TESOL Journal

WILEY

EMPIRICAL FEATURE ARTICLE

Scaffolding Multilingual Teaching Practices: Development and Validation of the SCOPE-ML Observation Instrument

Daniel Showalter[1] | Jihye Shin[2] | Jennifer R. Ottley[2] | Hye Pae[3] | Craig B. Howley[4] | Alisha Portolese[4] | Jennifer Kotler[4]

[1]Eastern Mennonite University, Harrisonburg, Virginia, USA | [2]Systems Development & Improvement Center, University of Cincinnati, Cincinnati, Ohio, USA | [3]School of Education, University of Cincinnati, Cincinnati, Ohio, USA | [4]WordFarmers Associates, Albany, Ohio, USA

Correspondence: Craig B. Howley (howleycb@gmail.com)

Received: 5 June 2024 | **Revised:** 28 September 2025 | **Accepted:** 6 November 2025

Keywords: instrument development | multilingual learner | observation | scaffolding

ABSTRACT

This article describes the development and validation of a new observation instrument, the Scaffolding Observation Protocol for Multilingual Learners (SCOPE-ML), designed to assess instructional changes resulting from two NPD projects at the University of Cincinnati. These projects focused on TESOL endorsement coursework and professional development for teachers of multilingual learners. The instrument was developed to measure differences in teaching practices between treatment and control groups in randomized controlled trials (RCTs), offering potential evidence of the effectiveness of the NPD interventions. The first project's treatment involved completion of the university's TESOL endorsement sequence, with participants recruited from upper elementary and secondary schools. The second project included two treatment conditions: (1) TESOL coursework and (2) completion of online modules with coaching and community of practice—with participants who taught in early elementary contexts. Development of the SCOPE-ML followed a rigorous, rapid process—literature review, conceptualization, item drafting, revision, pilot testing, and initial research use—guided by a pedagogical stance grounded in scaffolding, understood as the systematic support of student learning through responsive, asset-oriented instructional practice. This process produced evidence of high inter-rater reliability (percentage agreement and Cohen's kappa) and face validity (via peer debriefing and feedback). The study highlights the crucial role of teacher–student and peer interactions in fostering multilingual learners' literacy development and underscores the need for ongoing refinement of teaching practices for this population.

1 | Introduction

Instrument development work with two NPD projects at the University of Cincinnati (UC), housed at the Systems Development and Improvement Center (SDI Center) is the focus of this article. The instrument in question—the Scaffolding Observation Protocol for Multilingual Students (SCOPE-ML)—was developed to meet a specific evaluation need for these particular projects.[1] The work reported here supplies an additional alternative for judging teaching practice relevant to multilingual (ML) learners. Our practical problem was development within 1 year of a single instrument for observing teaching practice that would be suitable for the two projects' two randomized controlled trials (RCTs) encompassing ML students across the pre-K–12 age range. Our research question for this article, therefore, was "What evidence supports the reliability and validity of the instrument developed for the project?"

Despite the contextual limitations, instrument development was rigorous, including orientation to existing ML standards (nationally and in Ohio), research review, initial pilot testing including assessing interrater reliability (IRR), the development of training materials (to establish coding rules), and using initial administrations to confirm IRR and to improve the training materials (see Thompson et al. 2024). The two UC/SDI projects are part of a handful of NPD efforts incorporating

© 2025 TESOL International Association.

randomized controlled trials (RCTs) in their evaluation plans. The RCT, using the SCOPE-ML, will test (at the end of the project) whether or not improved teaching practice for ML students, including early literacy instruction in English (Morita-Mullaney et al. 2023) could be documented (via a post-only design). Each RCT tests the hypothesis differently.

2 | Background and Related Literature

The Center's organizational commitments and culture provide a logical home in Ohio for work to improve instruction for ML students. In brief, the mission of the SDI Center, and its varied projects, is to support equity efforts in districts and schools, with a primary focus on improving learning opportunities and outcomes for student groups that are traditionally marginalized in American schools.

2.1 | The Two NPD Projects at the University of Cincinnati

The two NPD projects at UC/SDI, Evidence-Based Practices for English Learners (E4L) and Early Learning for Multilinguals (ELM), help Ohio teachers earn Teachers of English to Speakers of Other Languages (TESOL) endorsements. Both 1-year programs deliver the specified TESOL coursework online.[2] Both programs also include a series of 10 online modules that focus attention on literacy instruction for MLs (see Appendix A). That is, on one hand, the coursework prepares students for teaching ML students regardless of discipline, and on the other hand the 10-module sequence prepares educators to deploy science-of-reading practices specifically with ML students, an adaptation strongly needed according to Morita-Mullaney et al. (2023).[3]

The first project (E4L) enrolls 160 participants, whereas ELM enrolls 240 participants in two separate treatment conditions (i.e., as compared to a single treatment condition in E4L). Participants include both special and general education teachers—grades Pre-K–12 for E4L and grades Pre-K–5 for ELM.

The RCTs for both projects compare control and treatment groups based on observed teaching practice (from videos submitted by teachers). The two UC/SDI projects together examine the possible influence on teaching practices in three combinations: (1) modules and coursework jointly (E4L), (2) coursework alone (ELM), and (3) modules alone (ELM). Beyond the RCT, both projects also collect pre- and post-participation data on literacy learning using two instruments developed by WordFarmers Associates, the third-party evaluator for both E4L and ELM.[4]

2.2 | Experimental Research on Online Courses and Professional Development Modules

In reviewing education databases for terms combining "online teacher education" with "randomized controlled trials," we located just one controlled study, which did not in fact use random assignment (Papageorgiou et al. 2023). A broader search without "online" yielded Lindell's dissertation (Lindell 2013), an RCT comparing role-play and online courses' effects on knowledge, competence, and practice, but not on actual teaching. Likewise, searches for online modules in professional development found two pertinent studies (Jimenez et al. 2016; Motoca et al. 2014) that focused on pedagogy but without analyzing instructional practice. The limited relevant literature suggests that the real-life change in teaching through either coursework or modules has received scant empirical attention (or none).

3 | Theoretical Framework

Although teaching practices are easy to observe impressionistically, systematic observation is much more challenging. But that was our evaluation task, and our team sought to develop an instrument that could characterize postparticipation group differences in teaching practice across classrooms with ML students.

Notably, to orient development, we also understood that practices responsive to ML students were *part of good teaching in general*. How to handle the overlap (i.e., with good teaching in general) was, therefore, of practical as well as theoretical interest.[5]

Although our challenge—as third-party evaluators and instrument developers—was practical and evaluative, the UC/SDI program team included experts in multilingual education and literacy education[6] who brought with them a shared commitment to the principles of scaffolding—not as a fixed methodology, but as an instructional outlook. That outlook holds that effective teaching for ML students depends on consistent, intentional supports that build on students' linguistic, cultural, and cognitive assets. Scaffolding becomes a way of framing instructional decisions: not merely a technique, but a disposition that resists deficit-based assumptions and challenges implicit biases. Although the SCOPE-ML does not operationalize "scaffolding" directly, it informed our efforts to identify observable classroom moves that can plausibly support multilingual learners' engagement with content and language together. For this work, then, the UC/SDI team's expertise and commitment to linguistic and cultural understanding motivated the development effort in the evaluation, but also, of course, the operation of its two NPD projects. All the work was oriented to prevailing standards, particularly Ohio's English Language Proficiency Standards (Ohio Department of Education 2015), *The 6 Principles for Exemplary Teaching of English Learners* (TESOL International Association 2018), the English Language Proficiency Assessment for the 21st Century Standards (ELPA21.org), and the relevant counsel from the What Works Clearinghouse (Baker et al. 2014; Gersten et al. 2007). These standards do not make artificial distinctions based on age or subject matter but rather assume that effective instructional strategies benefit multilingual learners in a broad range of contexts. They contributed appropriate guidance for item focus (e.g., teaching academic vocabulary, integrating language instruction into content-area teaching, structuring peer interaction).

3.1 | Related Measures

To expand and deepen our thinking, we first examined six existing observation instruments related to teaching ML students. These, in reverse chronological order were:

- *Modified Framework for Teaching* (MFT; Coady et al. 2020);
- *Teaching Observation Record* (Tang et al. 2020)
- *Classroom Quality for English Language Learners* (CQELL; Goldenberg et al. 2014);
- *Sheltered Instruction Observation Protocol* (SIOP; Guarino et al. 2001; Echevarria and Vogt 2010);
- *Timed Observations of Student Engagement/Language* (TO/SEL, Foorman and Schatschneider 2003); and
- *English Language Learner Classroom Observation Instrument* (ELLCOI, Haager et al. 2005).

This examination and reflection led to the following results: first, in Table 1, with details about each instrument, and then, narratively, the overall lessons from this exercise. These "lessons" are the practical pointers gleaned from the review, and show the relevance of the review to how we oriented our work to creating the SCOPE-ML. A short list of five lessons proved influential to our thinking, as follows:

1. Purposes (logically enough) orient to project-specific concerns (see also, Derrick 2016; Garza 2015; Mahan 2022).
2. Validity of use in other projects proceeds from purposes, and variable purposes make transfer across projects challenging (see also, Coady et al. 2020; Garza 2015).
3. Observation instruments for MLs' reading instruction are rare (see also, Whitacre et al. 2013).
4. Good instruction in general is part of good instruction for ML students (see also, Goldenberg 2013; Mahan 2022; Solano-Campos et al. 2020).
5. Technical decisions about inference level and rating categories varied (likely related to purpose).

The information in Table 1, the lessons derived from inspection of instruments, and the related reading and reflection supported the need to develop an instrument specifically designed for the two NPD projects, consistent with the commitments made in the UC/SDI Center proposals. What we sorely needed, however, was a theoretical framework to guide instrument development in our particular context.

3.2 | Selecting a Relevant Framework

The team members most directly responsible for instrument development (Showalter and Shin) discussed the insights from the instrument review and, on that basis, began by independently considering possible frameworks. Showalter is a mathematics educator; Shin, guided by her ML expertise, soon oriented to a promising new framework (Mahan 2022).

Indeed, the congruence of Mahan's conception with our commitments, PD purposes, and instrumentation needs seemed substantial (e.g., Shin et al. 2021). In the US, viewing the learning of an additional language *as an asset* is critical for helping ML students master English (Qin et al. 2025). In far too many US schools, speaking languages other than English marks students as socially, culturally, and intellectually deficient (Bertrand and Marsh 2021; Cummins 2000). It's a prejudice that derives from xenophobia and requires in-depth reflection of biases to identify and change (Orfield 2001; Spring 2016). Mahan's conceptualization of *scaffolding* in ML

TABLE 1 | Observation instruments examined.

Instrument	Relevant observations
MFT (Coady et al. 2020)	Adaptation of the Danielson framework, using 2 of 4 domains (instructional practice, classroom climate), to evaluate individual teachers' ML-related practice. Few items ($n = 10$); high $\alpha = 0.97$; 4-step rubric. Observations in situ; no training materials; average IRR (percentage) 81%
TOR (Tang et al. 2020)	Measures fidelity of implementation among teacher groups in RCTs. 9 items (one item related to ML students). 4-step rubric. IRR (as ICC) = 0.51. Considerable training documentation
CQELL (Goldenberg et al. 2014)	Research instrument to discover teaching practices related to MLs' achievement. Many observers involved ($n = 20$); many items ($n = 88$) and categories ($n = 14$); five-step rubric (0–4). Substantial training materials online. Observations in situ; training with video
SIOP (Echevarria and Vogt 2010)	Part of the SIOP professional development system for content-area teaching in general. Instrument used to assess fidelity of implementation among individual teachers. 8 domains; 30 items; 5-step rubric. $\alpha \geq 0.90$ (Guarino et al. 2001). Extensive training materials
TO/SEL (Foorman and Schatschneider 2003)	Not ML specific. Modification of 1996 instrument measuring time allotment; 2 domains (format, content); 25 items; no validity or reliability information, also included separate 10-item dichotomous scale to assess a whole lesson; IRR > 0.80 (type of IRR not specified)
ELLCOI (Baker et al. 2005)	Research instrument, described as "medium-inference," to investigate ML instruction and achievement in grade 1 classrooms. 6 categories; 29 items; 4-step rubric; median IRR = 74%; median $\alpha = 0.89$ (for the 6 categories)

Abbreviations: α, internal consistency reliability; ICC, intraclass correlation; IRR, inter-rater reliability.

instruction struck the full team as promising for four specific reasons, explained next.

First, Mahan oriented to an outlook uniquely appropriate to the NPD effort at the UC's SDI Center: "content and language integrated learning" (CLIL). Although CLIL is an outlook more European than American (Mahan 2022, 76), it represents well the approach of the ALL Ohio ML modules—teaching English (language) while also teaching early literacy (content).[7]

Second, "scaffolding"—as developed by Mahan—applies to both language and content learning (Bravo et al. 2014). Conversational fluency grows through socially situated interaction, which schooling can either support or inhibit (Cummins 1981, 2021; Pérez Castillejo 2023; Washington-Nortey et al. 2022). Effective ML instruction scaffolds this social accretion (Bravo et al. 2014; Madigan et al. 2019; Washington-Nortey et al. 2022), as reflected in the ML modules and supporting literature (Cummins 2021; Goldenberg 2013; Mahan 2022). This framing redefines teaching as an invested, facilitative act—contrary to the detached delivery model critiqued by Qin et al. (2025).

Third, scaffolding has a strong theoretical lineage rooted in constructivist thought, which holds that language and learning are socially constructed. The term was coined by Jerome Bruner (1978), drawing on the earlier work of Vygotsky (1962/1934), whose "zone of proximal development" remains widely cited. For Vygotsky, scaffolding could come from any more knowledgeable other—not just a teacher, but a peer or partner capable of supporting the learner's next steps.

Finally, Mahan (2022, 75) articulated her scaffolding framework as particularly suitable for coding videos of teaching practice. These four reasons convinced the team that Mahan's framework fit our design needs precisely. Additionally, apart from its suitability for our narrow purposes, Mahan's framework also aligned well with existing standards for effective instruction of ML students.[8]

3.3 | Description of Mahan's Scaffolding Framework

As just noted, Mahan argued the need for a scaffolding framework specifically for video observation (Mahan 2022, 76) and subsequently provided one (see the first two columns in Table 2), with two major divisions: (1) comprehension strategies and (2) task-solving strategies. The significance of the major divisions is transparent: (1) comprehension scaffolds help learners understand new material, and (2) task-solving scaffolds help learners complete tasks.

TABLE 2 | Mahan's scaffolding framework.

Category	Subcategory
Comprehension strategies	A. Connections to prior knowledge
	B. Supportive materials
	C. Academic language
Task-solving strategies	A. Uptake of student responses
	B. Strategy use & instruction
	C. Modeling and use of models

The Mahan framework notably emphasizes learners' agency and engagement: "students learning second languages are not empty vessels" (Mahan 2022, 76). In the context of industrial-style mass education (e.g., Fullan et al. 2019; Nguyen et al. 2018), this commitment has proven tremendously difficult to activate widely, particularly for the benefit of ML students (Mahan 2022). In this context, scaffolds are useful because, as Mahan (2022, 76–78) observed:

- students' existing knowledge merits teachers' respect, and drawing on it is a key instructional move that affirms what learners already bring to the classroom.
- "academic language" in fact represents unfamiliar vocabulary to *both* MLs and English-only learners: The difference is not vocabulary knowledge so much as it is cultural context—and, according to Cummins (2000), social and cultural power;
- ancillary materials (e.g., like those recommended by Universal Design for Learning—graphic organizers, video clips, and so forth) help supply a richer context that activates the assets that ML students bring to learning content;
- teacher-talk (revoicing, repeating, and elaborating) should anticipate the language needs of ML learners (not a disposition acquired in preparation programs for content teachers); and
- scaffolding is needed to help ML students acknowledge and *activate their own agency* as learners (thus closing the circle with "drawing on previous knowledge").

Scaffolding reflects a socially constructed view of teaching that emphasizes active instructional support. CLIL broadens this stance across subject areas by integrating language into content teaching. While such principles benefit all students, they are vital for MLs. Our observation items reflect this reality, aligning with standards that promote academic vocabulary, peer interaction, multimodal supports, oral and written language use, and encouragement of students' full linguistic repertoires.

Table 2 presents the entirety of Mahan's scaffolding framework, which we applied as a guide to specifying observable teaching practices. This effort—straightforward, but not easy—receives extended discussion in the Methodology section, which follows Table 2.

4 | Methods[9]

This section describes how we created the initial instrument using the guidance of Mahan's framework. First it details our instrumentation goals and related approaches to the specific design work. Next it considers the design process as executed and presents the set of items to be used in coding teacher videos. All

of this might, in some contexts, be considered "results," except that our presentation of findings centers on quantitative results from piloting and using the instrument. So, this section reports an account of the methods we used to create the SCOPE-ML for E4L and ELM.

Apart from creating items that would reflect contemporary standards of practice for ML teaching and learning (source of the instrument's face validity), we were most concerned to produce an observation instrument with arguable reliability. Two well-established metrics (applications of Cohen's kappa[10]) were used to measure reliability (Graham et al. 2012). The first measured inter-rater reliability (IRR) and the second was applied at the item level.

Developing a new instrument is a significant undertaking: it requires time, resources, and expertise in instrument development and psychometrics. The process involves piloting and refining the instrument, ensuring it meets the necessary standards for reliability and validity (Pae 2012). In our case—practical evaluation work—serious limitations applied (described in the discussion and in the limitations sections).

4.1 | Instrumentation Goals

Despite the theoretical and practical advantages, we needed to apply the framework to a context quite different from the one engaged when the Mahan framework was initially developed. We needed a wide-range instrument because our North American[11] context included multiple grade levels (pre-K–12), all content (e.g., mathematics, science), *and* it embedded a particular connection (via the online modules) with literacy instruction.

This observation raises the issue of pedagogical alignment, namely the nature of the NPD curriculum for these projects, particularly the grade levels and content involved. The first NPD grant enrolled upper elementary and secondary grade teachers across all subjects, whereas the second project enrolled early childhood and early elementary teachers. The main curriculum was UC's TESOL endorsement program, but for the second project, the module-only treatment condition focused on adapting a science-of-reading version of literacy to the ML context.

In short, we needed a very wide-range instrument. And we also hoped to develop and report an instrument that might prove useful beyond our own NPD projects. Our primary goal, then, was an observation instrument based on the Mahan framework that would prove simple and widely applicable. The primary goal articulated with two secondary goals: *accessibility* and *accuracy*, described as follows:

- Accessibility means (a) limiting the burden of data collection, (b) providing efficient and effective training for data analysts, (c) making reasonable the time requirements for coding, and (d) producing high interrater reliability (IRR).
- Accuracy means final ratings that validly represent the essential evidence-based practices for teaching. Arguing for accuracy, then, would rest on (a) the evident strength of the Mahan framework, (b) the standards-based nature of the items, and (c) evidence of IRR.[12]

Guided by the Mahan framework, we began to create an observation instrument to assess teachers' scaffolding of language input to promote ML students' understanding of content (e.g., use of visual aids like infographics) and learning process (e.g., use of modeling techniques). The TESOL International Association's principles (*The 6 Principles for Exemplary Teaching of English Learners* 2018) were a substantial help in the instrument development process, an advantage explained later in this section.

4.2 | Approaches to the Design Task

To address the context of multilingual learning, we extended Mahan's (2022) set of scaffolds. We retained her two broad types of strategies (comprehension and task-solving strategies). We also kept her related subcategories (see Table 2). The framework seemed close to mutually exclusive and jointly exhaustive. It did not, however, include items for the actual observation of teaching. So, we were left with the substantial task of translating these six theoretical divisions into a set of items (or "codes") that could be applied to our video-coding task. In particular, creating an *accessible instrument* meant mitigating the usual burdens of data collection, analyst training, and analysis.

Two approaches seemed useful: (1) framing items[13] as yes-no choices ('dichotomous coding') and (2) gathering data from teacher-made videos. Cost–benefit analysis favored dichotomous coding, as we explain next.

4.2.1 | Yes-No Approach

By comparison with individually oriented instruments designed to parse levels of a practice, a yes-no decision provides less nuance and less variability, but mitigates the training burden, the coding burden, and (because of reduced variability) facilitates inter-rater reliability. Limitations of budget and time were operant, as well: particularly in the context of an evaluation effort, where pedagogical nuance was not the focus. The trade-offs worked for our purposes.

4.2.2 | Video Approach

Given the projects' evaluation budgets, we realized from the beginning that in-person visits would be prohibitively expensive (hundreds of participants distributed across an entire state). Video observation was the logical alternative.

Although we anticipated the use of video data, we did not anticipate the possibility that teachers might be able to record video *themselves*. In our conversations, though, members of the project team *did* have experience with helping teachers record videos of their own teaching, and we leveraged these experiences along with a pilot video-capturing process to create a final process for obtaining high-quality videos.

Finally, although possibly useful, we decided against gathering and analyzing documentary data (e.g., lesson plans) from teachers. The increased burden on participating teachers would, we agreed, compromise the goal of accessibility.[14]

4.3 | Design Process

We aimed to create a concise, yes-no instrument (20–30 items) that captured Mahan's six subcategories. Observable scaffolds were derived from the TESOL International Association's 2018 teaching standards, which align with Mahan's strategies.

4.3.1 | Item Creation

The design subgroup[15] mapped Mahan's categories (Table 2) to observable practices (Table 3). After several revision cycles—adding, dropping, and coding scaffolds—we produced an initial item pool for the NPD projects at the SDI Center.

Three key issues guided the final refinement:

1. **Focus on evidence-based practices for multilingual (ML) learners**—We consulted the Ohio ELP, TESOL's six principles (2021), and language-acquisition norms. From 23 candidate practices, we retained 13 that are uniquely beneficial for ML students (e.g., use of non-English languages, cognates) and 10 general instructional practices (e.g., academic recasting, text-feature explanation).
2. **Weighting of items**—Mahan (2022) does not prescribe equal prevalence across subcategories, and our instrument is intended to capture broad instructional differences (e.g., between RCT groups) rather than individual profiles. Consequently, we omitted weighting, though it could be added in future, more granular applications.
3. **Item independence from supplemental teacher information**—To keep the tool usable without extra teacher input or observer inference, we accepted minor delimitations that limit student behavior analysis but serve the project's purpose.

4.3.2 | Pilot Testing

Three nonparticipant teachers each recorded two classroom videos (average 14 min, range 10–21 min) featuring ML students. Two team members coded the videos independently, reconciled discrepancies, and consulted a second language-acquisition expert to refine the rubric for greater validity and reliability. The resulting finalized items appear in Table 4; detailed justifications are in Appendix B (the revised SCOPE-ML).

4.3.3 | Reliability Assessment

Inter-rater reliability (IRR) statistics (see Findings) were promising. Following the pilot, we refined the coding manual and prepared training materials for coders and teachers, positioning the finalized SCOPE-ML for immediate research use.[16]

4.4 | Instrument Training for New Team Members

Development of the coding manual proved very useful. After the pilot phase and initial use with NPD participants, new members joined the team as additional coders. Their training regimen included an initial meeting with one of the developers, practice coding of three videos, and a second meeting to discuss the coding experience. On this basis small adjustments were made to the coding manual. A second practice round seemed advisable, this time with four videos. IRRs from the practice rounds were sufficiently high that further training was not needed.

TABLE 3 | Mahan's framework represented as specific scaffolds from the TESOL interventional association's (2018) teaching standards.

Category	Subcategory	Observable scaffolds
Comprehension strategies	A. Connections to prior knowledge	T links to another culture or language
		T prompts Ss to use background knowledge of another culture or language
	B. Supportive materials	T uses techniques to increase clarity
		T uses other tangible props/aids
	C. Academic language	T uses literacy techniques like word recognition or decoding strategies
		T preteaches or highlights key vocab/concepts needed to understand a lesson
Task-Solving Strategies	A. Uptake of student responses	T uses a variety of techniques to elicit language production
		T recasts students' responses using academic language
	B. Strategy use and instruction	T describes a metacognitive strategy that uses understanding of language structure to complete a task
		T prompts Ss to use a previously discussed metacognitive strategy related to language structure
	C. Modeling and use of models	T provides a sample or model response

Abbreviations: S, student; T, teacher.

TABLE 4 | Finalized item set.

Subcategory	Observable scaffold	Item
A. Connections to prior knowledge	T links to another culture or language	1. T uses a non-English language or mentions a difference between another language and English
		2. T uses an example from a culture where English is not the dominant language (aside from an example based solely on the language)
	T prompts Ss to use background knowledge of another culture or language	3. T prompts S to use a language other than English
		4. T prompts S to share their knowledge or experience of another culture where English is not the dominant language
B. Supportive materials	T uses techniques to increase clarity	5. T mimes or gestures in a purposeful way
		6. T speaks slowly and clearly
		7. T repeats or rephrases key information
	T uses other tangible props/aids	8. T uses a graphic organizer
		9. T gives instructions in both written and oral form
		10. T previews text features to increase understanding
		11. T highlights key words or phrases in *written or oral* form during the task
C. Academic language	T uses literacy techniques like word recognition or decoding strategies	12. T chunks a word into smaller pieces to sound out individual phonemes or helps students blend phonemes
		13. T helps students use morphemes to understand the meaning of a word
	T preteaches or highlights key vocab and/or key concepts needed to understand a lesson	14. T gives explicit vocabulary instruction for an upcoming task *with opportunity to practice*
		15. T identifies academic word as a cognate (or warns against false cognates)
A. Uptake of student responses	T uses a variety of techniques to elicit language production	16. T prompts S to elaborate on their answer
		17. T uses questions that elicit language production
	T recasts Ss' responses using academic language	18. T recasts S's response or question using academic language
B. Strategy use and instruction	T describes a metacognitive strategy...	19. T describes a language-based metacognitive strategy that students can use to complete a task
	T prompts S to use a ... previously discussed metacognitive strategy...	20. T prompts s to use previously discussed metacognitive strategy related to language
C. Modeling and use of models	T provides a sample or model response	21. T provides sentence stems or frames
		22. T provides an exemplar of the assignment
		23. T models language production for a task

Note: Refer to Table 2 for categories, subcategories, and observable scaffolds; some of the observable strategies in Table 4 have been shortened. Each item is coded dichotomously.
Abbreviations: S, student; T, teacher.

Nonetheless, further discussion of the two "metacognitive" items (#19 and 20)—a challenge for existing coders—yielded additional changes to the coding manual for these items, based partly on a review of all the previous coding related to them. Training the new team members was thus an additional opportunity in the instrument development process. Project time to accomplish this training was about 7 h for the trainer and about 12 h for each of the new coders.

5 | Results

This section reports the reliability results for three uses of the SCOPE-ML: The two uses reported in the previous section, and, most notably, the first use of the instrument at the end of the 1-year program for the first E4L cohort—thus inaugurating the accumulation of data for that project's RCT. After discussing our IRR metrics, full details of the results of these three

uses follow—first for our experience with the pilot videos, and second with the first E4L cohort.

As noted previously, two applications of Cohen's kappa were used to measure reliability: the first measuring IRR and the second applied at the item level. We hoped to achieve at least the value of $\kappa = 0.60$ recommended by Landis and Koch (1977) for the full instrument.

5.1 | Pilot Phase Results

In the initial rating of six videos during the pilot phase, two coders achieved an agreement rate of 84.1%. In other words, there were 23 dichotomous codes that could be triggered for each of the six videos, yielding a total of 138 possible points of agreement. The two raters agreed in 116 of these 138 instances (84%). Cohen's kappa for interrater reliability in the pilot phase was 0.57, just shy of the 0.60 suggested by Landis and Koch (1977) as acceptable. After this initial pilot coding, the research team modified the SCOPE-ML and coding notes.

The second training episode, for new team members, involved three coders: an anchor (who helped develop the instrument and participated in the initial coding) and the two trainees. Because Cohen's kappa requires comparison of two coders at a time, three pairwise comparisons were made: (1) anchor and trainee 1; (2) anchor and trainee 2; (3) trainee 1 and trainee 2. The agreement rates were 91% for the anchor and trainee 1, 88% for the anchor and trainee 2, and 88% for trainees 1 and 2.[17] The Cohen's kappa values for IRR were 0.82, 0.74, and 0.75, respectively, with an average of 0.77 across the three pairs—substantially greater than the 0.60 target. It seems likely that the further work to prepare the SCOPE-ML for initial use also strengthened it—and provided improved training materials.

5.2 | Initial Results With E4L's Cohort 1

During the development process, the research team modified the rubric and training manual notes. These changes likely helped improve it substantially, as reported shortly. For data gathered at the conclusion of E4L's first cohort, 43 instructors submitted video clips of their instruction. These clips ranged in length from 3 to 30 min, with an average length of around 17 min. Of the 43 videos, 35 (81%) had a duration between 15 and 25 min. Two trained team members then independently coded each of these 43 videos, blind to each other's coding and blind to each of the educators' randomly assigned research condition.

For this first research use (Summer 2023), the two coders agreed in 891 of the 989 instances, yielding a high agreement rate of 90%. Cohen's kappa for IRR in the first year was also high at 0.78, which is well above the target kappa reliability rating of 0.60. Data were archived for use in the RCT and will be formally analyzed after data from all four cohorts have been collected and analyzed.

5.2.1 | Item-Level Reliability From Cohort 1 Data

IRR, rather than item-level reliabilities, is the usual concern in this type of instrument development research. However, for the purpose of a thorough analysis, we turn our attention to this feature of the SCOPE-ML next.

Item-level reliability is typically lower than IRR. For reference, the average item-level kappa of Goldenberg (2013) items on the CQELL was 0.32. In the pilot phase of the current project, the average item-level kappa was 0.15. This statistic improved substantially in the Cohort 1 analysis. For the 43 videos in the analytic sample, the kappa values ranged from −0.04 to 1.00, with an average of 0.64. A complete listing of items and kappa values, along with frequency and rater agreement statistics appears in Appendix C.

5.2.2 | Reflections on E4L's First Cohort Results

While the kappa values for the SCOPE-ML suggested strong IRR, the item-level analysis seemed to reveal areas for potential improvement. The research team met to discuss whether the results warranted modifications to the instrument.

For example, two items (#2 and #4) had a negative kappa (see Appendix C). The team noted that the issue was not one of clarity of item wording; both of these items had percent agreement of above 90% between the researchers. Instead, the reason for the low kappa was that the items were observed in fewer than 4% of the videos. Because of the theoretical importance of valuing nondominant cultures (e.g., González et al. 2005), the team decided that the items should remain part of the SCOPE-ML.

The team also discussed items that were relatively low in terms of percent agreement (see Appendix C) between the raters (e.g., #8, #11, #23). To raise the percent agreement, clarifications were made to the training manual, resulting in the final version of the manual (available from the lead author). Thus, no changes were made to the actual items as a result of debriefing after the first research use of the instrument.

6 | Discussion

Effective instruments for promoting ML students' learning require careful consideration of the learner, learning goals, and desired data collection methods. Of importance are the instrument's trustworthiness and effectiveness. Existing observation instruments, such as MFT, TOR, CQELL, and TO/SEL, have laid foundations for the understanding of teacher–student interactions in classrooms. They were not, however, directly suitable for our two NPD projects because the objectives and learners were different. Moreover, as Whitacre et al. (2013) indicated, observation instruments for ML students' literacy instruction have been scarce.

To address this gap, we analyzed Mahan's (2022) framework along with the TESOL International Association's (2018) teaching standards as a guide for developing an observation instrument directly applicable to our two NPD projects that aim to enhance

ML students' learning experiences and outcomes. We constructed 4 items for the *Connection to Prior Knowledge* subcategory, 7 items for the *Supportive Materials* subcategory, and 4 items for the *Academic Language* subcategory, totaling 15 items under the Comprehension Scaffolding Strategies of the Mahan framework. We also developed 3 items for the *Uptake of Student Responses* subcategory, 2 items for the *Strategy Use and Instruction* subcategory, and 3 items for the *Modeling and Use of Models* subcategory, totaling 8 items, for the Task-Solving Scaffolding Strategies of the Mahan framework. This resulted in 23 items in total for the full observational measure. By recognizing the limitations of existing observation instruments and developing a new measure based on the unique objectives of our grant, we conclude that we can confidently use the new tool to determine whether our intervention had the desired impact on educators' teaching of ML students. These customization efforts were essential for the accurate assessment of target pedagogical practices at the classroom level.

In the observation measure's refinement stage, we examined the psychometric properties of the rubric through reliability and validity testing and face validity through peer debriefing and feedback. The results showed that the mean rate of IRR for the 23 items was 90%, ranging from 72% to 100%. After piloting, we revised the SCOPE-ML to ensure that it accurately measured the intended classroom activities and provided valuable data for project evaluation. Through this iterative process, a final set of items was constructed with better clarity, usability, and appropriateness. This refining process produced a robust instrument for use in our NPD projects and fills a gap in the literature by contributing another tool that can advance knowledge in the field to promote effective teaching and learning for ML students.

6.1 | Strengths

There were several strengths involved in the instrument development process. First, as existing instruments did not perfectly capture the specific demands of our two NPD projects, the development of the new tool allowed us to design a new measure to precisely assess classroom dynamics and teacher–student interactions in a way we aimed to understand. This new measure was guided by established frameworks and standards and was vetted by an external ML expert, which produced an instrument that was theoretically grounded and aligned with best practices for teaching ML students. Secondly, the existing observational tools relied on traditional data collection methods that were not necessarily accessible or feasible in applied research-based inquiries. Developing the new instrument allowed us to create an observation tool that was efficient and produced trustworthy data on the interactive experiences between the teacher and the student. Lastly, existing instruments did not cater to specific ML student populations and learning environments. Developing the new instrument allowed us to address these gaps and create a tool that is particularly relevant to ML students.

6.2 | Use in Other Contexts

Contemporary instrument development standards require both a validity and a use argument (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (U.S.) 2014). The prescribed use of the SCOPE-ML—as developed and reported here—does not extend to other contexts. It could, of course, be adapted to other such contexts, given additional development work.

Though the use of the instrument in its present form may be limited on this basis, the approach taken illustrates the path of possible adaptations. Furthermore, the SCOPE-ML has been developed as a sort of wide-range measure of content-area pedagogy (across all grades and school subjects). That feature (or bug) offers comparatively rich opportunities for extending the work if development rigor is applied. For instance, the dichotomous scheme could be rendered polytomous. The SCOPE-ML could be further adapted to in vivo settings instead of video recordings. Given far fewer time and money constraints, it could be adapted to investigating basic questions of language instruction. Thus, we see the uses as immediately applicable to designing evaluations of a range of ML projects, and less to applications such as leadership training or coaching. These would be worthy purposes, but they would require a far different development trajectory. The Mahan framework, however, and perhaps some of the items, might prove useful in such contexts.

6.3 | Limitations and Future Research Directions

Rather than a purely theoretical inquiry, the development and validation of this observation instrument emerged from the needs of evaluating a professional development program, necessitating trade-offs between ideal research conditions and the applied demands of educational settings. Fiscal and time limitations constrained development to essentials—ensuring face validity via adherence to standards and then prioritizing reliability. The constraints suggested dichotomous coding (accessibility) over polytomous coding (nuance). Accessibility facilitated the training of coders at the cost of losing theoretically (but not practically) useful data. This is a dilemma common to the realities of evaluating real-world implementations, especially with RCTs (Glennerster and Takavarasha 2013).

Since existing tools do not directly map onto new instruments, comparison of our results with previous research using established instruments is problematic. This conundrum limits the generalizability of our results. Since the existing tools did not, however, adequately address our team's needs, the benefits of the custom-designed instrument outweighed the potential limitations. But the limitations nonetheless persist.

Establishing validity and reliability for new instruments is not an easy feat. Researchers need to evaluate the instrument's psychometric properties rigorously to demonstrate its trustworthiness in measuring L2 learning (see Pae 2012). Additional psychometric work would be necessary when adapting the tool across diverse populations and in additional contexts. Although the work of the SDI Center's RCTs is statewide in nature, they involve schools in only one American state.

Moreover, although we used a rigorous process to develop and pilot the SCOPE-ML, the pilot data presented are from one

cohort of two four-cohort studies, and additional psychometric evaluations of the tool will be warranted with the full datasets as the NPD work wraps up.

The challenges described above point towards areas of improvement that we will seek to address throughout our current studies. During the rest of the NPD projects' life span, we will continue to use a cyclical process of refining the tool and its related materials to develop a more refined version of the instrument. This effort could involve such activities as modifying items, combining elements across items, more clearly defining operational definitions of items, or revising the training process. The multicohort design of our research affords the opportunity for us to thoughtfully consider the strengths and shortcomings of the tool and guide us towards the most effective approach for evaluating the effectiveness of our ML projects.

Specific to the generalizability of the SCOPE-ML, the diverse nature of ML student populations necessitates that the tool be used in other states and in other cultural and linguistic contexts. Multistate studies and studies that include students with a variety of home languages can shed light on the validity of the tool in other settings and with other populations.

Notably, there has been a lack of research regarding ML students with dyslexia. Evaluating and adapting this instrument to systematically address this group of learners is a meaningful next step for assessing the pedagogical practices needed to accommodate the students' learning needs while also assessing their English development.

Funding

This study was supported by grants T365Z210007 and T365Z220055 from the U.S. Department of Education. (Originating PI: Jennifer Ottley). The views expressed herein do not necessarily represent the positions or policies of the U.S. Department of Education. No official endorsement by the U.S. Department of Education of any product, commodity, or enterprise mentioned in this document is intended or should be inferred.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Endnotes

[1] WordFarmers Associates is the third-party evaluator for these projects, and it led the instrument development process reported here. Program leaders advised the effort and participated in the drafting of this article.

[2] *Foundations in Learning L1 & L2*; *Teaching Methods and Strategies for TESOL*; *Assessment and Evaluation for TESOL*; *Applied Linguistics*; *Practicum for TESOL* Spring; see https://cech.uc.edu/schools/education/certificates-and-endorsements/tesolpage.html.

[3] Morita-Mullaney and colleagues also clearly note what is missing from "science of reading" for ML students, and the Ohio modules address the missing features.

[4] The evaluation scope is predictably larger than the RCTs, involving formative feedback about events, as well as the summative effort (RCTs). One of the instruments is also used to establish the initial equivalency of the control and treatment groups.

[5] A full description of practices assessed appears in the Methodology section.

[6] The proposal and the subsequent award featured attention to the "science of reading" approach, though both the evaluation team and the program team understood the shortcoming of SOR with respect to ML students (cf. Morita-Mullaney et al. 2023).

[7] In CLIL students learn academic subjects—like science or history—through a language other than their first language. The approach aims to build communicative competence simultaneously with subject knowledge so as to cultivate deeper engagement among ML students, and intercultural awareness and intellectual development among all students. CLIL is widely used across Europe to support multilingualism and academic equity in diverse classrooms. For a detailed treatment of CLIL see Coyle (2018).

[8] See, for example, Ohio's English Language Proficiency Standards (Ohio Department of Education 2015), *The 6 Principles for Exemplary Teaching of English Learners* (TESOL International Association 2018), or the English Language Proficiency Assessment for the 21st Century Standards (ELPA21.org). These standards do not make artificial distinctions based on age or subject matter but rather assume that effective instructional strategies benefit multilingual learners in a broad range of contexts.

[9] This section transparently reports how we created the SCOPE-ML, but the discussion is technical. For readers unfamiliar with such technicalities, here is a short summary of this section. First, we used an established framework (CLIL, see Mahan 2022) for thinking about teaching practices for ML students. We used it to identify what the new instrument would measure in general and in specific for creating things to look for. This was a difficult task because the SCOPE-ML needed to work with all subjects and all ages. We focused on identifying concrete teaching practices—such as using visuals or modeling—that support ML students' understanding. Our goal was to keep the tool simple and practical: It uses yes/no ratings based on short video recordings that teachers made of their own lessons. This made training observers easier and reduced the burden on teachers. We tested and revised the tool through a pilot study, refined the coding instructions, and trained new observers to ensure they could use the instrument consistently.

[10] Although there is no established standard for a minimum threshold of Cohen's kappa, Landis and Koch (1977) suggested aiming for at least 0.60 on an entire instrument. An average kappa can also be calculated at the item-level for the purposes of item analysis. This average is expected to be lower than the kappa for the overall instrument (e.g., in Goldenberg et al.'s (2014) analysis of the CQELL, the average item-level kappa was 0.32).

[11] Mahan worked with Norwegian 11th-grade students in science, geography, and social science.

[12] Because the SCOPE-ML would address a wide-range context, levels, instructional content, and pedagogical styles, however, we did not anticipate or design for a high level of internal consistency reliability.

[13] Because our design goals aimed towards a short instrument with dichotomous coding, we are comfortable referring to them as *items*.

[14] See the pre- and postobservation protocol accompanying the CQELL for an example of the requirements for gathering such documentary evidence.

[15] Subgroup members (Showalter, Shin, and Pae) have expertise in applied linguistics, curriculum and quantitative data analysis, and instrument design.

[16] These adjustments indeed proved effective, as detailed in the Findings section.

[17] In other words, there were 23 dichotomous codes that could be triggered for each of the four videos, yielding a total of 92 possible points of agreement for each pair (276 total for the three pairs). For the anchor and trainee 1, there was agreement in 84 of those 92 instances, and for both other pairs, there was agreement in 81 of the 92 instances.

References

Ajayi, L. 2014. "Vocabulary Instruction and Mexican–American Bilingual Students: How Two High School Teachers Integrate Multiple Strategies to Build Word Consciousness in English Language Arts Classrooms." *International Journal of Bilingual Education and Bilingualism* 18, no. 4: 463–484.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (U.S.). 2014. *Standards for Educational and Psychological Testing*. American Educational Research Association.

Baker, S., N. Lesaux, M. Jayanthi, et al. 2014. *Teaching Academic Content and Literacy to English Learners in Elementary and Middle School (NCEE 2014-4012)*. Institute of Education Sciences. https://ies.ed.gov/ncee/wwc/Docs/PracticeGuide/english_learners_pg_040114.pdf.

Baker, S. K., R. Gersten, D. Haager, M. Dingle, and C. Goldenberg. 2005. "Assessing the Relationship Between Observed Teaching Practice and Reading Growth in First Grade English Learners: A Validation Study." https://eric.ed.gov/?id=ED486282.

Bertrand, M., and J. Marsh. 2021. "How Data-Driven Reform Can Drive Deficit Thinking." *Phi Delta Kappan* 102, no. 8: 35–39.

Bravo, M. A., E. Mosqueda, and J. L. Solís. 2014. "Possibilities and Limits of Integrating Science and Diversity Education in Preservice Elementary Teacher Preparation." *Journal of Science Teacher Education* 25, no. 5: 601–619. https://doi.org/10.1007/s10972-013-9374-8.

Bruner, J. S. 1978. "The Role of Dialogue in Language Acquisition." In *The Child's Conception of Language*, edited by A. Sinclair, R. J. Jarvella, and W. J. M. Levelt, 241–256. Springer-Verlag.

Coady, M., M. D. Miller, Z. Jing, et al. 2020. "Can English Learner Teacher Effectiveness Be Observed? Validation of an EL-Modified Framework for Teaching." *TESOL Quarterly* 54, no. 1: 173–200. https://doi.org/10.1002/tesq.544.

Coyle, D. 2018. "The Place of CLIL in (Bilingual) Education." *Theory Into Practice* 57, no. 3: 166–176. https://doi.org/10.1080/00405841.2018.1459096.

Cummins, J. 1981. *Bilingualism and Minority-Language Children*. OISE Press.

Cummins, J. 2000. *Language, Power, and Pedagogy: Bilingual Children in the Crossfire*. Multilingual Matters.

Cummins, J. 2021. *Rethinking the Education of Multilingual Learners: A Critical Analysis of Theoretical Concepts*. Channel View Publications.

Derrick, D. J. 2016. "Instrument Reporting Practices in Second Language Research." *TESOL Quarterly: A Journal for Teachers of English to Speakers of Other Languages and of Standard English as a Second Dialect* 50, no. 1: 132–153. https://doi.org/10.1002/tesq.217.

Echevarria, J., and M. Vogt. 2010. "Using the SIOP Model to Improve Literacy for English Learners." *New England Reading Association Journal* 46, no. 1: 8–15.

Echevarría, J., M. Vogt, D. Short, and K. Toppel. 2023. *Making Content Comprehensible for Multilingual Learners: The SIOP Model*. 6th ed. Pearson.

Foorman, B., N. Beyler, K. Borradaile, et al. 2016. *Foundational Skills to Support Reading for Understanding in Kindergarten Through 3rd Grade (NCEE 2016-4008)*. Institute of Education Sciences. http://whatworks.ed.gov.

Foorman, B. R., and C. Schatschneider. 2003. "Measurement of Teaching Practices During Reading/Language Arts Instruction and its Relationship to Student Achievement." In *Reading in the Classroom: Systems for the Observation of Teaching and Learning*, edited by S. Vaughn and K. L. Briggs, 1–30. Brookes.

Fullan, M., M. Gardner, and M. Drummy. 2019. "Going Deeper: What Today's Teens Need Most From Schools Is Learning That Fosters Engagement and Connection. That May Mean Changing Everything." *Educational Leadership* 76, no. 8: 64–69.

García, O., S. I. Johnson, K. Seltzer, and G. Valdés. 2017. *The Translanguaging Classroom: Leveraging Student Bilingualism for Learning*. Caslon.

Garza, T. 2015. "Examining Classroom Observation Instruments for English Language Learners: An Example of Proposing Causal Structure Pertaining to Pedagogy." Dissertation. Texas A&M University.

Gass, S. M. 2015. "Comprehensible Input and Output in Classroom Interaction." In *The Handbook of Classroom Discourse and Interaction*, edited by N. Markee, 182–197. John Wiley & Sons. https://doi.org/10.1002/9781118531242.

Gersten, R., S. K. Baker, T. Shanahan, S. Linan-Thompson, P. Collins, and R. Scarcella. 2007. *Effective Literacy and English Language Instruction for English Learners in the Elementary Grades: A Practice Guide (NCEE 2007-4011)*. Institute of Education Sciences. https://files.eric.ed.gov/fulltext/ED497258.pdf.

Glennerster, R., and K. Takavarasha. 2013. *Running Randomized Evaluations: A Practical Guide*. Princeton.

Goldenberg, C. 2013. "Unlocking the Research on English Learners: What We Know—and Don't Yet Know—About Effective Instruction." *American Educator* 37, no. 2: 4–11.

Goldenberg, C., R. Coleman, L. Reese, E. Haertel, and C. Rodriguez-Mojica. 2014. *Classroom Qualities for English Language Learners in Language Arts Instruction*. William T. Grant Foundation. https://wtgrantfoundation.org/classroom-qualities-for-english-language-learners-in-language-arts-instruction.

Goldenberg, C. 2020. "Reading Wars, Reading Science, and English Learners." *Reading Research Quarterly* 55: S131–S144.

González, N., L. C. Moll, and C. Amanti, eds. 2005. *Funds of Knowledge: Theorizing Practice in Households, Communities, and Classrooms*. L. Erlbaum Associates.

Goo, J., and A. Mackey. 2013. "The Case Against the Case Against Recasts." *Studies in Second Language Acquisition* 35, no. 1: 127–165.

Grabe, W., and J. Yamashita. 2022. *Reading in a Second Language: Moving From Theory to Practice*. 2nd ed. Cambridge.

Graham, M., A. Milanowski, and J. Miller. 2012. "Measuring and Promoting Inter-Rater Agreement of Teacher and Principal Performance Ratings (ED532068; Online Submission)." http://files.eric.ed.gov/fulltext/ED532068.pdf.

Guarino, A. J., J. Echevarria, D. Short, J. Schick, S. Forbes, and R. Rueda. 2001. "The Sheltered Instruction Observation Protocol." *Journal of Research in Education* 11, no. 1: 138–140.

Jimenez, B. A., P. J. Mims, and J. Baker. 2016. "The Effects of an Online Data-Based Decisions Professional Development for In-Service Teachers of Students With Significant Disability." *Rural Special Education Quarterly* 35, no. 3: 30–40. https://doi.org/10.1177/8756870516035003305.

Landis, J. R., and G. G. Koch. 1977. "An Application of Hierarchical Kappa-Type Statistics in the Assessment of Majority Agreement Among Multiple Observers." *Biometrics* 33: 363–374.

Lindell, M. A. 2013. "The Effects of Microteaching on Pre-Service Teachers' Knowledge and Implementation of the Concept Mastery Routine (Publication No. 3596337) [Doctoral Dissertation, University of Minnesota]." ProQuest Dissertations & Theses A&I. (1448556560).

Lyster, R., and L. Ranta. 1997. "Corrective Feedback and Learner Uptake: Negotiation of Form in Communicative Classrooms." *Studies in Second Language Acquisition* 19, no. 1: 37–66.

Madigan, S., H. Prime, S. A. Graham, et al. 2019. "Parenting Behavior and Child Language: A Meta-Analysis." *Pediatrics* 144, no. 4: 1–12. https://doi.org/10.1542/peds.2018-3556.

Mahan, K. R. 2022. "The Comprehending Teacher: Scaffolding in Content and Language Integrated Learning (CLIL)." *Language Learning Journal* 50, no. 1: 74–88. https://doi.org/10.1080/09571736.2019.1705879.

Masrai, A., and J. Milton. 2018. "Measuring the Contribution of Academic and General Vocabulary Knowledge to Learners' Academic Achievement." *Journal of English for Academic Purposes* 31: 44–57.

Morita-Mullaney, T., F. Gao, and R. Zhang. 2023. "When Literacy Ate Language: An Analysis of a Research-Based Literacy Intervention for Designated English Learners." *TESOL Journal* 14, no. 3. https://doi.org/10.1002/tesj.718.

Motoca, L. M., T. W. Farmer, J. V. Hamm, et al. 2014. "Directed Consultation, the SEALS Model, and Teachers' Classroom Management." *Journal of Emotional and Behavioral Disorders* 22, no. 2: 119–129. https://doi.org/10.1177/1063426614521299.

Nguyen, T. D., M. Cannata, and J. Miller. 2018. "Understanding Student Behavioral Engagement: Importance of Student Interaction With Peers and Teachers." *Journal of Educational Research* 111, no. 2: 163–174.

Ohio Department of Education. 2015. "English Language Proficiency Standards." https://education.Ohio.gov/getattachment/Topics/Student-Supports/English-Learners/Teaching-English-Learners/Ohio-English-Language-Proficiency-ELP-Standards/ELP-Content-Standards-20150824-1.pdf.aspx?lang=en-US.

Orfield, G. 2001. *Schools More Separate: Consequences of a Decade of Resegregation*. Civil Rights Project, Harvard University.

Pae, H. K. 2012. "A Psychometric Measurement Model for Adult English Language Learners: Pearson Test of English Academic." *Educational Research and Evaluation* 18, no. 3: 211–229.

Papageorgiou, E., C. Krommidas, N. Digelidis, L. Moustakas, and A. G. Papaioannou. 2023. "An Online PETE Course on Intercultural Education for Pre-Service Physical Education Teachers: A Non-Randomized Controlled Trial." *Teaching and Teacher Education* 121: 1–12. https://doi.org/10.1016/j.tate.2022.103920.

Pérez Castillejo, S. 2023. "Prior Processing, Foreign Language Classroom Anxiety, and L2 Fluency." *IRAL: International Review of Applied Linguistics in Language Teaching* 61, no. 2: 519–544. https://doi.org/10.1515/iral-2021-0091.

Pressley, M. 2002. "Metacognition and Self-Regulated Instruction." In *What Research Has to Say About Reading Instruction*, edited by A. Farstrup and S. J. Samuels, 3rd ed., 291–309. International Reading Association.

Qin, K., J. Wolf, and L. Llosa. 2025. "From Everyday Language to Disciplinary Discourse: Culturally and Linguistically Sustaining Dialogic Instruction With Multilingual Learners." *TESOL Journal* 16, no. 1. https://doi.org/10.1002/tesj.913.

Richards-Tutor, C., T. Aceves, and L. Reese. 2016. *Evidence-Based Practices for English Learners (Document No. IC-18)*. CEEDAR. http://ceedar.education.ufl.edu/tools/innovation-configurations/.

Shin, J. K., V. Savic, and T. Machida. 2021. *The 6 Principles for Exemplary Teaching of English Learners: Young Learners in a Multilingual World*. TESOL Press.

Solano-Campos, A., M. Hopkins, and L. Quaynor. 2020. "Linguistically Responsive Teaching in Preservice Teacher Education: A Review of the Literature Through the Lens of Cultural-Historical Activity Theory." *Journal of Teacher Education* 71, no. 2: 203–217. https://doi.org/10.1177/0022487118808785.

Spring, J. H. 2016. *Deculturalization and the Struggle for Equality: A Brief History of the Education of Dominated Cultures in the United States*. Routledge.

Tang, S., F. Tong, B. J. Irby, R. Lara-Alecio, and C. Guerrero. 2020. "Fidelity of Implementation in a Randomized Controlled Trial Study: The Effect of Virtual Professional Development on Bilingual Teachers." *Bilingual Research Journal* 43, no. 1: 111–124. https://doi.org/10.1080/15235882.2019.1711268.

TESOL International Association. 2018. "The 6 Principles for Exemplary Teaching of English." https://www.the6principles.org/the-principles/.

Thompson, K. D., J. Greenberg Motamedi, M. Vazquez Cano, and N. Patiño-Cabrera. 2024. "Broadening Definitions of Impact: A Case Study of Competing Tensions When Evaluating the Impact of National Professional Development Grants." *TESOL Journal* 16, no. 1: e893. https://doi.org/10.1002/tesj.893.

Vygotsky, L. 1962. *Thinking and Speaking* (E. Hanfmann, G. Vakar, & N. Minnick, Trans.). MIT Press. https://www.marxists.org/archive/vygotsky/works/words/index.htm. Original Work Published 1934.

Washington-Nortey, P.-M., F. Zhang, Y. Xu, A. B. Ruiz, C.-C. Chen, and C. Spence. 2022. "The Impact of Peer Interactions on Language Development Among Preschool English Language Learners: A Systematic Review." *Early Childhood Education Journal* 50, no. 1: 49–59. https://doi.org/10.1007/s10643-020-01126-5.

Whitacre, M., Z. Diaz, and J. Esquierdo. 2013. "Pre-Service Teachers: An Analysis of Reading Instruction in High Needs Districts Dual Language Classrooms." *International Journal of Instruction* 6, no. 1: 5–20.

Yoon, B. 2007. "Offering or Limiting Opportunities: Teachers' Roles and Approaches to English-Language Learners' Participation in Literacy Activities." *Reading Teacher* 61, no. 3: 216–225.

Appendix A

	E4L and ELM science-based ML literacy modules	
Sequence	**Module title**	**Content description**
ML1	An Introduction to Multilingual Learners	Backgrounds of MLs to help educators better understand and meet MLs' needs
ML2	Every Teacher is a Language Teacher	How English works, how rule systems affect literacy development; what teachers can do to scaffold MLs' literacy development
ML3	Word Recognition: A Prerequisite for Reading Comprehension	Foundational skills for strong reading; pedagogy to support MLs' literacy development
ML 4	Fostering Reading Fluency	Reading fluency = accuracy, speed, prosody; ways to foster MLs' English reading fluency at the word-reading and passage-reading levels
ML5	How Language Comprehension Leads to Reading Comprehension	Multifaceted nature of reading comprehension, including language comprehension in literacy instruction
ML6	Implementing Interventions	Effective literacy interventions for MLs who struggle in English literacy development
ML7	Explicit Vocabulary Instruction	Cyclical process of teaching vocabulary; ways to make vocabulary learning effective, efficient, and entertaining for MLs
ML8	Integrating Language into Content Instruction	Instructional strategies in content area to scaffold MLs' language and content learning
ML9	Writing Instruction	Writing development of MLs with varying proficiency levels; implications of major research findings on writing instruction
ML10	Culturally Responsive Teaching	Culturally responsive education for MLs; implications for teaching that serves MLs well in classrooms, schools, and communities

Note: The completed modules are housed on the ALL [Blinded State] website (https://all[Blinded State]literacy.org/learning-modules/multilingual-learners-learning-modules/).

Appendix B

Observable scaffold	Item	Justification
T links to another culture or language	1. T uses a non-English language or mentions a difference between another language and English 2. T uses an example from a culture where English is not the dominant language (aside from an example based solely on the language)	MLs know cultures and languages different from those prominently featured in schooling–which is nearly always English in the US. Their knowledge of different cultures and languages is an asset to be leveraged for learning (e.g., González et al. 2005). Teachers' attempts to use such assets are helpful—both for ML students and for monolingual students as well (e.g., demonstrating the importance of cultural knowledge in general). The second code's restriction means that mere reference does not justify use of either code: teachers must actually use a second language or describe a language difference with English
T prompts students to use background knowledge of another culture or language	3. T prompts students to use a language other than English 4. T prompts students to share their knowledge or experience of another culture where English is not the dominant language	Highlighting a language or culture is just one way to uplift the assets carried by MLs. Translanguaging pedagogy (García et al. 2017) is another approach. That is, teachers can *explicitly* invite ML students' *active deployment* of their linguistic repertoires (their linguistic funds of knowledge) as they make meaning of the lesson
T uses techniques to increase clarity	5. T mimes or gestures in a purposeful way 6. T speaks slowly and clearly 7. T repeats or rephrases key information	Clear teaching is arguably a foundation of instruction. In the case of MLs, relative lack of familiarity with English makes it more important for students and simultaneously more difficult for teachers. Using nonverbal cues and adjusting rate of speech, enunciation, and language complexity, teachers can make instruction more comprehensible to MLs (Echevarría et al. 2023)
T uses other tangible props/aids	8. T uses a graphic organizer 9. T gives instructions in both written and oral form 10. T explains text features to increase understanding 11. T highlights key words or phrases in written or oral form during the task	In addition to clarity, a teacher can scaffold comprehension for MLs by adding more tangible and visual support (TESOL International Association 2018), such as graphic organizers and written directions

(Continues)

Observable scaffold	Item	Justification
T uses literacy techniques like word recognition or decoding strategies	12. T chunks a word into smaller pieces to sound out individual phonemes or helps students blend phonemes 13. T helps students use morphemes to understand the meaning of a word	In the early stages of literacy development, all students, including MLs of course, need to learn to decode words with the goal of achieving automatic word recognition, the fundamental prerequisite for reading comprehension (Goldenberg 2020). Two ways of guiding students in this process involve (a) raising phonemic awareness (i.e., developing awareness of the segments of sounds in speech) by breaking a word into phonemes (i.e., individual sound units characteristic of the language) and (b) helping students analyze word parts (i.e., morphemes), especially in multisyllabic words (Foorman et al. 2016)
T preteaches or highlights key vocabulary and/or key concepts needed to understand a lesson	14. T gives explicit vocabulary instruction for an upcoming task with opportunity to practice. 15. T identifies academic word as a cognate (or warns against false cognates)	Vocabulary knowledge is one of the strongest predictors of reading comprehension and academic success (Grabe and Yamashita 2022; Masrai and Milton 2018). MLs, however, face higher cognitive-load demands when confronted with new material because they encounter unknown words while simultaneously learning subject content (Ajayi 2014). Therefore, teachers have the opportunity to facilitate MLs' content learning by covering or reviewing vocabulary words in advance of a task and by using MLs' knowledge of their own home language (Richards-Tutor et al. 2016). In particular, teachers may call attention to cognates to facilitate understanding or avoid misunderstanding in the case of false cognates
T uses a variety of techniques to elicit language production	16. T prompts students to elaborate on their answer 17. T uses questions that elicit language production	Although MLs' social and cultural needs have an effect on their participation (or lack thereof) in class (Yoon 2007), using language to demonstrate their understanding and to actively engage with the classroom task is critical for students who are developing language and literacy skills (Gass 2015). Teachers should make sure MLs progress towards fluency in English and make their classroom talk accountable, which also requires MLs to listen and respond constructively to others (TESOL International Association 2017). This can be done by posing open-ended questions or by encouraging students to elaborate on comments that have already been made (e.g., *Can you say more on that? Why do you think that?*)
T recasts students' responses using academic language	18. T recasts student's response or question using academic language	In some cases, the teacher may recognize that a student understands a concept but lacks the English relevant to the concept. The teacher might also modify the student's utterance using the appropriate English, but without altering the student's meaning. This type of feedback is a recast: "The teacher's reformulation of all or part of a student's utterance, minus the error" (Lyster and Ranta 1997, 46), which leads to substantial learner uptake (Goo and Mackey 2013)
T describes a metacognitive strategy that uses understanding of language structure to complete a task	19. T describes a language-based metacognitive strategy that students can use to complete a task	Metacognitive strategies help students monitor their own comprehension. This internal activity—an adjunct to engagement—supports their engagement with increasingly more challenging reading or academic tasks (Grabe and Yamashita 2022). Examples of specific strategies include noticing a signal word, adjusting a prediction, and reflection on a reading goal (Pressley 2002). In the mode of observable behavior, explicit explanation of specific metacognitive strategies builds MLs' learning capacities
T prompts students to use a previously discussed metacognitive strategy related to language structure	20. T prompts s to use previously discussed metacognitive strategy related to language	Teachers may have described language-based metacognitive strategies, but active use does not come naturally to most students (Pressley 2002). Recognizing this circumstance, teachers remind students to use them for a specific purpose (e.g., notice the structure of the text by looking at the title and headings)
T provides a sample or model response	21. T provides sentence stems or frames 22. T provides an exemplar of the assignment 23. T models language production for a task	Samples and models are a form of scaffolding that gives MLs access to content and language concepts that would otherwise be more difficult to grasp. For instance, teachers can supply partially completed sentences in an assignment or a fully completed assignment (a sample product to use as a model). Samples and models also illustrate a clear picture of what student need to accomplish (Echevarría et al. 2023)

Appendix C

Frequency and Reliability Statistics for Year One Videos ($n = 43$).

Item text	% yes	% Agmt	κ
1. T uses a non-English language or mentions a difference between another language and English	31%	98%	0.95
2. T uses an example from a culture where English is not the dominant language (aside from an example based solely on the language)	4%	93%	−0.04
3. T prompts (or allows) S to use a language other than English	17%	98%	0.92
4. T prompts S to share their knowledge or experience of another culture where English is not the dominant language	1%	98%	−0.01
5. T mimes or gestures in a purposeful way	92%	93%	0.53
6. T speaks slowly and clearly	78%	84%	0.53
7. T repeats key information	93%	100%	1.00
8. T uses a graphic organizer	49%	72%	0.44
9. T gives instructions in both written and oral form	51%	86%	0.72
10. T explains text features to increase understanding	12%	95%	0.77
11. T highlights key words or phrases in written or oral form during the task	44%	77%	0.53
12. T chunks a word into smaller pieces to sound out individual phonemes or helps students blend phonemes	22%	93%	0.80
13. T helps students use morphemes to understand the meaning of a word	9%	95%	0.72
14. T gives explicit vocabulary instruction for an upcoming task with opportunity to practice	34%	88%	0.74
15. T identifies academic word as a cognate (or warns against false cognates)	6%	98%	0.79
16. T prompts S to elaborate on their answer	51%	81%	0.63
17. T uses questions that elicit language production	73%	88%	0.70
18. T recasts S's response or question using academic language	14%	91%	0.61
19. T describes a language-based metacognitive strategy that students can use to complete a task	15%	88%	0.55
20. T prompts s to use previously discussed metacognitive strategy related to language	13%	88%	0.48
21. T provides sentence stems or frames	13%	98%	0.90
22. T provides an exemplar of the assignment	2%	100%	1.00
23. T models language production for a task	45%	70%	0.39