Addressing persistent inequities facing historically marginalized students in science education requires implementation efforts at a large scale. The *Framework* and the NGSS set the stage for large-scale implementation with the adoption or adaptation of the standards across 48 U.S. states (https://www.nsta.org/science-standards). This widespread adoption or adaptation has been accompanied by greater collaboration and mutual understanding across research communities, especially science education and language education (Lee & Grapin, 2024; National Academies of Sciences, Engineering, and Medicine, 2018). A crucial next step is to develop instruments that can measure the impacts of large-scale implementation on teachers and students. This study extends the emerging literature on NGSS-based instrumentation by reporting on the development of a teacher questionnaire that addresses instructional shifts in a comprehensive manner (i.e., phenomena, three-dimensional learning, coherence, registers, modalities, and interactions) in two domains (i.e., science and language) across three constructs (i.e., beliefs, preparedness, and instructional practices). The development of such instruments can contribute to painting a more complete portrait of teachers' instructional shifts, toward the ultimate goal of disrupting persistent inequities facing MLs in science education.

ACKNOWLEDGMENTS

This research was supported by a National Science Foundation grant (DRL-2035103).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data are not publicly available due to privacy or ethical restrictions. The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Scott E. Grapin http://orcid.org/0000-0002-8982-0771
Alycia J. Sterenberg Mahon http://orcid.org/0000-0003-0744-0917
Alison Haas http://orcid.org/0000-0002-8545-7346
Okhee Lee http://orcid.org/0000-0003-3551-1583

REFERENCES

Accurso, K. (2020). Bringing a social semiotic perspective to secondary teacher education in the United States. *Journal of English for Academic Purposes*, *44*, 100801. https://doi.org/10.1016/j.jeap.2019.100801

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing. American Educational Research Association*. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. https://www.aera.net/Publications/Books/Standards-for-Educational-Psychological-Testing-2014-Edition

Anderson, K. T., Ambroso, E., Cruz, J., Zuiker, S. J., & Rodríguez-Martínez, S. (2022). Complicating methods for understanding educators' language ideologies: Transformative approaches for mixing methods. *Language and Education*, *36*(1), 1–19. https://doi.org/10.1080/09500782.2021.1931296

Bacon, C. K. (2020). “It's not really my job”: A mixed methods framework for language ideologies, monolingualism, and teaching emergent bilingual learners. *Journal of Teacher Education*, *71*(2), 172–187. https://doi.org/10.1177/0022487118783188

Banilower, E., Smith, P. S., Malzahn, K., Plumley, C., Gordon, E., & Hayes, M. (2018). *Report of the 2018 National Survey of Science and Mathematics Educators plus*. Horizon Research, Inc. https://files.eric.ed.gov/fulltext/ED598121.pdf

Banilower, E., Smith, P. S., Weiss, I., Malzahn, K., Campbell, K., & Weis, A. (2013). *Report of the 2012 National Survey of Science and Mathematics Education*. Horizon Research, Inc. https://files.eric.ed.gov/fulltext/ED541798.pdf

Borg, S. (2018). Teachers' beliefs and classroom practices. In P. Garrett, & J. M. Cots (Eds.), *The Routledge handbook of language awareness* (pp. 75–91). Routledge. https://doi.org/10.4324/9781315676494-5

Brown, B. A., & Ryoo, K. (2008). Teaching science as a language: A “content-first” approach to science teaching. *Journal of Research in Science Teaching*, *45*(5), 529–553. https://doi.org/10.1002/tea.20255

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford.

Bunch, G. C. (2014). The language of ideas and the language of display: Reconceptualizing “academic language” in linguistically diverse classrooms.”. *International Multilingual Research Journal*, *8*(1), 70–86. https://doi.org/10.1080/19313152.2014.852431

Buxton, C. A., & Lee, O. (2023). Multilingual learners in science education. In N. G. Lederman, D. Zeidler, & J. Lederman (Eds.), *Handbook of research in science education* (3rd ed., pp. 290–323). Routledge.

Byrnes, D. A., & Kiger, G. (1994). Language attitudes of teachers scale. *Educational and Psychological Measurement*, *54*(1), 227–231. https://doi.org/10.1177/0013164494054001029

Campbell, T., Lee, H., Longhurst, M., McKenna, T. J., Coster, D., & Lundgren, L. (2021). Next generation science classrooms: The development of a questionnaire for examining student experiences in science classrooms. *School Science and Mathematics*, *121*(2), 96–109. https://doi.org/10.1111/ssm.12449

Campbell, T., & Lee, O. (2021). Instructional materials designed for a framework for K-12 science education and the next generation science standards: An introduction to the special issue. *Journal of Science Teacher Education*, *32*(7), 727–734.

Chen, Y. C., & Terada, T. (2021). Development and validation of an observation-based protocol to measure the eight scientific practices of the next generation science standards in K-12 science classrooms. *Journal of Research in Science Teaching*, *58*(10), 1489–1526. https://doi.org/10.1002/tea.21716

Couper, M. P., & Kreuter, F. (2013). Using paradata to explore item level response times in surveys. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *176*(1), 271–286. https://doi.org/10.1111/j.1467-985X.2012.01041.x

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, *38*(3), 181–199. https://doi.org/10.3102/0013189X08331140

Desimone, L. M., & Le Floch, K. C. (2004). Are we asking the right questions? Using cognitive interviews to improve surveys in education research. *Educational Evaluation and Policy Analysis*, *26*(1), 1–22. https://doi.org/10.3102/01623737026001001

Fulmer, G. W., Hwang, J., Ding, C., Hand, B., Suh, J. K., & Hansen, W. (2021). Development of a questionnaire on teachers' knowledge of language as an epistemic tool. *Journal of Research in Science Teaching*, *58*(4), 459–490. https://doi.org/10.1002/tea.21666

Gibbons, P. (2015). *Scaffolding language, scaffolding learning: Teaching English language learners in the mainstream classroom*. Heinemann.

González-Howard, M., & McNeill, K. L. (2016). Learning in a community of practice: Factors impacting English-learning students' engagement in scientific argumentation. *Journal of Research in Science Teaching*, *53*(4), 527–553. https://doi.org/10.1002/tea.21310

González-Howard, M., & Suárez, E. (2021). Retiring the term English language learners: Moving toward linguistic justice through asset-oriented framing. *Journal of Research in Science Teaching*, *58*(5), 749–752. https://doi.org/10.1002/tea.21684

Grapin, S. (2019). Multimodality in the new content standards era: Implications for English learners. *TESOL Quarterly*, *53*(1), 30–55. https://doi.org/10.1002/tesq.443

Grapin, S. E. (2021). Toward asset-oriented and definitionally clear terminology: A comment on González-Howard and Suárez (2021). *Journal of Research in Science Teaching*, *58*(5), 753–755. https://doi.org/10.1002/tea.21686

Grapin, S. E., Pierson, A., González-Howard, M., Ryu, M., Fine, C., & Vogel, S. (2023). Science education with multilingual learners: Equity as access and equity as transformation. *Science Education*, *107*(4), 999–1032. https://doi.org/10.1002/sce.21791

Haas, A., Januszyk, R., Grapin, S. E., Goggins, M., Llosa, L., & Lee, O. (2021). Developing instructional materials aligned to the next generation science standards for all students, including English learners. *Journal of Science Teacher Education*, *32*(7), 735–756. https://doi.org/10.1080/1046560X.2020.1827190

Hakuta, K., Santos, M., & Fang, Z. (2013). Challenges and opportunities for language learning in the context of the CCSS and the NGSS. *Journal of Adolescent & Adult Literacy*, *56*(6), 451–454. https://doi.org/10.1002/jaal.164

Harris, C., Feng, M., Murphy, R., & Rutstein, D. (2022). Curriculum materials designed for the next generation science standards show promise: Initial results from a randomized controlled trial in middle schools. *WestEd*. https://www.wested.org/resources/curriculum-materials-for-ngss/

Harris, C. J., Krajcik, J. S., Pellegrino, J. W., & DeBarger, A. H. (2019). Designing knowledge-in-use assessments to promote deeper learning. *Educational Measurement: Issues and Practice*, *38*(2), 53–67. https://doi.org/10.1111/emip.12253

Hayes, K. N., Lee, C. S., DiStefano, R., O'Connor, D., & Seitz, J. C. (2016). Measuring science instructional practice: A survey tool for the age of NGSS. *Journal of Science Teacher Education*, *27*, 137–164. https://doi.org/10.1007/s10972-016-9448-5

Horwitz, E. K. (1985). Using student beliefs about language learning and teaching in the foreign language methods course. *Foreign Language Annals*, *18*(4), 333–340. https://doi.org/10.1111/j.1944-9720.1985.tb01811.x

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.

Huerta, M., Garza, T., Jackson, J. K., & Murukutla, M. (2019). Science teacher attitudes towards English learners. *Teaching and Teacher Education*, *77*, 1–9. https://doi.org/10.1016/j.tate.2018.09.007

Jones, M. G., & Park, S. (2023). Science teacher attitudes and beliefs: Reforming practice. In N. G. Lederman, D. L. Zeidler, & J. S. Lederman (Eds.), *Handbook of Research on Science Education*. Routledge. https://doi.org/10.4324/9780367855758-40

Kane, M. (2013). The argument-based approach to validation. *School Psychology Review*, *42*(4), 448–457. https://doi.org/10.1080/02796015.2013.12087465

Kline, R. B. (2023). *Principles and practice of structural equation modeling*. Guilford.

Krajcik, J., & Czerniak, C. (2018). *Teaching science in elementary and middle school: A project-based learning approach* (5th ed.). Routledge. https://doi.org/10.4324/9781315205014

Krajcik, J., Schneider, B., Miller, E. A., Chen, I. C., Bradford, L., Baker, Q., Bartz, K., Miller, C., Li, T., Codere, S., & Peek-Brown, D. (2023). Assessing the effect of project-based learning on science learning in elementary schools. *American Educational Research Journal*, *60*(1), 70–102. https://doi.org/10.3102/00028312221129247

Kress, G., Jewitt, C., Ogborn, J., & Tsatsarelis, C. (2014). *Multimodal teaching and learning: The rhetorics of the science classroom* (2nd ed.). Bloomsbury Academic.

Lee, O. (2004). Teacher change in beliefs and practices in science and literacy instruction with english language learners. *Journal of Research in Science Teaching*, *41*(1), 65–93. https://doi.org/10.1002/tea.10125

Lee, O. (2021). Asset-orientedframing of science and language learning with multilingual learners. *Journal of Research in Science Teaching*, *58*(7), 1073–1079. https://doi.org/10.1002/tea.21694

Lee, O., & Grapin, S. (2024). English language proficiency standards aligned with content standards: How the next generation science standards and WIDA 2020 reflect each other. *Science Education*, *108*(2), 637–658. https://doi.org/10.1002/sce.21843

Lee, O., Grapin, S., & Haas, A. (2023). Teacher professional development programs integrating science and language with multilingual learners: A conceptual framework. *Science Education*, *107*(5), 1302–1323. https://doi.org/10.1002/sce.21807

Lee, O., Llosa, L., Grapin, S., Haas, A., & Goggins, M. (2019). Science and language integration with english learners: A conceptual framework guiding instructional materials development. *Science Education*, *103*(2), 317–337. https://doi.org/10.1002/sce.21498

Lee, O., Pasley, J., Banilower, E., Grapin, S. E., Plumley, C., Harper, L., Haas, A., Schwenger, A., & Sterenberg Mahon, A. (2024). Classroom observation protocol for NGSS-based science instruction with multilingual learners. Manuscript submitted for publication.

Lee, O., Quinn, H., & Valdés, G. (2013). Science and language for English language learners in relation to next generation science standards and with implications for common core state standards for English language arts and mathematics. *Educational Researcher*, *42*(4), 223–233. https://doi.org/10.3102/0013189X13480524

Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936–949. https://doi.org/10.3758/s13428-015-0619-7

Loewenberg Ball, D., & Forzani, F. M. (2009). The work of teaching and the challenge for teacher education. *Journal of Teacher Education*, *60*(5), 497–511. https://doi.org/10.1177/0022487109348479

Lowell, B. R., & McNeill, K. L. (2023). Changes in teachers' beliefs: A longitudinal study of science teachers engaging in storyline curriculum-based professional development. *Journal of Research in Science Teaching*, *60*(7), 1457–1487. https://doi.org/10.1002/tea.21839

Martínez, J. F., Kloser, M., Srinivasan, J., Stecher, B., & Edelman, A. (2022). Developing situated measures of science instruction through an innovative electronic portfolio app for mobile devices: Reliability, validity, and feasibility. *Educational and Psychological Measurement*, *82*(6), 1180–1202. https://doi.org/10.1177/0013164421106492

Messick, S. (1989a). Meanings and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11.

Messick, S. (1989b). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education.

Muthén, L. K., & Muthén, B. O. (1998). 2017 Mplus user's guide (8th ed.).

National Academies of Sciences, Engineering, and Medicine. (2018). *English learners in STEM subjects. Transforming classrooms, schools, and lives*. The National Academies Press. https://www.nap.edu/catalog/25182/english-learners-in-stem-subjects-transforming-classrooms-schools-and-lives

National Academies of Sciences, Engineering, and Medicine. (2020). *Changing expectations for the K-12 teacher workforce: Policies, preservice education, professional development, and the workplace*. National Academies Press. https://nap.nationalacademies.org/catalog/25603/changing-expectations-for-the-k-12-teacher-workforce-policies-preservice

National Academies of Sciences, Engineering, and Medicine. (2021). *Call to action for science education: Building opportunity for the future*. National Academies Press. https://www.nap.edu/catalog/26152/call-to-action-for-science-education-building-opportunity-for-the

National Academies of Sciences, Engineering, and Medicine. (2022). *Science and engineering in preschool through elementary grades: The brilliance of children and the strengths of educators*. National Academies Press. https://nap.nationalacademies.org/catalog/26215/science-and-engineering-in-preschool-through-elementary-grades-the-brilliance

National Center for Education Statistics. (2024). *The condition of education 2024 (NCES 2024–144)*. U.S. Department of Education. https://nces.ed.gov/pubs2024/2024144.pdf

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. The National Academies Press. https://nap.nationalacademies.org/catalog/13165/a-framework-for-k-12-science-education-practices-crosscutting-concepts

Neill, T., & Paulson, D. (2023). *State science standards: An overview and analysis of review, revision, and adoption processes in states in the U.S. [June 2023 draft]*. National Academies of Sciences, Engineering, and Medicine. https://www.nationalacademies.org/documents/embed/link/LF2255DA3DD1C41C0A42D3BEF0989ACAECE3053A6A9B/file/DB51C5F17C13BED8AB938B6BC188103DA22C20E8558F?noSaveAs=1

Next Generation Science Standards Lead States. (2013). *Next generation science standards: For states, by states*. The National Academies Press. https://www.nextgenscience.org

Nollmeyer, G., & Bangert, A. (2017). Measuring elementary teachers' understanding of the NGSS framework: An instrument for planning and assessing professional development. *Electronic Journal for Research in Science and Mathematics Education*, *21*(8), 20–45. https://ejrsme.icrsme.com/article/view/17887

Pajares, M. F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research*, *62*(3), 307–332. https://doi.org/10.2307/1170741

Peercy, M. M., Tigert, J., & Fredricks, D. (2023). *Core practices for teaching multilingual students: Humanizing pedagogies for equity*. Teachers College Press.

Penuel, W. R., Krumm, A. E., Pazera, C., Singleton, C., Allen, A. R., & Deverel-Rico, C. (2023). Belonging in science classrooms: Investigating its relation to students' contributions and influence in knowledge building. *Journal of Research in Science Teaching*, *61*(1), 228–252. https://doi.org/10.1002/tea.21884

Pierson, A. E., Clark, D. B., & Brady, C. E. (2021). Scientific modeling and translanguaging: A multilingual and multimodal approach to support science learning and engagement. *Science Education*, *105*(4), 776–813. https://doi.org/10.1002/sce.21622

Reiser, B. J., Michaels, S., Moon, J., Bell, T., Dyer, E., Edwards, K. D., McGill, T. A. W., Novak, M., & Park, A. (2017). Scaling up three-dimensional science learning through teacher-led study groups across a state. *Journal of Teacher Education*, *68*(3), 280–298. https://doi.org/10.1177/0022487117699598

Reiser, B. J., Novak, M., McGill, T. A. W., & Penuel, W. R. (2021). Storyline units: An instructional model to support coherence from the students' perspective. *Journal of Science Teacher Education*, *32*(7), 805–829. https://doi.org/10.1080/1046560X.2021.1884784

Schneider, B., Krajcik, J., Lavonen, J., Salmela-Aro, K., Klager, C., Bradford, L., Chen, I. C., Baker, Q., Touitou, I., Peek-Brown, D., Dezendorf, R. M., Maestrales, S., & Bartz, K. (2022). Improving science achievement—Is it possible? Evaluating the efficacy of a high school chemistry and physics project-based learning intervention. *Educational Researcher*, *51*(2), 109–121. https://doi.org/10.3102/0013189X211067742

Short, J., & Hirsh, S. (2020). *The elements: Transforming teaching through curriculum-based professional learning*. Carnegie Corporation of New York.

Smith, P. S., Smith, A., & Banilower, E. (2014). Situating beliefs in the theory of planned behavior: The development of the Teacher Beliefs about Effective Science Teaching questionnaire, *The role of science teachers' beliefs in international classrooms* (pp. 81–102). Brill.

Spillane, J. P., Reiser, B. J., & Reimer, T. (2002). Policy implementation and cognition: Reframing and refocusing implementation research. *Review of Educational Research*, *72*(3), 387–431. https://www.jstor.org/stable/3515992

Suárez, E. (2020). "Estoy explorando science": Emergent bilingual students problematizing electrical phenomena through translanguaging. *Science Education*, *104*(5), 791–826. https://doi.org/10.1002/sce.21588

Suárez, E., & Otero, V. (2023). Ting, tang, tong: Emergent bilingual students investigating and constructing evidence-based explanations about sound production. *Journal of Research in Science Teaching*, *61*(1), 137–169. https://doi.org/10.1002/tea.21868

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

U.S. Department of Education. (2015). *Every Student Succeeds Act*. U.S. Department of Education. https://www.gpo.gov/fdsys/pkg/BILLS-114s1177enr/pdf/BILLS-114s1177enr.pdf

Valdés, G., Kibler, A., & Walqui, A. (2014). Changes in the expertise of ESL professionals: Knowledge and action in an era of new standards. TESOL International Association. https://www.tesol.org/media/vh1pnlsi/professional-paper-26-march-2014.pdf

Viesca, K. M., Strom, K., Hammer, S., Masterson, J., Linzell, C. H., Mitchell-McCollough, J., & Flynn, N. (2019). Developing a complex portrait of content teaching for multilingual learners via nonlinear theoretical understandings. *Review of Research in Education*, *43*(1), 304–335. https://doi.org/10.3102/0091732X18820910

Windschitl, M., Thompson, J., Braaten, M., & Stroupe, D. (2012). Proposing a core set of instructional practices and tools for teachers of science. *Science Education*, *96*(5), 878–903. https://doi.org/10.1002/sce.21027

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Grapin, S. E., Plumley, C., Banilower, E., Sterenberg Mahon, A. J., Craven, L., Malzahn, K., Pasley, J., Schwenger, A., Haas, A., & Lee, O. (2025). Development of a questionnaire on teachers' beliefs, preparedness, and instructional practices for teaching NGSS science with multilingual learners. *Science Education*, *109*, 128–156. https://doi.org/10.1002/sce.21905