Global Business Languages

Volume 23 Article 2

Development and Effectiveness of a Rater Training Curriculum for Evaluating Student Medical Spanish Oral Proficiency Using the Physician Oral Language Observation Matrix

Pilar Ortega
University of Illinois College of Medicine

Reniell X. Iñiguez
University of Illinois College of Medicine

Steven E. Gregorich
University of California, San Francisco

Cristina Pérez-Cordón
United Nations

José Alberto Figueroa
Northwestern University Feinberg School of Medicine

Karen Izquierdo
College of Medicine, State University of New York at Downstate Health Sciences University

Javier González
Memorial Sloan Kettering Cancer Center

Leah Karliner
University of California, San Francisco

Lisa C. Diamond
Memorial Sloan Kettering Cancer Center

Recommended Citation

Ortega, P. et al. (2023). Development and Effectiveness of a Rater Training Curriculum for Evaluating Student Medical Spanish Oral Proficiency Using the Physician Oral Language Observation Matrix. *Global Business Languages*, 23, 14-39.
Available at (DOI): http://doi.org/ 10.4079/gbl.v23.3

Pilar Ortega
Accreditation Council for Graduate Medical Education
University of Illinois College of Medicine

Reniell X. Iñiguez
University of Illinois College of Medicine

Steven E. Gregorich
University of California, San Francisco

Cristina Pérez-Cordón
United Nations

José Alberto Figueroa
Northwestern University Feinberg School of Medicine

Karen Izquierdo
College of Medicine, State University of New York at Downstate Health Sciences University

Javier González
Memorial Sloan Kettering Cancer Center

Leah Karliner
University of California, San Francisco

Lisa C. Diamond
Memorial Sloan Kettering Cancer Center

Development and Effectiveness of a Rater Training Curriculum for Evaluating Student Medical Spanish Oral Proficiency Using the Physician Oral Language Observation Matrix

Abstract: To improve language-concordant, equitable healthcare for Spanish speakers, many United States medical schools offer medical Spanish education. However, there is no clinically contextualized, standardized approach to medical student language assessment. This article reports on the development and effectiveness of a training curriculum to prepare raters to use a new rubric, the Physician Oral Language Observation Matrix (POLOM)™, to reliably rate medical student Spanish oral proficiency after observing videorecorded encounters between students and standardized patients. Curriculum effectiveness was primarily evaluated by examining each rater trainee's level of agreement with expert consensus POLOM ratings (i.e., inter-rater reliability as measured by the generalizability theory dependability coefficient). Out of the study's nine rater trainees, who were from either medical or linguistic professional backgrounds, five proceeded to the calibration phase, and four attained the reliability threshold required for calibration. The paper concludes that this rater training curriculum can successfully train raters to use the POLOM reliably when evaluating medical student Spanish oral proficiency during videorecorded healthcare encounters. To allow for improved assessment of student

language skills prior to use in patient care, future efforts should focus on POLOM validity assessment and larger scale rater recruitment, training, calibration, and maintenance.

Keywords: doctor-patient communication, inter-rater reliability, language assessment, language proficiency, medical education, medical Spanish, rater training

Introduction

Approximately one in five people in the United States speaks a language other than English at home (Ryan, 2013) and Spanish is by far the most common non-English language spoken in the United States (Zong & Batalova, 2015). In a healthcare context, the large and growing numbers of US Spanish speakers with limited English proficiency (US Civil Rights Division, 2015) means that a significant subset of the population will have difficulty communicating about their health in English. Individuals who are unable to communicate directly with their clinicians in their preferred language have worse health outcomes compared to individuals who receive language-concordant care (Diamond et al., 2019).

To improve language-concordant healthcare, and thus, health equity for US Spanish speakers, many medical schools offer medical Spanish educational programs. In the context of training physicians, medical Spanish courses aim to teach "the use of Spanish in the practice of medicine for communication with patients" (Ortega et al., 2020a). All established core competencies and corresponding performance objectives for medical Spanish learners focus on oral communication skills (speaking and listening) during medical encounters (Ortega et al., 2020a). A recent national survey showed that 78% of US medical schools offer options for medical students to enhance their Spanish skills (Ortega et al., 2021a). However, 43% of those schools lack of a standard process for assessing medical Spanish proficiency and, therefore, forgo learner assessment altogether (Ortega et al., 2021a). Without formal guidance, students and physicians must typically decide for themselves whether and when to rely on their Spanish skills in patient care. Even among those schools that do incorporate an assessment, a wide variety of non-harmonized strategies are reported, including written examinations, oral interviews, commercially available exams, and objective structured clinical examinations with standardized patients (SPs: trained actors who learn to play the role of a particular patient in simulated medical encounters). The lack of a standard approach to medical Spanish proficiency assessment means that, to evaluate and provide students with feedback, educators often must create their own tools/rubrics and make their own determinations as to when students are ready to independently perform their clinical duties in Spanish.

A second, related challenge to assessment is that many medical Spanish educators feel ill-equipped to rate student medical Spanish-language skills. Faculty who teach medical Spanish in US medical schools vary in their prior training and qualifications; the majority are either physicians (78%) or language professors (Ortega et al., 2021a). Prior literature has highlighted the interdisciplinary nature of the field of medical Spanish, calling attention to the need to train educators in areas of medical Spanish to which they may not have previously been formally exposed (Hardin & Hardin, 2013; Ortega et al., 2021b; Ortega et al., 2021d). Physician educators may benefit from additional training regarding linguistic elements, and language educators may need supplemental preparation regarding clinical communication components.

Thus, to resolve the current gap in medical Spanish proficiency assessment, two issues must be addressed: 1) the development of a reliable and valid standardized approach for

assessing medical Spanish proficiency; and 2) the training of individuals who can reliably apply such a system to rate students' medical Spanish skills prior to independent patient care. To address the first issue, our research team developed a tool for medical Spanish proficiency assessment, the Physician Oral Language Observation Matrix (POLOM)™, the development of which has been described in detail elsewhere (Diamond et al., 2023). The POLOM defines six categories on which the student's medical Spanish skills are rated with five ordered levels of proficiency based on their performance during student-SP videorecorded encounters. The six categories are comprehension, fluency/fluidity, vocabulary, pronunciation, grammar, and communication. While a detailed description of the rating instrument has been described by Diamond et al. (2023), Table 1 provides a descriptive summary of the POLOM's six rating categories. In brief, our prior research demonstrated excellent reliability among four of the investigators who became expert raters using the POLOM. They derived their expertise via an iterative process of POLOM refinement using the POLOM to independently rate student-SP encounters and then meeting as a group to compare and discuss their scores and revise the tool. Before examining the validity of POLOM ratings, we turned our attention to the training of new raters in the reliable use of the POLOM.

Table 1

The Physician Oral Language Observation Matrix™ Rating Categories and Their Definitions

Category	**Definition**
Comprehension	The candidate's ability to understand the patient's speech, including the understanding of sounds, words, and phrases.
Fluency/Fluidity	The candidate's ability to make their speech in the tested language flow smoothly and with ease, without excessive pauses, stammering, or hesitation.
Vocabulary	The candidate's ability to use words/lexical units (including idioms or metaphors) appropriately to ask questions and/or provide explanations during the encounter.
Pronunciation	The candidate's production of words, which consists of: vocalization or articulation of sounds, and accentuation, rhythm, and intonation.
Grammar	The candidate's use of the rules and principles that determine the way in which words are combined to form and connect meaningful sentences (e.g., sentence construction, word order, verb conjugations, connectors).
Communication	The candidate's ability to successfully fulfill the task (e.g., conduct a patient interview), integrating language and social skills (e.g., rapport-building, appropriately adjusting register [such as explaining medical jargon and using appropriate formality in addressing the patient], respectfully addressing cultural or sensitive issues) in a correct and appropriate way.

Note. Categories and definitions are from Diamond et al., 2023.

Rater Training

A rater is defined as someone who uses a scoring rubric to measure a candidate's performance. Rater reliability is the extent to which two or more raters agree on each other's

scoring of the same candidate (Karuppaiah et al., 2020). One challenge to reliability is when raters differ in how they apply the rating rubric to candidate behaviors (e.g., whether a given student is rated as a level 3 or a level 4). In clinical assessment of medical learners, factors such as raters' emotions and their overall impression of a learner can also lead to discordant scores among raters, particularly when scoring criteria leave room for interpretation (Christensen et al., 2018; Gingerich et al., 2017; Gomez-Garibello & Young, 2018). For example, in a qualitative study of medical education raters, Christensen et al. (2018) found that raters spontaneously applied their *taste*, defined as characteristics of learners that raters were idiosyncratically drawn to, such as reflectivity, resilience, empathy, and alikeness. Rater training can improve rater accuracy, inter-rater reliability, and efficiency, and reduce rating bias due to taste (Davis, 2016; Kobak et al., 2005; Kogan et al., 2015).

Inaccurate rating of physician language skills is problematic because proficiency test results are intended to be used to determine whether a physician is allowed to use a specific language for direct patient care without a medical interpreter (Diamond et al., 2014). Mistakes in the determination of a physician or medical student's readiness for patient care in a specific language have implications for patient safety (e.g., if a physician or medical student makes communication errors due to limited language skills), utilization of limited hospital resources (e.g., language services use when they are actually not needed), patient satisfaction, and physician/medical student satisfaction.

Although rater training systems for language proficiency exist, they are not specific to healthcare. For example, the American Council on the Teaching of Foreign Languages (ACTFL) offers training and certification for raters using their oral proficiency interview (OPI). ACTFL OPI raters are required to have superior proficiency in the target language, an undergraduate degree in a related field, and be affiliated with an academic institution (ACTFL Language Connects, 2022). However, since the ACTFL OPI is not specific to healthcare, it may result in inaccurate characterization (under or over-rating) of individuals' communication skills in medical contexts. For example, a person may not pass a general Spanish advanced or higher level OPI, yet, in focused medical contexts and when properly trained for this purpose, be adept at clearly communicating the key information needed for patient care within their clinical specialty. Conversely, a person who grew up bilingual in Spanish and English but has no training in Spanish clinical skills (e.g., medical terminology, simplification of complex health concepts, etc.) could score well on a general Spanish OPI but may not be ready to apply Spanish-language skills to their professional medical responsibilities. The finding that general language proficiency testing is inadequate for determining language skills in a professional medical setting is supported in the language for specific purposes literature (O'Sullivan, 2012) as well as the medical literature (Friedman et al., 1991). Relatedly, if the goal of student language assessment in medical Spanish is limited to performance of their patient care duties, then raters should be trained in evaluating skills in that context specifically. In a report proposing language standards for healthcare practice in Canada, Watt and colleagues (2003) highlighted the importance of *situational authenticity* in all aspects of the medical language assessment, including the task the candidate has to complete, the interaction, as well as the rater and rating tool.

Training raters to evaluate English communication skills has been studied in US health professions programs (Kobak et al., 2005), medical schools (Yudkowsky et al., 2019), and residencies (Gardner et al., 2016). In these contexts, raters are typically trained to rate students' English communication skills while observing their performances during an SP encounter. SP encounters are a long-standing modality for assessment of communication skills in students

training to be physicians (Barrows, 1993). SP encounters most commonly occur in settings that resemble a clinic or hospital room housed within medical school simulation centers; the setting and scenario are intended to replicate an authentic patient-clinician encounter and are routinely incorporated throughout US medical education to teach and assess a variety of clinical skills in English, including history-taking, delivering serious news, counseling, physical examination, and procedural skills. The focus of evaluation is on the communication skills performed, which necessitate but go beyond English-language proficiency alone; other aspects of communication are also evaluated such as the learner's clinical reasoning (e.g., their ability to use the information gathered during the patient encounter for accurate medical decision-making). This is in contrast with the goals of rater training in the current study, in which the focus is limited to evaluating oral Spanish-language proficiency in the context of a medical encounter.

Some prior exams have evaluated language skills in clinical contexts to determine whether aspiring clinicians, such as international medical graduates, are competent in the national or *dominant* language. For example, in the former US Medical Licensing Examination Step 2 Clinical Skills, students' English-language skills were formally evaluated during SP encounters as a component required for passing. In the United Kingdom, New Zealand, and Australia, the Occupational English Test is one of the accepted examinations for overseas/immigrant clinicians to demonstrate English-language proficiency as part of the licensing process (Pill & Woodward-Kron, 2012). However, no work to date has addressed rater training for assessing proficiency in *non-dominant* languages common in local or regional populations, such as Spanish in the United States.

Some US medical schools that use SP encounters for medical Spanish assessment have reported using their own rating rubrics and a combination of faculty and SP ratings to provide learners with performance feedback (Morales et al., 2015; Ortega et al., 2017), but few have used validated tools or reported on the reliability of the rubrics they developed. While a rating rubric for trained SPs to evaluate medical student interpersonal skills has been adapted into Spanish and validated (Ortega et al., 2021c), this tool is not meant to directly rate students' medical language proficiency but rather their overall interpersonal communication skills as perceived by the SP. Other medical Spanish programs have circumvented the need to create a rating rubric or train raters by outsourcing assessment through the Clinical Cultural Linguistic Assessment (CCLA), a validated commercial phone-based examination in which candidates record responses to pre-recorded prompts (Tang et al., 2011). Although the CCLA is clinically contextualized, it has been critiqued both for failing to simulate an authentic patient-clinician dialogue and for its focus on medical content specific to certain specialties. Moreover, the CCLA is not designed for incremental formative feedback, making it challenging to align faculty's educational objectives and learner feedback with the CCLA as a summative assessment tool at the end of the course. Detailed information about how the CCLA is scored is not publicly available, and score reports do not describe how candidate behaviors (e.g., specific language-related attributes or mistakes) result in a given rating. If trained raters could provide reliable and valid ratings of students' medical Spanish proficiency using the POLOM, then the aforementioned challenges could be addressed.

Objective

In this article, we report on the development and effectiveness of a POLOM rater training curriculum. We first present the process of curriculum development and identification of trainee

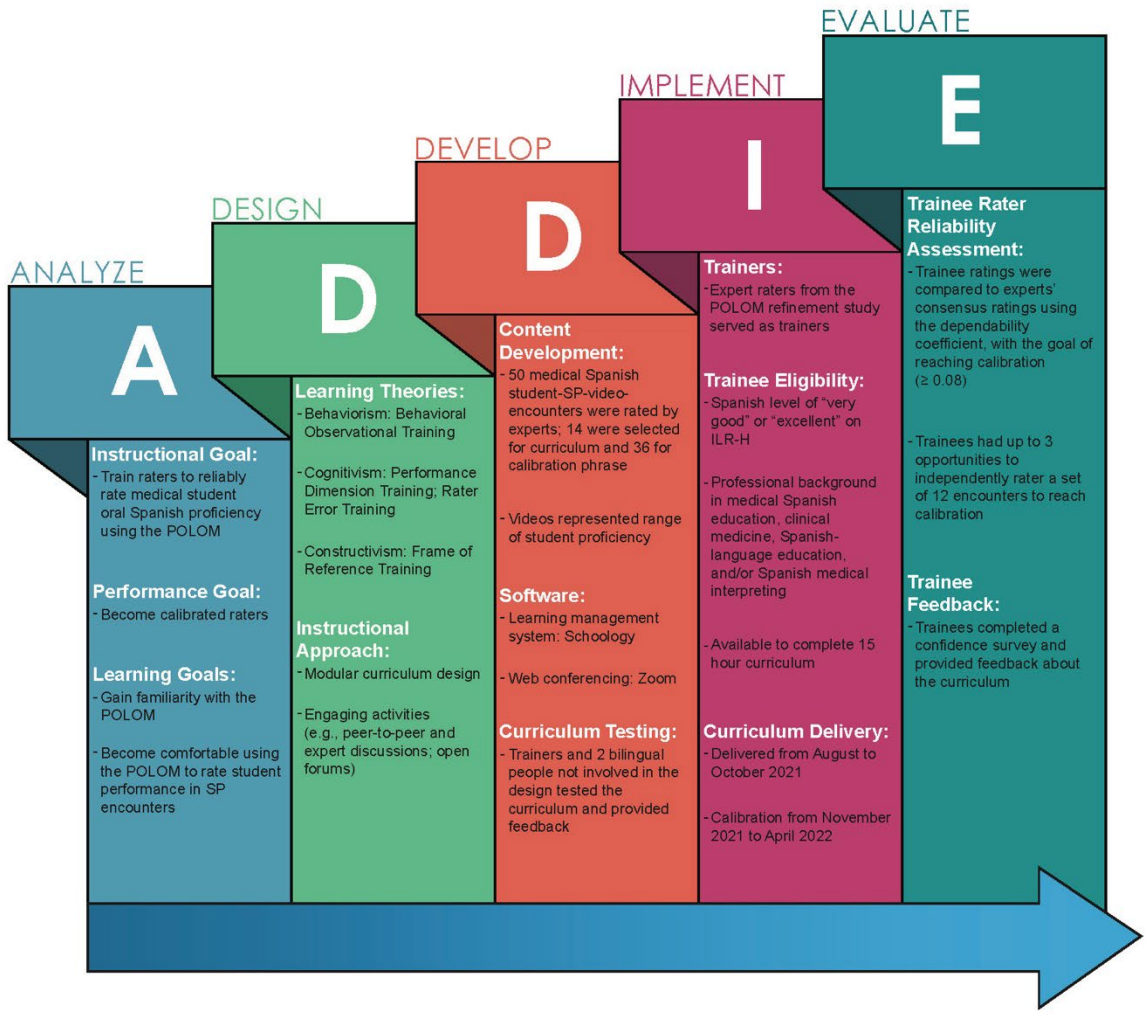
raters. Secondly, we report on curricular outcomes including trainee feedback, self-reported confidence levels, and trainee rater reliability.

Methods

To create the POLOM rater curriculum, we followed the Analysis, Design, Development, Implementation, and Evaluation (ADDIE) instructional design model (Bates, 2015). This model has been successfully used in both medical (Cheung, 2016) and language instruction (Zhang, 2020). The ADDIE acronym represents each of five stages of design and can be flexibly adapted (McIver et al., 2015). In what follows, we describe each stage of curriculum development for training POLOM raters, summarized in Figure 1 on following page (adapted from DeBell, 2020).

Figure 1

Summary of the Sequential Analysis, Design, Development, Implementation, and Evaluation (ADDIE) Instructional Design Model as Applied to the POLOM™ Rater Training Curriculum



Note. ILR-H=Interagency Language Roundtable healthcare scale; SP=Standardized patient; POLOM=Physician Oral Language Observation Matrix

ADDIE: Analysis Stage

We first established that the main instructional goal was to teach trainee raters to use the POLOM to reliably rate medical students' medical oral Spanish proficiency during videorecorded SP encounters. Next, we established that the performance goal was for trainees to become *calibrated* raters. We defined "calibration" as the ability to reliably rate students' ability to use Spanish in patient care as measured against expert rater consensus scores (determined by the aforementioned team of four expert raters) for the same SP encounters. Next, we identified the trainee learning goals required to achieve the performance objective (listed in Figure 1).

The learning environment was established as fully online using a web-based Learning Management System (LMS) software with asynchronous, self-paced activities, some autonomous and some interactive, as well as a synchronous live face-to-face component in the form of videoconferences with at least two expert raters who also served as trainers. Finally, we defined the trainee target population as language or medical professionals with some medical Spanish teaching experience. The trainee target population was defined based on the available literature describing the professional backgrounds of faculty who teach medical Spanish in US medical schools; the majority are physicians, some are language professionals with university teaching experience in Spanish for healthcare purposes, some are non-physician clinicians, and some are medical interpreters or translators (Morales et al., 2015; Ortega et al., 2021a). Notably, a significant subset of medical school Spanish courses lack a methodology for learner assessment (Ortega et al., 2021a). Thus, although it would have been ideal to select trainee raters with prior language assessment experience, such a requirement would not have realistically allowed for recruitment from among current medical school Spanish course faculty, yet these faculty members may greatly benefit from rater training to enhance their teaching roles.

ADDIE: Design Stage

We incorporated the learning theories of behaviorism, cognitivism, and constructivism, as well as previously published strategies for training raters to evaluate clinical communication skills. Behaviorist approaches integrated in our curriculum design included incorporating opportunities for multiple attempts to learn a new idea, breaking down complex ideas into concepts, and using positive reinforcement (Skinner, 1976); similar approaches have long been used in language acquisition (Aljumah, 2020; Broad, 2020). Cognitivism has also been used as a framework for language learning (Aljumah, 2020; Suharno, 2010). In our course, cognitivism-based strategies included progressing from general to more specific concepts, integrating new material from previously learned concepts, and using meaningful contexts that promote discovery and self-motivation (Piaget, 1968). Finally, we incorporated constructivist strategies to support building knowledge within social contexts by combining interactive content, trainee independent work, and supportive interactions with peers and instructors (Vygotsky, 1978). The constructivist approach has been previously applied to language teaching (Mvududu & Thiel-Burgess, 2012). Table 2 summarizes the theoretical principles and how we applied them in the POLOM rater training curriculum design.

Table 2

Summary of Rater Training Strategies for Evaluating Clinical Skills Performance and Their Application to the POLOM™ Rater Training Curriculum

Theoretical Principle	Rater Training Strategy	Description and Examples	Application to POLOM Rater Training Curriculum
Behaviorism	**Behavioral Observation Training**	Designed to condition raters to identify, store, and recall key behaviors during rating scenarios that are associated with an evaluated skill. Helps with rater recall and documentation of critical events that they can digest later when completing an evaluation. Examples: Ludbrook & Marshall, 1971; Rosen et al., 2008	Module 1: Quiz with key scenarios in which trainees had to choose an appropriate rating for a particular student behavior. Modules 5, 6, 7: In all rating practice activities, trainees completed a scoring sheet with designated spaces in which they recorded key student behaviors corresponding to each POLOM category. They then reviewed their recorded notes after watching the entire encounter and prior to assigning a score.
Cognitivism	**Performance Dimension Training**	Designed to increase trainee knowledge of the dimensions being targeted for evaluation by gradually scaffolding knowledge and skills as well as building concepts from general to specific. Examples: Evans et al., 2009; Feldman et al., 2012	Module 3, 4: After introducing the POLOM, each subsequent module gradually added to a more nuanced understanding of how to apply knowledge of the instrument in general to specific scenarios. Video encounters were given as examples and reviewed during live meetings.
	Rater Error Training	Designed to familiarize trainees with common rating errors. Based on these, they would be able to identify similar errors and avoid them. Examples: Fahim, 2011 ; Feldman et al., 2012; Iramaneerat & Yudkowsky, 2007	Module 2: Unconscious bias training Module 3: Trainees were provided with a document detailing tips and strategies on how to avoid future errors similar to the examples provided. Module 4: Training videos included explanations of why the student was given a specific rating in each scoring category versus another.

Constructivism	**Frame-of-Reference Training**	Designed to help trainees understand which student behaviors constitute specific levels of performance that correspond to a scoring category. Using this training approach, trainees have the opportunity to practice rating, discuss discrepancies among themselves and with trainers, and receive feedback. This allows trainees to collaboratively anchor to a common performance framework agreed upon by trainers that serves as a frame of reference when analyzing encounters. Example: Feldman et al., 2012	Module 2: Unconscious bias interactive reflection as relevant to specific POLOM categories (e.g., how a student's accent may intersect with how a rater scores their pronunciation) Modules 5, 6, 7: Prior to the calibration phase, trainees independently scored several encounters. Their scores were compared to expert consensus scores for that specific encounter. They discussed discrepancies in a group setting with peer and expert raters.

We sequenced the learning objectives and designed the instructional materials, delivery method, and formative and summative evaluation activities throughout the curriculum (Table 3). We chose a highly interactive instructional approach, including narrated sample SP encounters in which trainers walk trainees through the process in rating students who are at different proficiency levels, opportunities for peer-to-peer reviews, and several forums where trainees could ask questions and interact with other trainees.

Table 3

Structure and Description of the POLOM™ Rater Training Curriculum

Module Title	Module Description	Module Activities	Estimated Completion Time
0. Course Presentation and Personal Introductions	Introduction to trainers, peer trainees, and course objectives	Trainers and trainees introduce themselves, sharing their professional backgrounds. Trainees review the course objectives.	30 minutes
1. The POLOM	Introduction to the POLOM rating tool	Understanding the POLOM: Trainees review key documents explaining the POLOM purpose and rubric. Trainees complete multiple choice quiz with feedback.	2 hours

2. Addressing Unconscious Bias	Overview of unconscious bias and discussion of how it may influence ratings of student medical language proficiency	Trainees watch two videos explaining unconscious bias Trainees read an article about Spanglish followed with open forum. Reflection activity: Trainees write about how unconscious bias may relate to rating a students' medical language proficiency.	2 hours
3. Tips and Strategies	Self-paced review of expert scoring strategies	Trainees review a compilation document outlining several expert rater-based scoring strategies.	20–30 minutes
4. POLOM Rating: Observation	Narrated exemplar SP encounter videos displaying critical rating moments	Trainees watch a series of four narrated videos each displaying an encounter that corresponds to a given POLOM rating level. Trainees review the consensus rating written rationale for each video.	2 hours
5. Using the POLOM: Practice	SP encounter videos for trainees to practice and discuss ratings	Trainees rate one encounter and then meet with an assigned peer to discuss their scores. Trainees rate two additional encounters and meet as a group with expert raters and peers to discuss.* Trainees review the consensus rating written rationale for each video.	3 hours
6. Using the POLOM: Implementation	SP encounter videos for trainees to practice rating and receive expert feedback	Trainees sequentially rate three encounters, each followed by a group session with expert feedback.** Trainees review the expert consensus rating written rationale for each video.	3–4 hours
7. Independent POLOM Rater	SP encounter videos for trainees to rate independently	Trainees rate two encounters and submit ratings to trainers without discussing with peers/trainers or receiving feedback between ratings. If trainee moves on to	1 hour

		calibration phase, these two ratings are included as two of the 12 in the trainee's first reliability rating round.	

Note. *Attendance was required at one of two meetings and meeting recordings were made available to trainees who were absent; **Attendance was required at two of three meetings and recordings were made available to trainees who were absent.

ADDIE: Development Stage

We developed the content and necessary materials according to decisions made during the design stage. Some of the content development took place in the context of our prior study to refine the POLOM and evaluate expert rater reliability (Diamond et al., 2022). In that study, four expert raters attained excellent reliability (dependability coefficient, $\widehat{\Phi}$, of 0.926, defined below) with a sample of 50 videorecorded medical student-SP encounters. The 50 encounters were drawn from a repository of 356 video-recordings from a medical Spanish course for third- and fourth-year students at the University of Illinois College of Medicine. The videos included students with self-rated Spanish proficiencies ranging from "fair" to "excellent" on the Interagency Language Roundtable healthcare scale (ILR-H), a version of the ILR scale modified for clinician self-assessment. The ILR-H includes five proficiency level options ("poor," "fair," "good," "very good," and "excellent") and their descriptors in a healthcare context (Diamond et al., 2012). Of note, students were required to meet the ILR-H level of "fair" or higher to enroll in the medical Spanish course. The "fair" ILR-H level is approximately equivalent to low-intermediate level on ACTFL's proficiency scale. Following inter-rater reliability assessment, expert raters reconciled any differences in their ratings of each POLOM category within each encounter and developed consensus ratings. Additionally, trainers created a textual summary of each encounter that described the rationale for each consensus rating.

We determined that the training program would be imparted in two phases: a curriculum phase and a calibration phase. A purposive sample of 14 of the 50 videorecorded encounters as well as the corresponding consensus ratings and rationale were included in the curriculum phase to provide concrete examples of key instructional points; trainers intentionally selected encounters illustrating students at varying levels of proficiency that highlighted common rating challenges. The remaining 36 encounters were reserved for the calibration (reliability assessment) of trainees who completed the curriculum phase.

We selected accessible software that met our design criteria; we used Schoology as our LMS and Zoom as our teleconferencing platform for live meetings. We tested functionality by having trainers log in to the platform, test the interactive components, and provide feedback for improvement. Two bilingual study team members who were not involved in the design then tested the curriculum to identify any errors or potential challenges prior to implementation. Suggested modifications included clarifying the wording of some quiz questions and improving the placement of some resources for greater visibility/availability to trainees. The curriculum was then ready to be launched.

ADDIE: Implementation Stage

We delivered the POLOM rater training curriculum from August to October 2021 with the first cohort of trainees. The calibration phase took place from November 2021 to April 2022.

Context

Our rater training curriculum is situated in the context of a growing national effort to standardize medical Spanish courses in US medical schools (Ortega et al., 2020a). One of the outcomes of this recent effort has been the formation of the National Association of Medical Spanish (NAMS), an interdisciplinary non-profit organization with over 250 language and healthcare professional members. This national network facilitated access to potential trainee raters (e.g., medical Spanish faculty) from multiple institutions and materials—specifically, SP encounters—from one participating medical school.

Trainers

The four expert raters from our prior study agreed to serve as trainers. Two are language professionals whose first language is Spanish and subsequently learned English; specifically, one is an interpreter educator and a healthcare language access specialist and the other is a language assessment specialist with a focus on Spanish as a second language. The other two trainers are physicians; one is a clinician researcher who learned Spanish as a second language, and one is a clinician and medical Spanish educator who grew up speaking both Spanish and English. All trainers self-reported an ILR-H level of "excellent" in Spanish and "very good" or "excellent" in English.

Trainee Rater Eligibility and Recruitment

To be selected as trainee raters, individuals had to 1) self-report a Spanish ILR-H level of "very good" or "excellent"; 2) have a professional background and/or experience in medical Spanish education, clinical medicine, Spanish-language education, and/or Spanish medical interpreting; and 3) indicate willingness and availability to complete the curriculum requirements (an estimated 15-hour commitment). To recruit potential trainees, we sent an invitational email explaining the project and eligibility criteria to individuals involved in medical Spanish education through NAMS. The ILR-H as a self-reporting tool has been validated as comparable to a proficiency exam when individuals self-rate at the highest and lowest ends of the scale (Diamond et al., 2014); nonetheless, given the potential inaccuracy of self-reported language abilities, we verified eligibility by means of a 15-minute videoconference call with two trainers for any individuals who expressed interest in the trainee rater role. All prospective trainees who met the criteria were invited to participate as a trainee rater on a voluntary basis. No compensation was offered for curriculum participation. Upon successful curriculum completion, trainee raters were eligible to participate in the calibration phase (described next), for which a modest hourly compensation was offered up to a maximum of eight rating hours.

ADDIE: Evaluation Stage

We evaluated trainee performance by calculating their level of agreement with expert consensus ratings, a form of reliability assessment described below. Furthermore, we collected feedback from trainees to identify areas of curriculum improvement. This study was determined to meet criteria for exempt research by the Institutional Review Board of the University of Illinois on November 24, 2020 (Protocol#2019-0945).

Data Collection

A voluntary online pre-survey collected information about trainees including demographics, professional background, as well as prior experience teaching and assessing medical Spanish. Following the curriculum, trainees were invited to voluntarily complete a post-survey to gather self-rated confidence in using the POLOM, obtain feedback about the curriculum, and determine their interest in participating in the calibration phase.

Once in the calibration phase, each trainee rated encounters in a completely independent manner, never discussed their ratings with any other trainee, and only discussed their ratings with trainers *after* completing each rating round (described below). Once completing a round of ratings, each trainee sent their ratings to a designated member of the study team who entered them into a database for analysis.

Trainee Rater Calibration

For inter-rater reliability assessment, the expert rater consensus ratings reflecting POLOM total scores (summing all six POLOM category ratings) of 36 encounters served as gold standard ratings, which were compared to the corresponding ratings of each trainee. Trainee reliability assessment included between one and three rounds of ratings. Each round included 12 of the 36 videorecorded SP encounters, with each set selected to represent a range of POLOM consensus scores.

During the first round of ratings, each trainee independently used the POLOM to rate 12 encounters. Next, generalizability (G) theory was used to estimate dependability coefficients ($\widehat{\Phi}$) comparing each trainee's POLOM total score ratings to the expert consensus ratings. Application of G theory proceeded in two steps (Brennan, 2001; Shavelson & Webb, 1991). In the first step, for each trainee and rating round, a G study estimated three variance components of POLOM total scores attributable to medical students (*s*), trainee raters I, and residuals (*d*). Because the goal of the analyses was to generalize to the populations of potential students and raters, the corresponding modeled effects were regarded as random. In the second step, a decision (D) study used the G study variance component estimates to estimate the dependability coefficient for each trainee and rating round.

Dependability is a type of reliability that reflects absolute agreement across raters; that is, high dependability required a trainee to provide POLOM total scores that were highly similar to the corresponding expert consensus scores. In contrast, some other reliability coefficients focus on relative agreement, only requiring raters to agree on the rank ordering of students with respect to their POLOM scores. Because the POLOM is intended to assess students' medical Spanish oral language proficiency in an absolute sense, we did not consider relative agreement coefficients. Dependability coefficients were estimated via Equation 1, where, e.g., $\hat{\sigma}_s^2$ represents

the variance component estimate for students. The reported dependability coefficients represent the proportion of total variation in POLOM scores attributable to between-student variation, and have a possible range of [0,1]. Some examples follow. When $\widehat{\Phi} = 1$, agreement is perfect; there is positive estimated between-student variation ($\hat{\sigma}_s^2$), but the between-rater ($\hat{\sigma}_r^2$) and residual ($\hat{\sigma}_d^2$) variance component estimates both equal zero, i.e., all variation in POLOM total scores is attributable to between-student differences. When $\widehat{\Phi} = 0$, there is no systematic agreement between raters; variation attributable to between-student differences equals zero and all variation in POLOM total scores is attributable to between-rater and/or residual variation.

Equation 1

$$\widehat{\Phi} = \frac{\hat{\sigma}_s^2}{\hat{\sigma}_s^2 + \hat{\sigma}_r^2 + \hat{\sigma}_d^2}$$

We defined the threshold for calibration as $\widehat{\Phi} \geq 0.80$ for the POLOM total score because it was deemed both a reasonably stringent and attainable initial training goal. Any trainee with $\widehat{\Phi} < 0.80$ for the POLOM total score received individualized feedback and rated a new round of 12 encounters, for up to three total reliability assessment rating rounds. Feedback to each trainee included a detailed report comparing their ratings to the consensus scores (e.g., tabular summaries and Bland-Altman plots [Altman & Bland, 1983; Bland & Altman, 1986]); trends in their ratings across rating rounds (if applicable), a written rationale for the consensus rating for each encounter, and a 1-hour videoconference with two trainers to review challenging cases and address questions.

Results

In sum, nine trainees participated in the rater training curriculum and five proceeded to the calibration phase. We now present the results of our trainee recruitment, trainee feedback about the curriculum, and reliability for the five trainees who proceeded to the calibration phase.

Trainee Rater Participants

Ten individuals expressed interest in training as medical Spanish oral proficiency raters and were scheduled for 15-minute interviews with two trainers. All 10 met eligibility criteria, were invited to enroll in the curriculum, and nine accepted (one person declined due to lack of time). All nine enrolled trainees completed the pre-survey, self-identified as Hispanic/Latinx, reported speaking primarily Spanish at home during their upbringing, completed their elementary, secondary, and higher education with Spanish as the dominant language of instruction, and self-reported their ILR-H Spanish level as “excellent” and their English level as “very good” or “excellent.” All trainees reported having at least one higher education degree, including seven trainees with doctoral degrees, one master’s, and one bachelor’s degree. Table 4 details the trainees’ professional backgrounds and previous training relevant to medical Spanish assessment.

Table 4

Trainee Rater Professional Background and Experience Relevant to Medical Spanish Assessment

Trainee ID	Professional Background	Professional Roles	Prior Training in Teaching Medical Spanish	Prior Training in Language Assessment	Prior Experience in Medical Spanish Assessment
A	Medical	Medical Spanish educator, Physician, Researcher	NAMS Train-the-Trainer pilot online course "Enseñar español médico"	None	None
B	Linguistic	Medical Spanish educator, Spanish language professor	Mentorship/coaching from a colleague	Coursework at Graduate School	SP encounters; Observed patient encounters
C	Linguistic	Medical translator, Spanish language educator	Mentorship/coaching from a colleague	Mentorship/coaching from a colleague	None
D	Linguistic	Medical Spanish educator, Spanish language professor	Mentorship/coaching from a colleague	Mentorship/coaching from a colleague	None
E	Medical	Medical Spanish educator, Physician	None	None	SP encounters; Observed medical encounters
F	Medical and linguistic	Medical Spanish educator, Medical interpreter, Medical translator, Physician (not practicing)	Mentorship/coaching from a colleague	CCLA rater training	SP encounters; Observed medical encounters
G	Medical	Medical Spanish	NAMS Train-the-Trainer pilot online	None	None

		educator, Clinical psychologist	course "Enseñar español médico"		
H	Linguistic	Medical Spanish educator, Medical interpreter, Spanish language professor	None	None	None
I	Linguistic	Medical Spanish educator, Medical interpreter, Medical translator, Spanish language professor	NAMS Train-the-Trainer pilot online course "Enseñar español médico"	ACTFL OPI rater training; MELAB rater training; Mentorship/coaching from a colleague	Observed medical encounters; Oral examinations of resident physicians

Note. ACTFL=American Council for the Teaching of Foreign Languages; CCLA=Clinician Cultural and Linguistic Assessment; MELAB=Michigan English language assessment battery; NAMS=National Association of Medical Spanish; OPI=Oral Proficiency Interview; SP=Standardized patient

Trainee Feedback and Self-Reported Confidence

Upon completing the curriculum, but prior to calibration, all 9 trainees responded to the post-survey. Trainees reported total time spent on the curriculum, ranging between 11-30 hours; this variation is to be expected because several elements were self-paced and two meetings were optional. The post-survey also asked trainees six questions to indicate their self-confidence in using the POLOM to independently rate a student's comprehension, fluidity/fluency, vocabulary, pronunciation, grammar, and communication (corresponding to each POLOM category) during an SP encounter. Self-confidence was measured using 4-point ordinal response options ranging from "strongly disagree" to "strongly agree." Eight of the 9 trainees either agreed or strongly agreed that they were confident in independently rating student performance on all six POLOM categories. One trainee expressed lack of confidence in rating the communication and comprehension categories; although all other trainees reported confidence in rating those categories, several indicated in free-text comments that those POLOM categories were the most challenging to understand as a rater.

We reviewed trainee rater feedback to understand the elements of the curriculum that worked best and those that could be improved for future trainings. In general, trainees valued the many opportunities to get to know both trainers and peers, starting with Module 0 and throughout the course. One trainee described that "knowing my classmates made me feel more

comfortable and helped create a trusting learning environment," and another said the discussion forum was helpful because "reading my classmates' posts helped me better understand their backgrounds and how we can complement each other." Trainees unanimously rated the peer-peer practice and expert-led group discussions (corresponding to Modules 5 and 6) as the most effective and enjoyable course components. One trainee wrote that these modules "informed me of the nuances and elements that are important to consider to rate fairly." Another person explained that the live meeting discussions about specific cases "strengthened our focus on the objectives [medical] students need to achieve while communicating with Spanish-speaking patients."

There were some elements of the peer-peer and group discussion sessions for which trainees suggested improvements. For instance, several trainees suggested to extend the duration and quantity of the live meetings to enable more discussion and answer questions about rating difficult cases. One person also suggested additional practice cases, explaining that "more practice over an extended period of preparation will certainly help me and possibly any rater to become more consistent and fluid at rating." Another trainee commented that although they found the unconscious bias training (Module 2) conceptually valuable, they would have liked to understand better how those concepts can affect rating the students' performance during medical encounters.

One trainee did not complete all required elements of the curriculum and was therefore excluded from the calibration phase. The eight trainees who completed the curriculum were invited to participate in calibration, and three declined to participate due to lack of time. Those who proceeded to calibration reported spending an average of 45 minutes rating each encounter, adding up to approximately nine rating hours per calibration round (each round consisting of 12 video encounters).

Trainee Rater Reliability

Five trainees (A through E in Table 4) proceeded to the calibration phase. POLOM total scores have a theoretical range from 6-30. In this context, total scores below 12 are unlikely because medical Spanish training programs select for students with a *minimum* Spanish ILR-H level of "fair" (approximately equivalent to low-intermediate level on ACTFL's proficiency scale). Table 5 summarizes the POLOM expert consensus ratings for the 36 encounters reserved for the calibration phase, stratified by rating round. Consensus total scores ranged from 14 to 29 or 30 in all rounds. Round 3 had a slightly higher total score consensus mean of 20.4 versus 18.8 and 19.0 for rounds 1 and 2, respectively.

Table 5

Summary of Expert Consensus Scores for Three Reliability Assessment Rating Rounds, Each with 12 Videorecorded Encounters

	POLOM™ Rating Category	Mean	Standard Deviation	Minimum	Maximum
Round 1	Comprehension	4.2	0.8	3	5
	Fluency/Fluidity	2.9	1.0	2	5
	Vocabulary	2.8	0.9	2	5
	Pronunciation	3.3	0.9	2	5

	Grammar	3.0	1.0	2	5
	Communication	2.6	1.0	2	5
	Total Score	18.8	4.8	14	30
Round 2	Comprehension	4.1	0.9	3	5
	Fluency/Fluidity	3.1	1.1	2	5
	Vocabulary	2.8	1.1	2	5
	Pronunciation	3.5	0.8	2	5
	Grammar	3.0	1.0	2	5
	Communication	2.6	1.0	2	5
	Total Score	19.0	5.0	14	30
Round 3	Comprehension	4.3	0.8	3	5
	Fluency/Fluidity	3.4	1.1	2	5
	Vocabulary	2.9	0.9	2	4
	Pronunciation	3.6	0.9	2	5
	Grammar	3.2	1.1	2	5
	Communication	3.0	1.3	2	5
	Total Score	20.4	5.6	14	29

Table 6 reports the POLOM total score dependability coefficients for each trainee at each completed round as well as the corresponding mean POLOM total score difference (trainee minus expert consensus). Four of the five trainees eventually attained the $\widehat{\Phi}\geq 0.80$ threshold required for calibration: trainees A and B in the first round, and trainees C and D in the third round. Upon passing the reliability threshold, the four trainees' average POLOM total scores tended to be somewhat higher than consensus scores, ranging from 0.92 to 1.83 points higher out of a total possible of 30 points.

Table 6

Dependability Coefficient Estimates and Average POLOM™ Total Score Difference at Each Reliability Assessment Rating Round for Five Trainee Raters

Trainee ID	Trainee Professional Background	Round 1		Round 2		Round 3	
		$\widehat{\Phi}$	$\bar{\Delta}$	$\widehat{\Phi}$	$\bar{\Delta}$	$\widehat{\Phi}$	$\bar{\Delta}$
A	Medical	**0.933**	**+0.92**	--	--	--	--
B	Linguistic	**0.840**	**+1.83**	--	--	--	--
C	Linguistic	0.460	+3.58	0.718	+2.67	**0.845**	**+1.58**
D	Linguistic	0.658	+3.75	0.762	+0.83	**0.808**	**+1.83**
E	Medical	0.331	+5.33	0.709	-1.33	0.651	+4.08

Note. $\widehat{\Phi}$, dependability coefficient estimate; $\bar{\Delta}$, average POLOM total score difference (trainee minus expert consensus); --, trainee previously met reliability threshold. Table entries for the rating round where the trainee met the reliability threshold are in boldface.

Discussion

This study demonstrates that it is possible to train raters to reliably use the POLOM when evaluating medical student Spanish oral proficiency during SP encounters. We created a curriculum grounded in communication and learning theory that was successful in calibrating the majority of trainees. Trainees particularly appreciated the ample opportunities to practice using the rating instrument to evaluate medical student performance using videorecorded medical encounters and to discuss their ratings with peers and trainers. The curriculum was highly interactive and accounted for varied trainee backgrounds and learning needs; if a trainee did not reach the reliability threshold after one round, we provided remediation by means of individualized feedback and an opportunity to complete up to three total scoring rounds. The POLOM rater training program was attentive to situational authenticity with regards to trainee raters; we intentionally recruited individuals who were already involved in medical Spanish teaching and thus motivated to enhance their skills in assessing students' medical language proficiency. Moreover, the training materials were authentic in representing medical students who would be seeking evaluation of their medical Spanish skills for patient care. Overall, the majority of trainees considered the required time commitment to be reasonable, despite the complexities of rating language proficiency in a medical context.

Nonetheless, identifying and preparing raters for this task is not without challenges. First, identifying individuals to train as raters is difficult. Medical Spanish educators represent a potential pool of raters who may further benefit from POLOM rater training to improve their teaching and the feedback they provide students throughout their courses. However, data show that medical schools primarily rely on physician faculty to teach medical Spanish (Ortega et al., 2021a), and these individuals may lack training in language teaching or assessment as well as have significant time constraints due to clinical responsibilities. Some medical schools have successfully incorporated non-faculty raters into their pool of trained raters for learner clinical skills evaluation (Yudkowsky et al., 2019). In Canada, health professionals in nursing, pharmacy, occupational therapy, and other fields serve as raters for medical licensing exams (Humphrey-Murto et al., 2005) and SPs have been successfully used as raters of English-language skills for international medical graduates (Rothman & Cusimano, 2001). In our study, we successfully recruited trainees from NAMS, whose members have skills in language and medicine. For future trainee recruitment, collaboration with this and other aligned professional organizations may be fruitful to establish a pool of trained POLOM raters.

Second, even once trainee raters are identified, it is important to acknowledge that rater training of physicians' language skills can have high stakes implications if the results will impact certification or licensure; following a validation study and standard-setting (currently in progress) for determining the score necessary to determine readiness for Spanish-language patient care, the POLOM has the potential for being used as part of a credentialing exam. Thus, very little rater error can be tolerated. In our POLOM rater training curriculum, we incorporated several approaches to enhance reliability, including rater error training and frame of reference training. Our program set the trainee rater reliability threshold for calibration at $\widehat{\Phi}\geq 0.80$, yet this level may not suffice for trainers to be qualified for high stakes evaluation. If applying the POLOM to determine student certification for patient care, a higher rater reliability threshold (e.g., $\widehat{\Phi}\geq 0.90$) may be deemed more appropriate for assuring patient care quality and safety. A higher level of reliability can be attained either through additional training or by averaging

ratings from multiple raters per student. Our data from expert raters and one trainee rater (who achieved $\widehat{\Phi}$>0.90) indicates that such a threshold is attainable.

Third, unconscious bias, a type of inconsistency rater error, must be carefully addressed when developing training programs for raters of medical oral language proficiency. Unconscious bias is increasingly recognized as a source of rating variance that is often structurally embedded in educational systems and disadvantages medical trainees who identify with racial or ethnic groups underrepresented in medicine (Klein et al., 2022). Certain linguistic attributes, such as a speaker's accent, the sound of their voice, or even their physical features, may cause the observer to attribute a higher or lower value to different linguistic varieties of the same language (Lippi-Green, 1997; Ortega et al., 2022). For example, candidates with lower proficiencies or "non-native" phonological patterns may be more susceptible to negatively biased ratings when assessed by an examiner who is a "native" speaker (Kang et al., 2019). In Spanish in particular, attributing higher value to some varieties of the language than to others (e.g., based on accent, sentence structure, terminology, or other language practices that may vary nationally, regionally, or locally) is a known source of potential bias (Ortega et al., 2020b; Ortega et al, 2022) that may influence ratings of oral proficiency. For example, some raters may have a tendency to rate heritage speakers (Martínez, 2010; Prada, 2019) and second language learners differently even when they are performing at the same proficiency level. In the design of our rater training curriculum, we addressed unconscious bias in general, and we also engaged trainee raters in reflective discussions about the diverse linguistic features of the students in the sample videos, emphasizing the importance of raters valuing all varieties of Spanish equally. Based on the trainee feedback, further enhancement of this section of the training may better support trainees to make explicit connections to their own rating patterns.

Limitations

Our sample of trainee raters was small, which limits the generalizability of our results to future groups of trainees. Also, our POLOM rater training curriculum may require modifications for different groups of trainees, depending on factors such as trainee professional background, number of trainees, trainer availability, and feedback from new trainee cohorts.

Future Directions

Future work should involve recruitment and preparation of additional trainers as well as curriculum refinements to enhance effectiveness of the program for a broader set of potential trainees, improve program efficiency, and increase rater reliability (agreement with expert ratings). Tracking rating metrics will be important for quality improvement to ensure that raters maintain their rating skills longitudinally. Additional research is needed to modify and evaluate the POLOM and the rater training curriculum for rating medical oral language proficiency in languages other than Spanish. Moreover, while our data show that our curriculum can effectively yield calibrated POLOM raters, another study is concurrently evaluating the validity of the POLOM for rating medical student Spanish oral proficiency during SP encounters.

If shown to be valid, the POLOM has the potential to serve as a practical rating tool to verify medical student qualifications prior to independent use of Spanish in patient care. Future efforts should also be directed toward establishing a sustainable and impactful system for training POLOM raters and maintaining calibration over time.

References

ACTFL Language Connects. (2022). *ACTFL tester and rater certifications*. https://www.actfl.org/assessment-research-and-development/tester-rater-certifications

Aljumah, F. H. (2020). Second language acquisition: A framework and historical background on its research. *English Language Teaching, 13*(8), 200–207.

Altman D. G., Bland J. M. (1983). Measurement in medicine: the analysis of method comparison studies. *The Statistician, 32*(3), 307–317.

Barrows, H. S. (1993). An overview of the uses of standardized patients for teaching and evaluating clinical skills. AAMC. *Academic Medicine: Journal of the Association of American Medical Colleges, 68*(6), 443–453. https://doi.org/10.1097/00001888-199306000-00002

Bates, A. W. (2015) The ADDIE model. In *Teaching in a digital age*. Tony Bates Associates LTD. https://opentextbc.ca/teachinginadigitalage/chapter/6-5-the-addie-model/

Bland J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet, 327*(8476), 307–310.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Broad, D. (2020). Literature review of theories of second language acquisition. *Journal of Applied Linguistics and Language Research, 7*(1), 80-86.

Cheung, L. (2016). Using the ADDIE model of instructional design to teach chest radiograph interpretation. *Journal of Biomedical Education,* 1–6.

Christensen, M. K., Lykkegaard, E., Lund, O., & O'Neill, L. D. (2018). Qualitative analysis of MMI raters' scorings of medical school candidates: A matter of taste? *Advances in Health Sciences Education: Theory and Practice, 23*(2), 289–310. https://doi.org/10.1007/s10459-017-9794-x

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, *33*(1), 117–135. https://doi.org/10.1177/0265532215582282

DeBell, A. (2020) ADDIE Model of Instructional Design. Water bear learning. https://waterbearlearning.com/addie-model-instructional-design/

Diamond, L., Chung, S., Ferguson, W., Gonzalez, J., Jacobs, E. A., & Gany, F. (2014). Relationship between self-assessed and tested non-English-language proficiency among primary care providers. *Medical Care, 52*(5), 435–438. https://doi.org/10.1097/MLR.0000000000000102

Diamond, L. C., Gregorich, S. E., Karliner, L., González, J., Pérez-Cordón, C., Iniguez, R., Figueroa, J. A., Izquierdo, K., & Ortega, P. (2022). Development of a tool to assess medical oral language proficiency. *Academic Medicine: Journal of the Association of American Medical Colleges,* 10-1097. https://doi.org/10.1097/ACM.0000000000004942

Diamond, L., Izquierdo, K., Canfield, D., Matsoukas, K., & Gany, F. (2019). A systematic review of the impact of patient-physician non-English language concordance on quality of care and outcomes. *Journal of General Internal Medicine, 34*(8), 1591–1606. https://doi.org/10.1007/s11606-019-04847-5

Diamond, L. C., Luft, H. S., Chung, S., & Jacobs, E. A. (2012). "Does this doctor speak my language?" Improving the characterization of physician non-English language

skills. *Health Services Research, 47*(1 Pt 2), 556–569. https://doi.org/10.1111/j.1475-6773.2011.01338.x

Evans, L. V., Morse, J. L., Hamann, C. J., Osborne, M., Lin, Z., & D'Onofrio, G. (2009). The development of an independent rater system to assess residents' competence in invasive procedures. *Academic Medicine: Journal of the Association of American Medical Colleges, 84*(8), 1135–1143. https://doi.org/10.1097/ACM.0b013e3181acec7c

Fahim M. (2011) The effects of rater training on raters' severity and bias in second language writing assessment. *Iranian Journal of Language Testing, 1*(1):1–16.

Feldman, M., Lazzara, E. H., Vanderbilt, A. A., & Diaz Granados, D. (2012). Rater training to support high-stakes simulation-based assessments. *The Journal of Continuing Education in the Health Professions, 32*(4), 279–286. https://doi.org/10.1002/chp.21156

Friedman, M., Sutnick, A. I., Stillman, P. L., Norcini, J. J., Anderson, S. M., Williams, R. G., Henning, G., & Reeves, M. J. (1991). The use of standardized patients to evaluate the spoken-English proficiency of foreign medical graduates. *Academic Medicine: Journal of the Association of American Medical Colleges, 66*(9 Suppl), S61–S63. https://doi.org/10.1097/00001888-199109000-00042

Gardner, A. K., Russo, M. A., Jabbour, I. I., Kosemund, M., & Scott, D. J. (2016). Frame-of-reference training for simulation-based intraoperative communication assessment. *American Journal of Surgery, 212*(3), 548–551.e2. https://doi.org/10.1016/j.amjsurg.2016.02.009

Gingerich, A., Ramlo, S. E., van der Vleuten, C., Eva, K. W., & Regehr, G. (2017). Inter-rater variability as mutual disagreement: identifying raters' divergent points of view. *Advances in Health Sciences Education: Theory and Practice, 22*(4), 819–838. https://doi.org/10.1007/s10459-016-9711-8

Gomez-Garibello, C., & Young, M. (2018). Emotions and assessment: considerations for rater-based judgements of entrustment. *Medical Education, 52*(3), 254–262. https://doi.org/10.1111/medu.13476

Hardin, K. J., & Hardin, D. M. (2013). Medical Spanish programs in the United States: a critical review of published studies and a proposal of best practices. *Teaching and Learning in Medicine, 25*(4), 306–311. https://doi.org/10.1080/10401334.2013.827974

Humphrey-Murto, S., Smee, S., Touchie, C., Wood, T. J., & Blackmore, D. E. (2005). A comparison of physician examiners and trained assessors in a high-stakes OSCE setting. *Academic Medicine: Journal of the Association of American Medical Colleges, 80*(10 Suppl), S59–S62. https://doi.org/10.1097/00001888-200510001-00017

Iramaneerat, C., & Yudkowsky, R. (2007). Rater errors in a clinical skills assessment of medical students. *Evaluation & the Health Professions*, 30(3), 266–283. https://doi.org/10.1177/0163278707304040

Kang, O., Rubin, D., & Kermad, A. (2019). The effect of training and rater differences on oral proficiency assessment. *Language Testing, 36*(4), 481–504. https://doi.org/10.1177/0265532219849522

Karuppaiah, S., & Raof, A. H. A. (2020). The impact of rater training on rater reliability in an English oral test. *Asian Journal of Assessment in Teaching and Learning, 10*(2), 94-105. https://doi.org/10.37134/ajatel.vol10.2.10.2020

Klein, R., Ufere, N. N., Schaeffer, S., Julian, K. A., Rao, S. R., Koch, J., Volerman, A., Snyder, E. D., Thompson, V., Ganguli, I., Burnett-Bowie, S. M., & Palamara, K. (2022). Association between resident race and ethnicity and clinical performance assessment

scores in graduate medical education. *Academic Medicine: Journal of the Association of American Medical Colleges*. Advance online publication. https://doi.org/10.1097/ACM.0000000000004743

Kobak, K. A., Lipsitz, J. D., Williams, J. B., Engelhardt, N., & Bellew, K. M. (2005). A new approach to rater training and certification in a multicenter clinical trial. *Journal of Clinical Psychopharmacology, 25*(5), 407–412. https://doi.org/10.1097/01.jcp.0000177666.35016.a0

Kogan, J. R., Conforti, L. N., Bernabeo, E., Iobst, W., & Holmboe, E. (2015). How faculty members experience workplace-based assessment rater training: A qualitative study. *Medical Education, 49*(7), 692–708. https://doi.org/10.1111/medu.12733

Lippi-Green, R. (2012). *English with an accent: Language, ideology, and discrimination in the United States*. Routledge.

Ludbrook, J., & Marshall, V. R. (1971). Examiner training for clinical examinations. *British Journal of Medical Education, 5*(2), 152–155. https://doi.org/10.1111/j.1365-2923.1971.tb02020.x

Martínez, G. A. (2010). Medical Spanish for heritage learners: A prescription to improve the health. *Building Communities and Making Connections,* 2–15.

McIver, D., Fitzsimmons, S., & Flanagan, D. (2015). Instructional design as knowledge management. *Journal of Management Education, 40*(1), 47–75. https://doi.org/10.1177/1052562915587583

Morales, R., Rodriguez, L., Singh, A., Stratta, E., Mendoza, L., Valerio, M. A., & Vela, M. (2015). National Survey of Medical Spanish Curriculum in US Medical Schools. *Journal of General Internal Medicine, 30*(10), 1434–1439. https://doi.org/10.1007/s11606-015-3309-3

Mvududu, N., & Thiel-Burgess, J. (2012). Constructivism in practice: The case for English language learners. *International Journal of Education, 4*(3), 108.

O'Sullivan, B. (2012). Assessment issues in languages for specific purposes. *The Modern Language Journal, 96*, 71–88. https://doi.org/10.1111/j.1540-4781.2012.01298.x

Ortega, P., Diamond, L., Alemán, M. A., Fatás-Cabeza, J., Magaña, D., Pazo, V., Pérez, N., Girotti, J. A., Ríos, E., & Medical Spanish Summit (2020). Medical Spanish standardization in US medical schools: Consensus statement from a multidisciplinary expert panel. *Academic Medicine: Journal of the Association of American Medical Colleges, 95*(1), 22–31. https://doi.org/10.1097/ACM.0000000000002917

Ortega, P., Francone, N. O., Santos, M. P., Girotti, J. A., Shin, T. M., Varjavand, N., & Park, Y. S. (2021a). Medical Spanish in US medical schools: A national survey to examine existing programs. *Journal of General Internal Medicine, 36*(9), 2724–2730. https://doi.org/10.1007/s11606-021-06735-3

Ortega, P., Hardin, K., Pérez-Cordón, C., Cox, A. O., Kim, K. C., Truesdale, D., Chang, R., Martínez, G. A., Miller De Rutté, A. M., Pérez-Muñoz, C., Rolón, L., & Shin, T. M. (2021b). An overview of online resources for medical Spanish education for effective communication with Spanish-speaking patients. *Teaching and Learning in Medicine*, 1–13. Advance online publication. https://doi.org/10.1080/10401334.2021.1959335

Ortega, P., Martínez, G., Alemán, M. A., Zapién-Hidalgo, A., & Shin, T. M. (2022). Recognizing and dismantling raciolinguistic hierarchies in Latinx health. *AMA Journal of Ethics, 24*(4), E296–E304. https://doi.org/10.1001/amajethics.2022.296

Ortega, P., Moxon, N. R., Chokshi, A. K., Pérez-Cordón, C., & Park, Y. S. (2021c). Validity evidence supporting the Comunicación y Habilidades Interpersonales (CAI) scale for medical Spanish communication and interpersonal skills assessment. *Academic Medicine: Journal of the Association of American Medical Colleges, 96*(11S), S93–S102. https://doi.org/10.1097/ACM.0000000000004266

Ortega, P., Park, Y. S., & Girotti, J. A. (2017). Evaluation of a medical Spanish elective for senior medical students: improving outcomes through OSCE assessments. *Medical Science Educator, 27*(2), 329–337. https://doi.org/10.1007/s40670-017-0405-5

Ortega, P., & Prada, J. (2020b). Words matter: Translanguaging in medical communication skills training. *Perspectives on Medical Education, 9*(4), 251–255. https://doi.org/10.1007/s40037-020-00595-z

Ortega, P., Shin, T. M., Francone, N. O., Santos, M. P., Girotti, J. A., Varjavand, N., & Park, Y. S. (2021d). Student and faculty diversity is insufficient to ensure high-quality medical Spanish education in us medical schools. *Journal of Immigrant and Minority Health, 23*(5), 1105–1109. https://doi.org/10.1007/s10903-021-01198-4

Piaget, J. (1968). *Six psychological studies*. Anita Tenzer (Trans.), New York: Vintage Books

Pill, J., & Woodward-Kron, R. (2012). How professionally relevant can language tests be?: A response to Wette (2011). *Language Assessment Quarterly, 9*(1), 105–108. https://doi.org/10.1080/15434303.2011.637263

Prada, J. (2019). Exploring the role of translanguaging in linguistic ideological and attitudinal reconfigurations in the Spanish classroom for heritage speakers. *Classroom Discourse*, *10*(3–4), 306–322. https://doi.org/10.1080/19463014.2019.1628793

Rosen, M. A., Salas, E., Silvestri, S., Wu, T. S., & Lazzara, E. H. (2008). A measurement tool for simulation-based training in emergency medicine: the simulation module for assessment of resident targeted event responses (SMARTER) approach. *Simulation in Healthcare: Journal of the Society for Simulation in Healthcare, 3*(3), 170–179. https://doi.org/10.1097/SIH.0b013e318173038d

Rothman, A. I., & Cusimano, M. (2001). Assessment of English proficiency in international medical graduates by physician examiners and standardized patients. *Medical Education, 35*(8), 762–766. https://doi.org/10.1046/j.1365-2923.2001.00964.x

Ryan, C. (2013, August) *Languages in the United States: 2011*. US Census Bureau. https://www2.census.gov/library/publications/2013/acs/acs-22/acs-22.pdf

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park: Sage Publications.

Skinner, B. F. (1976). *About behaviorism*. New York: Vintage Books

Suharno, S. (2010). Cognitivism and its implication in the second language learning. *Parole: Journal of Linguistics and Education, 1*, 72–96.

Tang, G., Lanza, O., Rodriguez, F. M., & Chang, A. (2011). The Kaiser Permanente clinician cultural and linguistic assessment initiative: Research and development in patient-provider language concordance. *American Journal of Public Health*, *101*(2), 205–208. https://doi.org/10.2105/AJPH.2009.177055

United States Civil Rights Division. (2015). 2015 US National Limited English Proficient (LEP) Population Maps: Number by State. https://www.lep.gov/sites/lep/files/resources/US_state_LEP_count.ACS_5yr.2015.pdf

Vygotsky, L. (1978). *Mind in society*. London: Harvard University Press.

Watt, D., Lake, D., Cabrnoch, T., & Leonard, K. (2003). Assessing the English language proficiency of international medical graduates in their integration into Canada's physician supply. Report commissioned by the Canadian Task Force on Licensure of International Medical Graduates, Ottawa, Ontario, Canada. Retrieved November 18, 2022: https://silo.tips/download/assessing-the-english-language-proficiency-of-international-medical-graduates-in

Yudkowsky, R., Hyderi, A., Holden, J., Kiser, R., Stringham, R., Gangopadhyaya, A., Khan, A., & Park, Y. S. (2019). Can nonclinician raters be Ttained to assess clinical reasoning in postencounter patient notes? *Academic Medicine: Journal of the Association of American Medical Colleges, 94*(11S Association of American Medical Colleges Learn Serve Lead: Proceedings of the 58th Annual Research in Medical Education Sessions), S21–S27. https://doi.org/10.1097/ACM.0000000000002904

Zhang, J. (2020). The construction of college English online learning community under ADDIE Model. *English Language Teaching, 13*(7), 46–51.

Zong, J. & Batalova, J. (2015). The limited English proficient population in the United States in 2013. *Migration Information Source.* https://www.migrationpolicy.org/article/limited-english-proficient-population-united-states-2013