Title

Development of a Tool to Assess Medical Oral Language Proficiency.

Journal

Academic Medicine, 98(4)

Authors

González, Javier
Pérez-Cordón, Cristina
Iniguez, Reniell
et al.

Publication Date

2023-04-01

DOI

10.1097/ACM.0000000000004942

Peer reviewed

HHS Public Access

Author manuscript

Acad Med. Author manuscript; available in PMC 2024 April 01.

Published in final edited form as:

Acad Med. 2023 April 01; 98(4): 480–490. doi:10.1097/ACM.0000000000004942.

Development of a Tool to Assess Medical Oral Language Proficiency

Lisa C. Diamond, MD, MPH [associate attending physician],
Immigrant Health and Cancer Disparities Service, Hospital Medicine Service, Departments of Medicine and Psychiatry and Behavioral Sciences, Memorial Sloan Kettering Cancer Center, New York, New York;

Steven E. Gregorich, PhD [professor emeritus],
Division of General Internal Medicine, Department of Medicine, University of California, San Francisco, San Francisco, California.

Leah Karliner, MD, MAS [professor],
Division of General Internal Medicine, Center for Aging in Diverse Communities, Multiethnic Health Equity Research Center, Department of Medicine, University of California, San Francisco, San Francisco, California;

Javier González, MFA [program manager],
Language Initiatives Program, Immigrant Health and Cancer Disparities Service, Department of Psychiatry and Behavioral Sciences, Memorial Sloan Kettering Cancer Center, New York, New York.

Cristina Pérez-Cordón, PhD,
language assessment specialist in the Language and Communication Training Unit, United Nations Headquarters, New York, New York.

Reniell Iniguez [medical student],
University of Illinois College of Medicine, Chicago, Illinois.

José Alberto [Figueroa medical student],
Northwestern University Feinberg School of Medicine, Chicago, Illinois.

Karen Izquierdo [medical student],

Correspondence should be addressed to Lisa Diamond, Immigrant Health and Cancer Disparities Service, Memorial Sloan Kettering Cancer Center, 485 Lexington Avenue, 2nd Floor, New York, New York 10017; telephone: (646) 888-8061; diamondl@mskcc.org; Twitter: @DrLisaDiamond.

Ethical Approval: The initial study where the POLOM was designed was approved by the Institutional Review Board (IRB) Committee on Human Research at the University of California, San Francisco, on November 24, 2015 (study #15–16762). The use of the videotapes for research was determined to meet criteria for exemption by the IRB at the University of Illinois on November 24, 2020 (protocol # 2019–0945).

Disclaimers: Design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, and approval of the manuscript; and decision to submit the manuscript for publication were independent of Memorial Sloan Kettering Cancer Center, University of California, San Francisco, University of Illinois at Chicago, and the National Board of Medical Examiners. The statements presented are solely the responsibility of the author(s) and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors, or Methodology Committee.

Previous presentations: P.O. and R.I. orally presented portions of this work at the Sixth International Symposium on Languages for Specific Purposes Conference in Chicago, Illinois, on April 22, 2022.

Supplemental digital content for this article is available at [LWW INSERT LINK]

College of Medicine, State University of New York at Downstate Health Sciences University, Brooklyn, New York.

Pilar Ortega, MD [clinical assistant professor]
Departments of Medical Education and Emergency Medicine, University of Illinois College of Medicine, Chicago, Illinois;

Abstract

Purpose: To communicate with linguistically diverse patients, medical students and physicians often use their non-English language skills. However, there is no standard protocol to determine whether those skills are adequate prior to patient care. This causes many physicians, institutions, educators, and learners to forgo non-English language proficiency assessment altogether. The purpose of this study is to report on the development, refinement, and interrater reliability of the Physician Oral Language Observation Matrix (POLOM), a rater-based tool assessing 6 language skill categories observed during clinical interactions: comprehension, fluency/fluidity, vocabulary, pronunciation, grammar, and communication. This study focused on the use of the POLOM in Spanish interactions.

Method: The authors adapted an existing language observation tool for use in clinical settings, creating the preliminary POLOM. Next, they iteratively refined the tool from April to July 2021 using videorecorded medical student-standardized patient encounters from a U.S.-based medical Spanish program. In each refinement iteration, 4 bilingual raters (2 physicians and 2 linguists) independently rated 3 to 6 encounters and convened to discuss ratings with the goals of improving instrument instructions, descriptors, and subsequent rater agreement. Using the final POLOM, raters independently rated 50 videos in rotating interdisciplinary pairs. Generalizability theory was applied to estimate reliability via interrater agreement (dependability) coefficients (range 0–1) for each POLOM category and the total score.

Results: POLOM total score dependability equaled 0.927 (single rater) and 0.962 (averaged across 2 raters). The highest mean score was observed for the comprehension category (4.15; range: 1–5) while the lowest was for communication (3.01; range 1–5).

Conclusions: Raters achieved a high level of agreement on POLOM assessments of students' medical oral Spanish proficiency. The POLOM is the first such assessment tool that provides examinees and instructors with both a holistic and detailed review of clinician non-English oral language skills as contextualized for patient care.

More than 67 million people in the United States speak a non-English language at home, the majority speaking Spanish.[1] Of those who speak a language other than English at home, 42% report limited English proficiency (LEP).[2] Even individuals with non-English language preference who speak some English may encounter difficulties communicating health concepts in English.[3] Title VI of the 1964 U.S. Civil Rights Act, as implemented by Executive Order 13166[4] and the Affordable Care Act,[5] established that individuals must have meaningful access to federal services, including health care, regardless of language abilities. The National Standards for Culturally and Linguistically Appropriate Services (CLAS) provide guidelines for implementing language-appropriate services in health care, including through medical interpreters and bilingual providers.[6] Despite these federal

protections and guidelines, however, patients with LEP frequently report dissatisfaction with health care[7]; receive worse care than their English-speaking counterparts in the American health care system[8] due to health-related misunderstandings[9]; and experience more harmful adverse events.[10]

There are 2 methods to address language-appropriate care for patients with LEP: (1) using a professional interpreter and (2) matching a patient with a language-concordant clinician (a clinician who is proficient in the patient's preferred language). A vast majority of U.S. medical residency applicants report some skills in at least 1 non-English language,[11] and research shows that medical students and physicians often use their non-English skills to communicate directly with linguistically diverse patients, regardless of proficiency.[12,13] Using nonproficient language skills to communicate with patients has implications for patient safety and quality of care.[10,14] However, there is no standard protocol to determine whether clinician non-English skills are adequate to communicate with patients without an interpreter. Although 78% of U.S. medical schools report offering medical Spanish education (courses that aim to teach Spanish-speaking clinicians to use Spanish with patients), 43% of programs do not include any assessment of learner medical oral language proficiency prior to patient care.[15] Similarly, although the CLAS Standards define a "bilingual provider" as an "individual with proficiency in more than one language," no guidance is given as to how proficiency should be assessed or what level is sufficient for direct patient care. The lack of proficiency assessment highlights a concerning gap in U.S. medical education and health care systems because students and clinicians learning medical Spanish may not receive sufficient feedback on their performance to know when they can safely and accurately use their language skills with patients.

Medical oral language proficiency can be defined as how clinicians communicate in a particular language with patients in real-world, spontaneous, nonrehearsed contexts through speaking and listening, with this definition adapted from the American Council on the Teaching of Foreign Languages (ACTFL) general language proficiency definitions.[16] Potential approaches to evaluate medical student and physician non-English skills include self-assessment, oral proficiency interviews (OPIs), objective structured clinical examinations (OSCEs), and direct observation of clinical encounters.

One applicable self-assessment tool is the Interagency Language Roundtable (ILR) scale, which has a version modified for health care.[17] The ILR health care scale has been used by the Association of American Medical Colleges for documenting the language skills of medical school and residency candidates and by hospitals for documenting skills of practicing physicians.[18] The ILR has been shown to be as accurate as an OPI and, therefore, valid for those who self-assess on the low and high ends of the scale.[18] For those who self-assess in the middle of the scale, further assessment is required to determine skill level. Thus, the ILR has been recommended as a screening tool for students enrolling in medical language courses rather than a certification tool for independent patient care.[19]

OPIs offer another method of assessing language proficiency. The Clinician Cultural and Linguistic Assessment (CCLA)[20] is a validated OPI specific to assessing clinician language skills. This telephone-based exam evaluates proficiency, fluency, pronunciation, customer

service, and cultural proficiency. Clinician tasks involve listening to prerecorded clinical scenarios and recording a verbal response. Other clinically relevant OPIs exist but target individuals who wish to become certified as medical interpreters,[21,22] and thus are not appropriate for assessing the proficiency levels of medical students or physicians. While there are other validated OPIs, such as one by ACTFL,[23] these are not clinically relevant. In general, OPIs present limitations for medical language assessment due to a lack of interactivity and authenticity.[24]

OSCEs represent an opportunity to evaluate student language proficiency in a high-fidelity clinical setting. OSCEs provide a simulated clinical encounter in which learners interview trained standardized patients (SPs), and they are an accepted methodology for medical student formative and summative assessment. Presently, there is no validated tool to assess student language proficiency in non-English OSCEs. Among medical schools that offer Spanish courses, 29% report using SP encounters, but they lack a standardized rubric to rate student non-English language skills.[15] Only 1 previous study examined the implementation of medical school OSCEs with Spanish-speaking SPs as part of a longitudinal medical Spanish curriculum. The study found a lack of agreement between faculty and SP ratings of learners' performance,[25] suggesting the need for an objective medical language scoring tool.

To address this gap, we adapted an existing language observation tool for use in clinical settings. We report here on the tool adaptation process, its refinement using videorecorded OSCE interactions between medical students and Spanish-speaking SPs, the refined tool—the Physician Oral Language Observation Matrix (POLOM), and the reliability (interrater agreement) obtained by a group of experienced raters using the POLOM. The purpose of the current study was to present the POLOM's development, refinement, and interreliability achieved by experienced raters as a first step in rigorously evaluating a new rating tool for assessing physician medical oral non-English proficiency.

Method

The preliminary POLOM

The POLOM began as an adaptation of the Student Oral Language Observation Matrix (SOLOM) (see Supplemental Digital Appendix 1 at [LWW INSERT LINK]). The SOLOM was developed by the California Department of Education in 1978 to allow instructors to rate students' listening and speaking abilities in any language. Instructors use the SOLOM to rate language proficiency using 5 ordinal options for each of 5 categories: comprehension, fluency, vocabulary, pronunciation, and grammar. Each SOLOM response option includes a textual description that is tailored to the category and proficiency level. The SOLOM is commonly used in educational and research settings.[26,27]

Initially, an interdisciplinary team comprising 2 bilingual physicians (L.C.D. and L.K.), 1 linguist (J.G.), and 1 psychometrician (S.E.G.) adapted the SOLOM for use in patient-physician interactions and named the adapted instrument the POLOM. Like the SOLOM, the POLOM allows raters to assess spoken and receptive language but contextualizes the observation to clinical settings. The preliminary POLOM evaluated physician skills in the same 5 categories as the SOLOM, with category-specific scores ranging from 1 (not

proficient) to 5 (fully proficient). The adaptation entailed modifying the textual descriptions of each ordinal rating option. The initial adaptation was approved by the Institutional Review Board (IRB) at the University of California, San Francisco, on November 24, 2015 (study #15–16762).

Initial evaluation and iterative adaptation of the preliminary POLOM

We piloted the preliminary POLOM with 13 audiorecorded Spanish-speaking patient primary care visits collected as part of a larger study on language access from March to October 2018.[28] A linguist (J.G. or C.P.C.) and/or a physician (L.C.D. or P.O.) used the POLOM to rate each clinician's performance. The development team met and made iterative changes to the POLOM based on raters' reports. POLOM changes included: (1) adaptations to the qualifiers that captured frequency of errors (e.g., "rarely"; "often"); (2) addition of items to assess dynamics influencing communication (e.g., "use of English by patient or physician"); and (3) introduction of the idea of "repairing" (i.e., the physician's ability to recognize and fix communication errors), which resulted in the addition of a sixth POLOM category: "communication." Both raters scored 9 encounters and achieved low levels of category-specific agreement. This suggested that the preliminary POLOM required significant refinement to improve the likelihood of reaching an acceptable level of interrater agreement, and that raters required more training and experience using the POLOM. In addition, 3 audiorecordings were deemed insufficient for evaluating the physician's Spanish level due to: extensive use of English by physician and patient during the visit, a family member frequently acting as an ad hoc interpreter, or little direct conversation with the patient because of impaired cognition. The complexities and unpredictability of real clinical encounters evident in these recordings highlighted that the POLOM needed to be further refined in more standardized encounters prior to further study in the clinical environment, particularly when assessing trainees. Figure 1 shows the POLOM development and subsequent refinement process.

Refinement of the POLOM

From April to July 2021, we refined the POLOM using medical student-SP encounters that had been previously videorecorded for educational purposes. The advantages of this approach included standard scenario content, sufficient quantities of videos, and varied student Spanish levels. We drew our sample from videorecorded student-SP encounters (n = 356) collected from a medical Spanish course for third- and fourth-year medical students at the University of Illinois College of Medicine from 2013–2020. Encounters included 4 standardized scenarios: pelvic pain, upper abdominal pain, chest pain, and shortness of breath. The use of the videotapes for research was determined to meet criteria for exemption by the University of Illinois IRB on November 24, 2020 (protocol #2019–0945).

The bilingual rating team included 2 physicians (1 medical Spanish educator and clinician who grew up speaking English and Spanish [P.O.] and 1 clinician who is a Spanish as a second-language speaker (L.C.D.]) and 2 linguists (1 expert in language access [J.G.] and 1 language teaching/assessment specialist [C.P.C.]; both were raised speaking Spanish and report advanced-level English proficiency). All raters self-reported a Spanish level of "excellent" on the ILR health care scale.[17]

At any iteration of the refinement process (Figure 1), we used the most recent POLOM version; each rater independently viewed and rated 3 to 6 preassigned SP encounters. Raters then met as a group to compare scores, resolve inconsistencies, consider any proposed POLOM revisions, and reach consensus on any revisions via discussion. If all raters approved, we updated, reviewed, and modified the POLOM, and then we used the newest version in the next rating round.

Demonstration of interrater agreement

Raters discarded 10 videos during rating rounds: 8 due to concerns that the SP's Spanish was insufficient for controlled examination purposes (n = 3 SPs), 1 due to poor audio, and 1 due to concerns that the student was reading from a script rather than speaking spontaneously. We excluded 3 other videos because they were additional encounters from students already in the sample. Ultimately, following POLOM refinement, alternating pairs of linguist and physician raters rerated 50 of the 63 originally rated SP encounters.

Each week, raters were assigned to pairs that consisted of 1 physician and 1 linguist. The pairing assignments rotated so that all 4 physician-linguist pairings occurred with similar frequency. Members of each rater pair independently viewed randomly selected videos and used the POLOM to rate students' Spanish proficiency. Thus, the data included 100 data records representing 50 encounters, each independently rated by 2 raters. These POLOM ratings are the focus of the quantitative analyses below. The primary quantitative aim of the process was to determine whether experienced raters could use the POLOM to provide reliable ratings. If the reliability of ratings was demonstrated, a secondary aim was to report upon the descriptive statistics of the ratings.

Statistical analysis

We used generalizability (G) theory to estimate dependability (agreement) coefficients for ratings on each of the 6 POLOM categories as well as the POLOM total score.[29] Application of G theory proceeds in 2 steps.[30,31] In the first step, a G study estimates variance components of outcome response (POLOM scores) that are attributable to the sources under investigation (i.e., students, raters, residual). In the second step, a decision (D) study uses the G study variance component estimates to calculate agreement coefficients.

Seven G study models—1 per POLOM category, plus total score—estimated variance components for students (s), raters (r), and residual (d). In this study, the residual confounds the students-by-raters and random error sources of variation. In G theory parlance, students are the objects of measurement and both students, and raters are the facets of measurement. In the analyses, students and raters were regarded as random facets because the goal of the analyses was to generalize to the populations of potential students and raters. All G study models were fit using SAS PROC MIXED with restricted maximum likelihood (SAS/Stat 15.1, SAS Institute Inc., Cary, NC). For each G study, we descriptively report the percentage of total variation attributable to each estimated variance component.

D studies estimated a type of agreement known as the dependability coefficient (Φ), which is akin to reliability but reflects absolute agreement across raters; that is, a high level of dependability would require independent raters to provide highly similar POLOM 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. A future validity study would test the degree to which raters' POLOM scores accurately measure medical students' Spanish-language proficiency. If validated, the POLOM is intended to assess medical oral language proficiency in an absolute sense, not simply the relative standing of students. Therefore, we did not consider relative agreement coefficients. D studies can estimate the dependability of ratings from a single rater as well as the dependability of ratings averaged across any number of raters. We report dependability coefficients assuming POLOM scores are both provided from a single rater and averaged across 2 raters. Dependability coefficients were estimated via Equation 1, where, e.g., $\hat{\sigma}_s^2$ represents the variance component estimate for students and n_r equaled 1 or 2 for the dependability of ratings from a single rater versus averaging across 2 raters, respectively. Dependability coefficients have a possible range of 0–1.

$$\hat{\Phi} = \frac{\hat{\sigma}_s^2}{\hat{\sigma}_s^2 + \frac{\hat{\sigma}_r^2}{n_r} + \frac{\hat{\sigma}_d^2}{n_r}} \quad \text{Eq. 1}$$

Results

From the selected videos of 50 encounters, 64% (n = 32) of students identified as female and 36% (n = 18) male, 14% (n = 7) Asian, 14% (n = 7) Black, 30% (n = 15) White, and 42% (n = 21) Hispanic/Latinx. In all, 38% (n = 19) were Spanish heritage speakers (i.e., they grew up speaking Spanish at home). Based on the ILR health care scale,[17] candidate self-reported Spanish levels were: 18% (n = 9) "excellent"; 38% (n = 19) "very good"; 24% (n = 12) "good"; and 20% (n = 10) "fair."

Qualitative results: Refinement of the POLOM

Refinements included modifying the textual descriptions of category-specific rating options and clarification of instructions in scoring each category (e.g., guidance on what to do if unsure between adjacent rating options or if 1 error could be attributed to more than 1 category). This iterative process repeated until all raters agreed that no further changes were required. During POLOM refinement, raters evaluated 63 encounters over 4 months (April-July 2021). We selected these 63 encounters to represent a range of student Spanish-language proficiency. It is important to note that the extensive process of POLOM refinement also provided the raters with intensive training in using the POLOM. By the end of the refinement process, each of the raters had a high level of experience and expertise using the POLOM.

The final version of the POLOM is shown in Table 1 and includes the finalized definitions of each of the 6 categories and the textual descriptions of observable behaviors that determine level 1–5 ratings in each category. The first 5 categories of comprehension, fluency/fluidity, vocabulary, pronunciation, and grammar point to specific linguistic features of the candidate's use of Spanish with the patient. The sixth category, communication, is not summative of all the other categories but rather represents how well the candidate is able

to integrate Spanish in the context of the social interaction with the patient. Depending on the encounter and the patient, this may, for example, involve adjusting the communication register (e.g., level of formality, use of technical vocabulary) to ensure that the patient is able to understand a complex medical concept or respectfully addressing topics that feel sensitive to the patient. These adjustments can have a significant impact on the effectiveness of communication, yet they are not captured by other linguistic features. For example, vocabulary use can be linguistically correct, but if the content was delivered in a way that was unclear to the patient, then the communication will have been impeded. The intent of the POLOM is to address language proficiency in a medical setting that closely resembles the candidate's typical job duties, not their medical knowledge. Table 2 summarizes the 22 iterative POLOM revisions.

Quantitative results: Rater agreement

Table 3 shows the decomposition of POLOM score variation by source (student, rater, residual) and dependability coefficients. Overall, between-student variability dominated, ranging from 74% (comprehension) to 93% (total score), suggesting that score variation predominantly reflected differences in student performance. In contrast, between-rater variation was negligible (0% to 3%), suggesting no substantial systematic differences between raters in terms of their POLOM ratings. In this design, residual variation, which confounds variation attributable to random noise and students-by-rater interaction, was small-to-moderate (7% to 25%). Correspondingly, dependability coefficients had very high values: 0.926 and 0.961 for POLOM total scores from a single rater and a 2-rater average, respectively. In the context of our trained raters who developed a high degree of expertise using the POLOM during the refinement process, the dependability coefficients suggest that a single individual's rating of a particular encounter will be 93% similar to that of another rater. The reliability is even higher, 96%, when the ratings of 2 expert raters are averaged for the same encounter. Dependability of individual POLOM category ratings were good (0.738: comprehension, single rater) to very high (0.941: communication, 2-rater average), with all category-specific dependability coefficients greater than or equal to 0.849 for ratings averaged across 2 raters. POLOM mean scores by SP scenario are provided in Supplemental Digital Appendix 2 at [LWW INSERT LINK].

Quantitative results: POLOM descriptive statistics

Table 4 shows descriptive statistics for the ratings that the trained raters generated, the POLOM category-specific scores, and the total scores across 100 ratings. The highest mean score was observed for the comprehension category (4.15; range 2–5), while the lowest was for communication (3.01; range 1–5). Since the rationale for adding the communication category to the rubric was to identify any factors that impeded communication with the patient (including but not limited to limitations or errors identified in comprehension, fluency/fluidity, vocabulary, pronunciation, or grammar), communication scores may be affected by the candidate's scores in any of the other 5 categories. This explains why the communication category had the lowest mean score.

Discussion

In this study, we developed and extensively refined the POLOM to create a tool in which experienced raters obtained high interrater agreement.. We established that experienced raters can provide highly reliable scores, which is an important first step in the programmatic development of the POLOM as a tool to assess clinical Spanish language proficiency. Our goal is that the POLOM will be validated in a future study and adapted to other languages, so that it can become the first direct observation, standardized rating tool for medical oral language proficiency in a non-English language.

Providing medical students with not only a total score but also a category-specific breakdown of their medical oral language proficiency has important implications for medical education. First, the lack of an evidence-based curriculum is a recognized challenge for medical Spanish educators.[19] By identifying learner strengths and weaknesses, educational curricula can be tailored to address the most commonly challenging areas or customized to meet individual students' needs. Additionally, once validated, the POLOM potentially could be administered at multiple points of a course to assess learner needs and progress. It could thus be used as a tool for formative assessment, which is recommended practice for medical school assessments.[31] Repetitive assessments, especially when given with appropriate performance feedback, can promote active learning and skill development[32] and positively impact long-term retention.[33]

Our study presents the results of an interdisciplinary collaboration involving both medical and language experts in POLOM development. Teaching and assessing medical Spanish ideally requires clinical knowledge and experience along with language pedagogy and evaluation training, yet few professionals meet all characteristics.[19] Thus, educators and researchers have called for interdisciplinary collaborations in creating educational tools.[34] Having both medical and language expert input in POLOM development and refinement was particularly valuable because these are the most common backgrounds of medical Spanish faculty, who represent potential future users of this rating tool.[15] As expected, ratings averaged across experienced physician-linguist rater pairs were more reliable, but POLOM scores from a single rater also were highly reliable. This suggests that in the future, once validity testing for the POLOM has been completed, should resources only allow for 1 trained rater, the dependability of the POLOM is still adequate, permitting flexibility with application of the tool by either an experienced physician or a linguist rater.

If, in the future, the POLOM is validated and then employed by properly trained raters, it will have potential health equity implications for linguistically diverse populations. Until now, clinical communication skills have only been consistently assessed in a standardized fashion to prepare U.S. medical students to care for English-speaking patients.[35] Thus, use of the POLOM could reduce structural barriers to care for the U.S. Spanish speaking population. A future goal is to review, refine, evaluate, and test the reliability and validity of the POLOM for use in other languages. Furthermore, the lack of valid tools to assess clinician language is an important limitation in prior language concordance research.[36] The POLOM is a rigorously developed tool that currently allows trained raters to reliably assess

medical oral language proficiency; therefore, it may contribute to a standardized definition of language-concordant care in future outcomes research.

Finally, once the POLOM has been validated, it has implications for patient safety. Providing medical students and physicians with medical Spanish learning opportunities without appropriate assessment can lead to a false sense of confidence related to non-English language use with patients despite insufficient skills.[37] Accordingly, clinicians may choose to "get by" with limited language skills, an approach that increases the risk of communication-related medical errors.[13] The POLOM could improve clinician decision-making around use of non-English skills, and thereby patient safety, by helping clinicians better understand when their skills suffice and when they should request a professional interpreter. Moreover, if validated, the POLOM could be used as a resource for more explicitly operationalizing "bilingual provider proficiency" using the CLAS Standards. The lack of a clear definition of proficiency level for bilingual providers is a current limitation to the current CLAS Standards that may contribute to low rates of compliance with recommended standards by U.S. health care institutions.[38]

Our study has limitations. First, the 50 videos selected for assessing interrater agreement had been previously rated by the same raters during POLOM refinement, which could have affected their scores. Raters did not think their prior exposure had substantial impact on their final ratings due to the 4 months that passed between ratings. Second, the tool refinement process used in this study effectively served as an intensive rater training program for the POLOM. It is uncertain whether other raters can be efficiently trained to achieve levels of interrater agreement similar those attained by study team members for reasonable POLOM deployment at scale. Third, demonstrations of interrater agreement are a necessary, but not sufficient, requirement of this type of tool. POLOM validity must also be investigated. Thus, questions about whether the POLOM is a valid measure of medical Spanish oral proficiency and, if so, how it might be applied to determine an individual's level of proficiency, are beyond the scope of this study. Tool validation will necessarily include comparing POLOM ratings to other measures of proficiency and setting proficiency thresholds. This will require careful consideration of the implications of particular scores, such as the minimum score needed for independent direct patient care in Spanish without a professional interpreter. Finally, unconscious biases about students could have affected ratings, and there is no way to capture this in our scoring system.

The POLOM, when utilized by sufficiently trained and skilled raters, allows for reliable assessment of medical oral language proficiency. Our team's next steps include developing an online module to train raters in the use of the POLOM (which includes unconscious bias training); assessing interrater agreement for newly trained raters; and conducting a validation study. Reliable assessment of Spanish-language use in clinical settings is an important step toward characterizing clinicians' medical language proficiencies and improving language-appropriate communication for diverse populations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding/Support:

L.C.D. was supported by Memorial Sloan Kettering Cancer Center Support Grant/Core Grant (P30 CA008748). L.C.D., S.F.G., L.K., J.G., C.P.C., and P.O. were supported by a grant from the National Board of Medical Examiners Stemmler Fund. Initial POLOM development was funded through a Patient-Centered Outcomes Research Institute (PCORI) Award (AD-1409-23627).

Other Disclosures:

L.C.D. receives author royalties from Multilingual Matters. P.O. receives author royalties from Elsevier.

References

1. U.S. Census Bureau, Characteristics of People by Language Spoken at Home, 2019 American Community Survey 1-Year Estimates Subject Tables. 2019. https://data.census.gov/cedsci/table?q=S1603&tid=ACSST1Y2019.S1603. Accessed July 25, 2022.
2. Ryan C Language Use in the United States: 2011. U.S. Census Bureau Report #ACS-22. Published April 1, 2010. https://www.census.gov/library/publications/2013/acs/acs-22.html. Accessed July 25, 2022.
3. Ortega P, Shin TM, Martínez GA. Rethinking the term “limited English proficiency” to improve language-appropriate healthcare for all. J Immigr Minor Health. 2022;24(3):799–805. doi:10.1007/s10903-021-01257-w [PubMed: 34328602]
4. Executive Office of the President. Improving access to services for persons with limited English proficiency. Executive Order 13166. Fed Regist 2000;65:50119–50122. https://www.federalregister.gov/d/00-20938. Accessed July 25, 2022.
5. Office for Civil Rights, Office of the Secretary, U.S. Department of Health and Human Services. Nondiscrimination in health programs and activities: final rule. Federal Register. 2016;81(96);Part IV. https://www.govinfo.gov/content/pkg/FR-2016-05-18/pdf/2016-11458.pdf. Accessed July 25, 2022.
6. U.S. Department of Health and Human Services, Office of Minority Health: National Standards for Culturally and Linguistically Appropriate Services (CLAS) in Health and Health Care. Published April 2013. https://thinkculturalhealth.hhs.gov/assets/pdfs/EnhancedCLASStandardsBlueprint.pdf. Accessed July 25, 2022.
7. Weech-Maldonado R, Elliott MN, Morales LS, Spritzer K, Marshall GN, Hays RD. Health plan effects on patient assessments of Medicaid managed care among racial/ethnic minorities. J Gen Intern Med. 2004;19(2):136–145. doi:10.1111/j.1525-1497.2004.30235.x [PubMed: 15009793]
8. Ngo-Metzger Q, Sorkin DH, Phillips RS, et al. Providing high-quality care for limited English proficient patients: the importance of language concordance and interpreter use. J Gen Intern Med. 2007;22 (2 suppl 2):S324–S330. doi:10.1007/s11606-007-0340-z
9. Solis JM, Marks G, Garcia M, Shelton D. Acculturation, access to care, and use of preventive services by Hispanics: findings from HHANES 1982–84. Am J Public Health. 1990;80 (uppl):S11–S19. doi:10.2105/ajph.80.suppl.11
10. Divi C, Koss RG, Schmaltz SP, Loeb JM. Language proficiency and adverse events in US hospitals: a pilot study. Int J Qual Health Care. 2007;19(2):60–67. doi:10.1093/intqhc/mzl069 [PubMed: 17277013]
11. Diamond L, Grbic D, Genoff M, et al. Non–English-language proficiency of applicants to US residency programs. JAMA. 2014;312(22):2405–2407. doi:10.1001/jama.2014.15444 [PubMed: 25490332]
12. Vela MB, Fritz C, Press VG, Girotti J. Medical students’ experiences and perspectives on interpreting for LEP patients at two US medical schools. J Racial Ethn Health Disparities. 2016;3(2):245–249. doi:10.1007/s40615-015-0134-7 [PubMed: 27271065]
13. Diamond LC, Schenker Y, Curry L, Bradley EH, Fernandez A. Getting by: Underuse of interpreters by resident physicians. J Gen Intern Med. 2009;24(2):256–262. doi:10.1007/s11606-008-0875-7 [PubMed: 19089503]

14. Prince D, Nelson M. Teaching Spanish to emergency medicine residents. Acad Emerg Med. 1995;2(1):32–37. doi:10.1111/j.1553-2712.1995.tb03076.x [PubMed: 7606608]
15. Ortega P, Francone NO, Santos MP, et al. Medical Spanish in US medical schools: A national survey to examine existing programs. J Gen Intern Med. 2021;36(9):2724–2730. doi:10.1007/s11606-021-06735-3 [PubMed: 33782890]
16. American Council on the Teaching of Foreign Languages (ACTFL) Language Connects. ACTFL Proficiency Guidelines 2012. https://www.actfl.org/resources/actfl-proficiency-guidelines-2012. Accessed July 25, 2022.
17. Diamond LC, Luft HS, Chung S, Jacobs EA. "Does this doctor speak my language?" Improving the characterization of physician non-English language skills. Health Serv Res 2012;47(1 Pt 2):556–569. doi:10.1111/j.1475-6773.2011.01338.x [PubMed: 22091825]
18. Diamond L, Chung S, Ferguson W, Gonzalez J, Jacobs EA, Gany F. Relationship between self-assessed and tested non-English-language proficiency among primary care providers. Med Care. 2014;52(5):435–438. doi:10.1097/MLR.0000000000000102 [PubMed: 24556893]
19. Ortega P, Diamond L, Alemán MA, et al. Medical Spanish standardization in U.S. medical schools: Consensus statement from a multidisciplinary expert panel. Acad Med. 2020;95(1):22–31. doi:10.1097/ACM.0000000000002917 [PubMed: 31365394]
20. Tang G, Lanza O, Rodriguez FM, Chang A. The Kaiser Permanente Clinician Cultural and Linguistic Assessment Initiative: Research and development in patient-provider language concordance. Am J Public Health. 2011;101(2):205–208. doi:10.2105/AJPH.2009.177055 [PubMed: 21228282]
21. Certification Commission for Health Care Interpreters. http://www.cchicertification.org. Accessed July 25, 2022.
22. The National Board of Certification for Medical Interpreters. https://www.certifiedmedicalinterpreters.org. Accessed July 25, 2022.
23. American Council on the Teaching of Foreign Languages (ACTFL) Language Connects. Oral Preficiency Interview (OPI). https://www.actfl.org/assessment-research-and-development/actfl-assessments/actfl-postsecondary-assessments/oral-proficiency-interview-opi. Accessed on July 25, 2022.
24. Shrum J, Glisan E. Teacher's handbook: Contextualized language instruction, 4th ed. Boston, MA: Heinle Cengage Learning; 2010.
25. Ortega P, Park YS, Girotti JA. Evaluation of a medical Spanish elective for senior medical students: Improving outcomes through OSCE assessments. Med Sci Educ. 2017;27(2):329–337. doi:10.1007/s40670-017-0405-5 [PubMed: 29910972]
26. Dennis LR, Krach SK, McCreery MP, Navarro S. The student oral Language observation matrix: A psychometric study with preschoolers. Assess Eff Interv. 2019;45(1):65–72. doi:10.1177/1534508418782624.
27. McConkey Robbins A, Green JE, Waltzman SB. Bilingual oral language proficiency in children with cochlear implants. Arch Otolaryngol Head Neck Surg. 2004;130(5):644–647. doi:10.1001/archotol.130.5.644 [PubMed: 15148191]
28. Karliner L, Diamond L, Toman J, et al. Testing a program to improve patient-clinician communication for patients who speak limited English. Patient-Centered Outcomes Research Institute (PCORI). 2021. 10.25302/02.2021.AD.140923627. Accessed August 9, 2022.
29. Shavelson RJ, Webb NM. Generalizability Theory: A Primer. Newbury Park, CA: SAGE Publications, Inc; 1991
30. Brennan RL. Generalizability Theory. New York, NY: Springer-Verlag; 2001.
31. Albanese MA. Commentary: Measurement and interpretation challenges in comparing student performance outcomes from different medical schools. Acad Med. 2011;86(9):1073–1075. doi:10.1097/ACM.0b013e318226340c [PubMed: 21865904]
32. Augustin M How to learn effectively in medical school: test yourself, learn actively, and repeat in intervals. Yale J Biol Med. 2014;87(2):207–212. [PubMed: 24910566]
33. Karpicke JD, Butler AC, Roediger HL 3rd. Metacognitive strategies in student learning: do students practise retrieval when they study on their own? Memory. 2009;17(4):471–479. doi:10.1080/09658210802647009 [PubMed: 19358016]

34. Shin TM, Hardin K, Johnston D, et al. Scoping review of textbooks for medical Spanish education. Med Sci Educ. 2021;31(4):1519–1527. doi:10.1007/s40670-021-01333-8 [PubMed: 34457990]

35. Ortega P Spanish language concordance in U.S. medical care: A multifaceted challenge and call to action. Acad Med. 2018;93(9):1276–1280. doi:10.1097/ACM.0000000000002307 [PubMed: 29877912]

36. Diamond L, Izquierdo K, Canfield D, Matsoukas K, Gany F. A systematic review of the impact of patient-physician non-English language concordance on quality of care and outcomes. J Gen Intern Med. 2019;34(8):1591–1606. doi:10.1007/s11606-019-04847-5 [PubMed: 31147980]

37. Lion KC, Thompson DA, Cowden JD, et al. Clinical Spanish use and language proficiency testing among pediatric residents. Acad Med. 2013;88(10):1478–1484. doi:10.1097/ACM.0b013e3182a2e30d [PubMed: 23969350]

38. Diamond LC, Wilson-Stronks A, Jacobs EA. Do hospitals measure up to the national culturally and linguistically appropriate services standards? Med Care. 2010;48(12):1080–1087. doi:10.1097/MLR.0b013e3181f380bc [PubMed: 21063229]

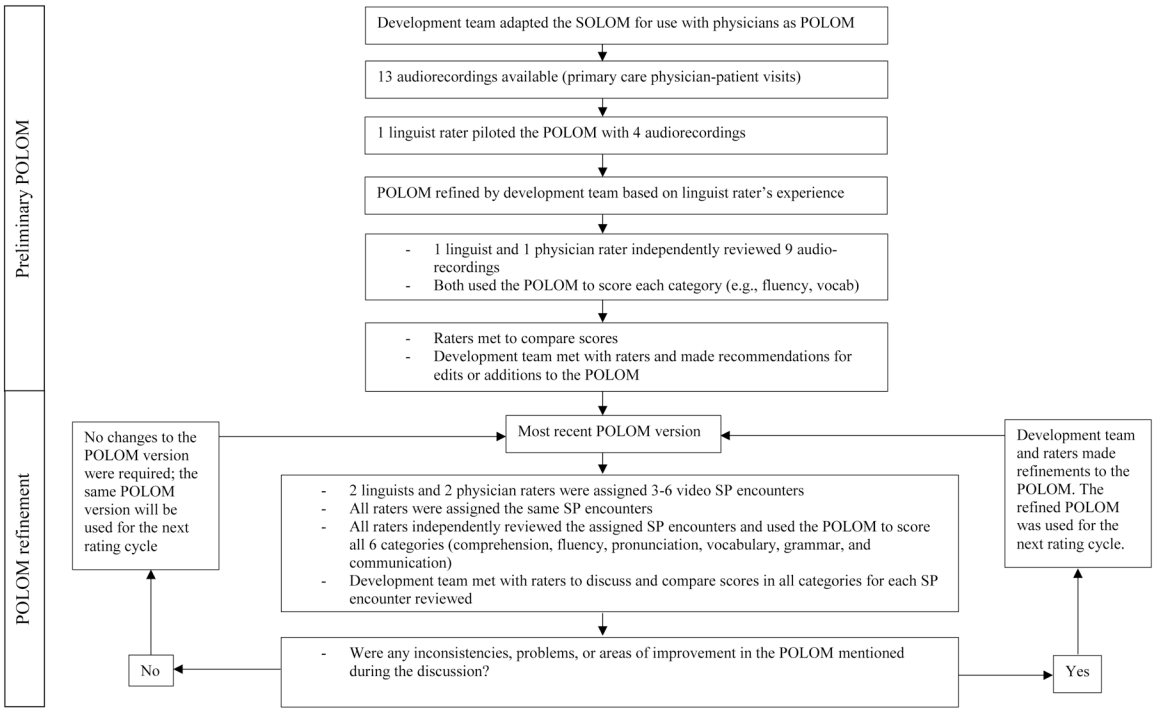

Figure 1.

Process for development and iterative refinement of the Physician Oral Language Observation Matrix (POLOM). Abbreviations: POLOM, Physician Oral Language Observation Matrix, SP, standardized patient; vocab, vocabulary.

Table 1

Final Version of the Physician Oral Language Observation Matrix (POLOM)[a]

Category and definition	Level 1	Level 2	Level 3	Level 4	Level 5
Comprehension: The candidate's ability to understand the patient's speech, including the understanding of sounds, words, and phrases.	The candidate cannot be said to understand even simple conversation. AND/OR The candidate's speech lacks acknowledgement of any of what the patient says or asks.	The candidate has difficulty following what is said and can understand only conversation spoken slowly with very frequent repetitions or use of English. AND/OR The candidate misunderstands or does not understand or acknowledge a portion of the conversation (e.g., a pivotal sentence, idea, or word) or most of the conversation, impeding communication or misleading the interview at some point during the encounter. AND This misunderstanding significantly impacts the content as observed by the rater.	The candidate understands what is said at slower than normal speed or with frequent repetitions. AND/OR The candidate does not understand a portion of the conversation (e.g., a pivotal sentence, idea, or word) and either does not acknowledge or ask for clarification or, after clarification, still does not understand, but the misunderstanding does not impact the content as observed by the rater.	The candidate understands nearly everything at normal speed, although occasional repetition at a slower speed or explanation may be necessary. If the candidate requests and receives clarification, explanation, or repetition from the patient, the candidate is then able to understand.	The candidate understands conversation at normal speed without difficulty (e.g., occasional requests for clarification of regionalisms or repetition of the same word or phrase may be acceptable in the context of overall excellent comprehension).
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.	The candidate's speech is so halting and fragmentary as to make conversation impossible or nearly impossible. AND/OR The candidate is unable to communicate beyond the sentence level.	The candidate's speech in conversation is markedly not fluid or smooth, such that it impairs or strains the conversation at some point. Indicators may include: • candidate is hesitant most of the time. • candidate is repeatedly stuck. • candidate is clearly forced into English (or another nontested language) or into unnatural pauses, silence, or excessive use of pet words. • candidate is unable to find the correct manner of expression at some specific point during the conversation.	The candidate's speech in conversation is not fluid or smooth, yet it still allows conversation to occur. AND The candidate's speech in conversation is disrupted by lapses due to the search for the correct manner of expression but eventually is able to express it. The occasional use of English (or another nontested language) to string words together might be acceptable if it is corrected to Spanish and/or explained and does not strain the conversation.	The candidate's speech in conversation is mostly fluid and smooth, with occasional natural, short lapses while they search for the correct manner of expression. Occasional use of English (or another nontested language) might be acceptable if it is quickly corrected to Spanish and does not strain the conversation.	The candidate's speech is fluid, smooth, and effortless.
Vocabulary: The candidate's ability to use "words" (including idioms or metaphors) appropriately to ask questions and/or provide explanations during the encounter.	The candidate's vocabulary is so poor as to make conversation impossible or nearly impossible.	The candidate's vocabulary limitations show a clear inability to define, clarify, or explain the concept(s) in the tested language, including using inadequate, incorrect, or insufficient vocabulary most of the time and needing to use English (or another nontested language). AND/OR The candidate makes a critical mistake that impedes communication at some	The candidate demonstrates some difficulty with vocabulary including using inadequate or incorrect vocabulary or using technical jargon or an acronym (including occasional English medical terms) and has some difficulty clarifying it or does not clarify it using other terms	The candidate occasionally uses inadequate or incorrect vocabulary or jargon or uses occasional acronyms or English medical terms but easily provides a clear explanation; or the candidate makes a mistake(s) so minor that it/they can be easily	The candidate uses a wide variety of vocabulary, including synonyms, and avoids technical jargon as needed; if jargon is used, the candidate provides a clear explanation and confirms or checks for

Category and definition	Level 1	Level 2	Level 3	Level 4	Level 5
		point during the patient encounter or significantly impacts its content as observed by the rater.	or explanation. YET This difficulty does not impact the medical content as observed by the rater nor does it impede communication with the patient at any point.	understood (e.g., adding or changing a letter in a vocabulary term).	patient understanding. If occasional acronyms or English medical terms are used, the candidate explains why English is being used and easily provides the equivalent term in the tested language and a clear explanation.
Pronunciation: The candidate's production of words, which consists of: • vocalization or articulation of sounds. • accentuation, rhythm, and intonation.	The candidate's pronunciation problems are so severe as to make speech impossible or nearly impossible to understand.	The candidate's speech is very hard for the rater to understand because of pronunciation problems. AND/OR The candidate makes a significant mistake that impedes communication or causes misunderstanding at some point during the encounter with the patient, or that significantly impacts the content as observed by the rater.	The candidate's pronunciation patterns require some concentration from the rater and/or sometimes strain the conversation with the patient.	The candidate's pronunciation is consistently intelligible. Pronunciation is clear enough, with occasional inappropriate intonation patterns that do not cause misunderstandings or strain the conversation.	The candidate's pronunciation is totally clear.
Grammar: The candidate's knowledge and 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).	The candidate's errors in grammar, sentence construction (syntactic parsing), and word order are so severe as to make speech impossible to understand.	The candidate's grammar, sentence construction, and word order errors make some sentences difficult or nearly impossible for the rater to understand; the candidate too often must rephrase and/or restrict themselves to basic patterns. AND/OR The candidate makes a significant mistake that impedes communication or causes misunderstanding at some point during the encounter with the patient.	The candidate's sentence structure is not limited to basic patterns but makes errors of grammar, word order, and sentence construction (syntactic parsing) that strain the conversation with the patient at some point, yet they can be understood by the rater.	The candidate makes grammatical and/or word order errors that do not strain the conversation with the patient.	The candidate's grammar and word order are always correct or the candidate very rarely makes a mistake, and if a mistake is made, it does not strain the conversation.
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.	The candidate builds rapport with the patient using basic greetings but lacks skills to address cultural issues, simplify complex matters, or present sensitive aspects. Limitations were so severe as to make communication impossible or nearly impossible.	The candidate builds rapport with the patient using basic greetings and encouraging questions but lacks skills to address all or most cultural issues, simplify complex matters, or present sensitive aspects. AND The candidate's limitations impede communication with the patient at some point during the encounter and/or significantly impact the content as observed by the rater.	The candidate builds rapport with the patient using basic greetings and encouraging questions and adjusts communicative register most of the time and addresses many cultural issues but lacks skills to simplify complex matters or present sensitive aspects. AND/OR The candidate's limitations strain the conversation but do not impede communication with the patient or significantly impact the content as observed by the rater.	The candidate builds rapport with the patient using basic greetings, encouraging questions, and showing empathy in an appropriate way; adjusts communicative register easily and addresses most cultural issues; and simplifies complex matters and presents sensitive aspects in an appropriate way but with occasional difficulties. Communication happens with few limitations and these limitations do not negatively affect the conversation with the patient.	The candidate builds rapport using basic greetings, encouraging questions, and showing empathy in an appropriate way; adjusts communicative register easily and addresses all cultural issues, checking for or confirming patient comprehension; and simplifies complex matters and presents sensitive aspects in an appropriate way and with no difficulties. Communication happens fully appropriately and without difficulties.

[a]The use of the POLOM requires rater training to obtain reliable scores. Use of the POLOM by raters without sufficient training is expected to result in scores with low reliability. We are currently testing and refining a POLOM rater training program.

Table 2

Iterative Revisions to the Physician Oral Language Observation Matrix (POLOM)

POLOM category	Revisions 1–7	Revisions 8–14	Revisions 15–22
Overall/across all categories	• Added and revised definitions of each POLOM category so that raters with varying medical or linguistic expertise can understand • Developed anchoring characteristics (e.g., types of errors by category) • Replaced the term "native" with observable descriptors of skills (L5) • Added a visual cue to rater by shading cells for L4 and L5 in a different color from L1, L2, and L3 • Clarified that L2 is defined by error(s) that cause miscommunication or affect the medical content of encounter (compared to L3)	• Improved consistency between the terms used to describe the errors in each category across categories and across levels (e.g., terms like "impedes communication" [L2] and "strains the conversation" [L3] are applied consistently) • Reduced focus on frequency of errors and instead focused on whether they impacted the medical communication • Created a list of factors that should prompt rating a second encounter (e.g., audio difficulties, standardized patient errors or low proficiency, candidate factors)	• Reviewed all level descriptors in each category to clarify whether a particular error should be determined "as observed by the rater" or based on the patient's reaction. The rationale for this edit was that SPs may not always react or demonstrate a lack of understanding (e.g., due to cultural factors inhibiting the patient from speaking up or because the SP contextually or due to prior experience or knowledge is able to "guess" what the candidate meant)
Comprehension	• Clarified acceptable requests for "repetition" (L5) • Changed the word "interpret" to "understand" for clarity	• Improved characterization of the observable deficits in comprehension to better distinguish candidates at each level. (e.g., L2: misunderstanding impacts content, L3: misunderstanding happens or goes unclarified but does not impact content, L4: repetition or explanation is needed but no misunderstanding happens)	• Added "acknowledgement of what the patient says/asks" as an observable element (L1, L2, L3)
Fluency/fluidity	• Quantified meaning of "frequent" and "occasional" (L3, L4) • Clarified "use of English" (L2) • Clarified distinction between fluency and vocabulary to avoid double penalization; added descriptors "to string words together," "hesitant," and "forced into English" (L2) • Changed "use of English" to "use of a non-tested language"	• Clarified description of this category due to colloquial understanding of the word "fluent" differing from its linguistic meaning • Removed error quantity and instead focused on whether errors "strained the conversation" (L3, L4) • Added examples of indicators of fluency deficits (L2)	• Improved characterization of candidates who were unable to communicate beyond the sentence level (L1)
Vocabulary	• Quantified meaning of "frequent" and "occasional" (L3, L4) • Clarified acceptable uses of "jargon" (e.g., when appropriately explained)	• Added the term "impeding communication" instead of "frustrating the conversation" (L2, L3) • Clarified descriptors of vocabulary errors due to use of a nontested language (e.g.,	• Clarified descriptors of errors to distinguish between L2 and L3 and restructured descriptors as a list for ease of rating

POLOM category	Revisions 1–7	Revisions 8–14	Revisions 15–22
	• Changed “words” to “words/lexical units” with description	inadequate words) vs. fluency errors (e.g., hesitancy when searching for the right word or using nontested language to string words together)	• Removed error quantity and instead focused on whether errors “strained the conversation” (L3) or were “minor and easily explained” (L4) • Added “insufficient vocabulary” as descriptor (L2)
Pronunciation	• Clarified distinction between “causing misunderstanding” and “straining conversation” (L2, L3) • Changed “virtually impossible” to “impossible or nearly impossible to understand” (L1)	• Removed references to “native” level since these can mistakenly equate nativity or origin (e.g., the lack of a mixed accent) with clarity of pronunciation	• Removed error quantity and instead focused on whether errors “strained the conversation” (L3, L4)
Grammar	• Added descriptors/types of errors so that nonlinguists can rate them more accurately (e.g., syntactic parsing)	• Removed error quantity and instead focused on whether errors “strained the conversation” (L3, L4)	• Clarified the descriptors “impossible” and “nearly impossible to understand” to distinguish levels (L1, L2)
Communication	• Changed from a 4-point to a 5-point scale to match other categories • Proposed descriptors of all communication levels (L1 to L5)	• Added the term “impeding communication” instead of “frustrating the conversation” (L2, L3)	• N/A

Abbreviations: L, level; N/A, not applicable, meaning no significant changes made in the given category; SP, standardized patient.

Table 3

Variance Component Estimates and Dependability Coefficients from Experienced Raters Using the Physician Oral Language Observation Matrix (POLOM) to Assess Medical Spanish Students (n = 50 students[a])

	POLOM rating variation by source, %[b]			Dependability coefficients[c]	
POLOM score	**Student**	**Rater**	**Residual**	**Single rater**	**2-rater average[d]**
Comprehension	73.8	1.2	25.0	.738	.849
Fluency/fluidity	88.1	0.7	11.2	.881	.937
Vocabulary	77.7	3.4	18.8	.777	.875
Pronunciation	82.1	1.3	16.6	.821	.902
Grammar	83.4	0.0	16.6	.834	.910
Communication	88.8	0.2	11.1	.888	.941
Total	92.6	0.7	6.8	.926	.961

[a] Each student was rated by 2 raters.

[b] Within each row, the percentages of student, rater, and residual variation sum to 100, within rounding error. The residual confounds the student-by-rater and random error sources.

[c] Possible range, 0–1.

[d] Averaged across ratings by 2 raters, including 1 physician and 1 linguist.

Table 4

Descriptive Statistics from Experienced Raters Using the Physician Oral Language Observation Matrix (POLOM) to Assess Medical Spanish Students (n = 50 students[a])

	POLOM score			
Variable	**Mean rating**	**SD**	**Min**	**Max**
Comprehension[b]	4.15	0.95	2.0	5.0
Fluency/fluidity[b]	3.30	1.07	1.5	5.0
Vocabulary[b]	3.05	1.04	1.0	5.0
Pronunciation[b]	3.62	0.78	2.5	5.0
Grammar[b]	3.28	0.96	2.0	5.0
Communication[b]	3.01	1.13	1.0	5.0
Total[c]	20.42	5.35	10.5	30.0
Total rescaled[d]	3.40	0.89	1.8	5.0

[a]Ratings for each student were averaged across 2 raters prior to computing summary statistics.

[b]Possible range, 1–5.

[c]Possible range, 6–30.

[d]Total POLOM score rescaled to have a possible range of 1–5.

Abbreviations: SD, standard deviation; min, minimum rating; max, maximum rating.