ABSTRACT

Title of Dissertation:	SIMULATED REAL-LIFE TASKS AS A TOOL TO INVESTIGATE THE EXTRAPOLATION INFERENCE OF LANGUAGE ASSESSMENTS FOR PROFESSIONAL PURPOSES
	Rosario de Fatima Montero Navarro, Doctor of Philosophy, 2025
Dissertation directed by:	Professor Kira Gor, Second Language Acquisition
	Dr. Martyn K. Clark, Second Language Acquisition

In language test validity, the extrapolation inference refers to test scores being representative of the target linguistic performance in the target context of language use. This inference is investigated through criterion-related studies that compare test scores/performance with other external evaluation criteria such as actual samples from the target domain. However, obtaining these samples can be difficult in certain professional fields like the medical one. In these cases, simulated real-life tasks could be a suitable alternative.

Following the argument-based approach to test validity (Chapelle et al., 2008), this study used a simulated real-life task specifically designed for this project to examine the extrapolation inference of the medical Spanish speaking section of the

CanopyCredential exam—a language assessment for the medical domain. The simulated task represents a doctor-patient interaction, and was developed employing the theoretical frameworks of task-based language teaching (Long, 2015), task-based language assessment (Norris, 2016), and the language proficiency interview (Ross, 2017). The medical Spanish speaking section of the *CanopyCredential* exam measures examinees' medical Spanish speaking ability to communicate with Spanish speaking patients in the U.S. medical workplace.

Upon self-assessing their medical Spanish listening and speaking ability, participants ($N = 82$) completed the exam and the simulated task in a counterbalanced order, within two weeks from each other. Their performance on each instrument was rated using linguistic-criteria rubrics. The many-facet Rasch model, ordinal regression analyses, and Spearman's correlation were used to examine the appropriateness of the simulated task as external evaluation criterion, and the extrapolation inference of the medical Spanish speaking section of the exam.

The results indicated that the simulated task was a suitable external evaluation criterion. Furthermore, significant associations were found between participants' scores in the exam and (1) their scores in the simulated task, (2) the linguistic characteristics of their production in the simulated task, and (3) their self-assessment of medical Spanish listening and speaking ability. This suggests that simulated tasks can be an appropriate instrument to examine the extrapolation inference of language assessments for professional purposes, and provide support for the extrapolation inference of the medical Spanish speaking section of the *CanopyCredential* exam.

SIMULATED REAL-LIFE TASKS AS A TOOL TO INVESTIGATE THE EXTRAPOLATION INFERENCE OF LANGUAGE ASSESSMENTS FOR PROFESSIONAL PURPOSES

by

Rosario de Fatima Montero Navarro

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2025

Advisory Committee:
Professor Kira Gor, Chair
Dr. Martyn K. Clark, Co-Chair
Professor Nan Jiang
Dr. Megan C. Masters
Professor Donald J. Bolger, Dean's Representative

© Copyright by
Rosario de Fatima Montero Navarro
2025

Dedication

To Don and Beth Van Roekel, for all their love and support.

Acknowledgements

I would have not been able to successfully complete my Ph.D. program at the University of Maryland without the help and support of so many wonderful individuals. Each one of them inspired me on a daily basis to become a better researcher and person. I owe them my accomplishment and will always keep them close to my heart.

I am extremely grateful to my dissertation co-chair, Dr. Martyn Clark, for his mentorship. Dr. Clark did all the advising for this dissertation project. He reviewed multiple drafts for the proposal and the final manuscript, helped me with the data analyses, and taught me about the Rasch model. He also checked on me periodically as I was working through different stages of the project and made himself available anytime I needed help. His kindness and work ethic always motivated me to continue doing my best and I am so thankful for everything he has taught me.

I am also immensely grateful to my dissertation chair, Dr. Kira Gor, and to all members of my dissertation committee. Dr. Gor provided insightful feedback on various drafts of the manuscript, and shared wonderful ideas for future research projects that I am looking forward to pursuing. Dr. Nan Jiang's patience and research rigor have always been a model to me. I enjoyed learning about these and so many other concepts and research approaches in his courses, by having him as my QP2 advisor, and through his thoughtful comments and feedback on my dissertation manuscript. Dr. Megan Masters and Dr. Donald J. Bolger attention to detail helped me shape my project as well as the final manuscript, and I am thankful and fortunate to have had them in my dissertation committee.

This project would have not been possible without the generosity of Canopy Innovations and its wonderful team. I am deeply grateful to Bill Tan, Caroline Stone, Dr. Kate Riestenberg, and Dr. Payman Vafaee for allowing me to use their *CanopyCredential* exam as one of my research

instruments, and for all the help and support they provided. All of them went above and beyond to assist me with my project and I hope the results from this study are helpful to them as well.

I am deeply grateful to the experts who assisted me in the development of the simulated-real life task and the rating rubric that I used in this project. Dr. Brandon Berger, Dr. Evelyn Canabal-Torres, Dr. Katie Brown, Dr. Lourdes Ortega, Mariana Alvarez Torres, and Mary Beth Van Roekel assisted me in every step of this process and helped me recruit study participants as well. I am also grateful to the team of raters who evaluated participants' performance in the simulated real-life task. Dr. Cecilia Battauz, Dr. Daniela Hernández Rodríguez, and Dr. Victoria Herrera Arvay took time out of their busy schedules to help me with this.

My deepest gratitude to all the volunteers and study participants who put time aside to complete all my study tasks within the required time frame, and to all those who helped me with the recruitment process. Their stories, commitment to their work, and unwavering willingness to help was moving and inspiring. I learned many things from each one of them and have been able to better understand and have even more respect than I already had towards the hard work that all healthcare professionals do.

I would also like to thank the entire team of the Office of Dean of the College of Arts and Humanities for all their help and support. It was an honor to work with them for several years and I am grateful for everything that I have learned from them as well. They not only provided the support that I needed to complete my studies, but also helped me in moments of serious difficulties. I am especially thankful to Sherita Huff and Angela Smith for their love and friendship. It has been a pleasure to work in tandem with them during my time at the office.

I am equally thankful to the entire team of the Office of the Director of the School of Languages, Literatures, and Cultures where I also had the honor to work. My deepest gratitude to

Dr. Mary Ellen Scullen and the late Dr. Lauretta Clough for their kindness and wise advice. I am also thankful to the administrative team of the Department of Spanish and Portuguese, where I had the opportunity to teach Spanish and collaborate with their talented professors for several years. A special thank you to Dr. Elisa Gironzetti and Dr. Manel Lacorte for all their teachings, and for helping and supporting me in so many different ways.

A heartfelt thank you to all my professors and classmates in the Second Language Acquisition Program. I enjoyed every moment with them and deeply value everything I learned with and from them. I am especially grateful to Dr. Mike Long and Dr. Steve Ross whose teachings and ideas guided me even when they were not among us any longer. This dissertation is also dedicated to them and to their memory.

I was also fortunate to receive the help and support of many friends in the DMV area and beyond during these wonderful years of studying and learning. I am thankful to my Georgetown University friends Dr. Brandon Tullock, Christina Butler, Dr. Julia Goetze, Dr. Lara Bryfonski, Dr. Sakol Suethanapornkul, and Dr. Young-A Son. I am also grateful to my friends at the University of Maryland, Dr. Alia Lancaster, Bradford Salen, Dr. Eric Pelzl, Eunsoo Kang, Jackson Monzón, Dr. Jonathan Malone, Dr. Jiyong Lee, Dr. Kichan Park, Dr. Kyoko Kobayashi Hillman, Mireia Toda Cosi, Dr. Nick Pandža, Dr. Nur Başak Karataş, Dr. Qi Zheng, Dr. Ryo Maie, Sanshiroh Ogawa, Dr. Stephen O'Connell, Tetiana Tytko, Yoonjee Hong, Dr. Yi Wei, and Zhiyi Wu. I am also thankful to my dear friends Andy Hill, Deja Williams, Feibe Boyle, Kelly Livingston, Laura Vergés Oller, Lauren Campbell, Dr. Mariana Reyes, Matthew Fletcher, Dr. Nélida Devesa Gómez, and Dr. Sarah Dowman.

Last but not least, I would like to thank my parents Aurelio and Fátima, and my sister Teresa for teaching me about life. This dissertation is dedicated to you as well.

This dissertation was supported by the 2025 TOEFL® Grant for Doctoral Research in Language Assessment.

Table of Contents

Dedication ii
Acknowledgements iii
Table of Contents vii
List of Tables ix
List of Figures x
Chapter 1: Introduction 1
1.1 Significance of this study 4
Chapter 2: Literature review 7
2.1 Test users' perceptions about language assessments for professional purposes 7
2.2 Validity research in language testing 14
2.3 The argument-based approach to test validity 17
2.4 Use of the argument-based approach to investigate the extrapolation inference of language assessments 23
2.5 Use of simulations in learning and assessment 29
2.6 Context of this study 37
Chapter 3: Methodology 48
3.1 Participants 49
3.2 Instruments 51
3.2.1 The *CanopyCredential* exam 51
3.2.2. The simulated real-life task 54
3.3 Procedures 66
3.4 Data analyses 75
3.4.1 Analyses conducted to examine the simulated real-life task's functionality 75
3.4.2 Analyses conducted to examine the extrapolation inference of the *CanopyCredential* exam 80
Chapter 4: Results 83
4.1 Suitability of the SRLT as external evaluation criterion 83
4.2 The extrapolation inference of the CanopyCredential exam 90
4.2.1 Relationship between participants' scores in the CCE and the SRLT 90
4.2.2 Relationship between participants' scores in the CCE and linguistic measures of intelligibility 93
4.2.3 Relationship between participants' scores in the CCE and the linguistic measure of fluency 97
4.2.4 Relationship between participants' scores in the CCE and linguistic measures of grammar 99
4.2.5 Relationship between participants' scores in the CCE and the linguistic measure of vocabulary 108
4.2.6 Relationship between participants' scores in the CCE and linguistic measures of coherence 110
4.2.7 Relationship between participants' scores in the CCE and linguistic measures of effectiveness 113
4.2.8 Relationship between participants' scores in the CCE and their self-assessment of medical Spanish listening and speaking ability 117

Chapter 5: Discussion 120
5.1 RQ1: Do the SRLT facets (i.e., participant's ability, rating criteria, rating scale, and rater severity) function appropriately? 122
5.2 RQ2: Do the SRLT raters and linguistic scoring criteria reliably differentiate participants' proficiency level? 123
5.4 RQ3: To what extent are participants' scores in the speaking section of the Spanish module of the CCE related to their scores in the SRLT? 125
5.5 RQ4: To what extent are participants' scores in the speaking section of the Spanish module of the CCE related to their production in the SRLT on linguistic measures of intelligibility, fluency, grammar, vocabulary, coherence, and effectiveness? 126
5.6 RQ5: To what extent are participants' scores in the speaking section of the Spanish module of the CCE related to their self-assessment of medical Spanish speaking ability? 135
Chapter 6: Conclusion 137
Appendices 141
References 150

List of Tables

Table 1 *Interpretation/Use Argument for the CanopyCredential Exam* 40
Table 2 *Participants by Self-Assessment of Medical Spanish Listening and Speaking Ability Options* 50
Table 3 *Characteristics of Tasks in the Speaking Section of the Spanish Module of the CCE* 54
Table 4 *Correspondence between the LPI and the Study's SRLT* 60
Table 5 *Scoring Rubric for the SRLT* 62
Table 6 *Study's Linguistic Criteria/Measures* 74
Table 7 *Descriptive Statistics for the Linguistic Scoring Criteria* 84
Table 8 *Linguistic Scoring Criteria Difficulty Measures (logits), Fair Average Measures, and Fit Statistics Estimated Based on the Raters' Ratings* 87
Table 9 *Scoring Rubric Scale Functionality* 89
Table 10 *Participants by CCE Medical Spanish Speaking Proficiency Level* 90
Table 11 *Descriptive Statistics for Participants' Scores (log-odds) in the SRLT* 91
Table 12 *Results Summary for the Relationship between Participants Medical Spanish Speaking Scores in the CCE and the SRLT* 92
Table 13 *Descriptive Statistics for the Intelligibility Criterion's Linguistic Measures* 94
Table 14 *Results Summary for the Relationship between Linguistic Measures of Intelligibility and CCE Scores* 97
Table 15 *Descriptive Statistics for the Fluency Criterion's Linguistic Measure* 98
Table 16 *Results Summary for the Relationship between the Linguistic Measure of Fluency and CCE Scores* 99
Table 17 *Descriptive Statistics for the Grammar Criterion's Linguistic Complexity Measures* 100
Table 18 *Results Summary for the Relationship between Linguistics Measures of Grammar Complexity and CCE Scores* 104
Table 19 *Descriptive Statistics for the Grammar Criterion's Linguistic Accuracy Measures* ... 105
Table 20 *Results Summary for the Relationship between Linguistic Measures of Grammar Accuracy and CCE Scores* 108
Table 21 *Descriptive Statistics for the Vocabulary Criterion's Linguistic Measure* 108
Table 22 *Results Summary for the Relationship between the Linguistic Measure of Vocabulary and CCE Scores* 109
Table 23 *Descriptive Statistics for the Coherence Criterion's Linguistic Measures* 110
Table 24 *Results Summary for the Relationship between Linguistic Measures of Coherence and CCE Scores* 112
Table 25 *Descriptive Statistics for the Effectiveness Criterion's Linguistic Measures* 114
Table 26 *Results Summary for the Relationship between Linguistic Measures of Effectiveness and CCE Scores* 117
Table 27 *Participants Distribution across CCE Score Categories and Self-Assessment Categories* 118

List of Figures

Figure 1 *Interpretive Argument for the TOEFL iBT* 20
Figure 2 *Argument-Based Validity Structure Model* 21
Figure 3 *Interpretation/Use Argument Structure for the CanopyCredential Exam* 39
Figure 4 *Simulated Real-Life Task Prompt* 57
Figure 5 *Guideline for the SRLT's Simulated Patient* 58
Figure 6 *Example of a Scored SRLT Performance* 62
Figure 7 *Qualtrics Template used to Rate Participants' SRLT Performance* 70
Figure 8 *Wright Map Displaying the Location of the Elements in Each Facet* 86
Figure 9 *Category Probability Curves for the Scoring Rubric Scale* 89
Figure 10 *Relationship between Participants Medical Spanish Speaking Scores in the CCE and the SRLT* 91
Figure 11 *Relationship between Pronunciation Errors (Phoneme) per AS-Unit and CCE Scores* 94
Figure 12 *Relationship between Pronunciation Errors (Syllable) per AS-unit and CCE Scores* .95
Figure 13 *Relationship between Pronunciation Errors (Word Stress) per AS-unit and CCE Scores* 96
Figure 14 *Relationship between Pauses per AS-unit and CCE Scores* 98
Figure 15 *Relationship between Clauses per AS-unit and CCE Scores* 101
Figure 16 *Relationship between Clitic Pronouns per AS-unit and CCE Scores* 102
Figure 17 *Relationship between Subjunctive Tense Forms per AS-unit and CCE Scores* 102
Figure 18 *Relationship between Lexical Errors per AS-unit and CCE Scores* 105
Figure 19 *Relationship between Morphological Errors per AS-unit and CCE Scores* 106
Figure 20 *Relationship between Syntactic Errors per AS-unit and CCE Scores* 106
Figure 21 *Relationship between Medical Terms per AS-unit and CCE Scores* 109
Figure 22 *Relationship between Coordinating Conjunctions per AS-unit and CCE Scores* 111
Figure 23 *Relationship between Subordinating Conjunctions per AS-unit and CCE Scores* 111
Figure 24 *Relationship between Explained Medical Terms by Definition per AS-Unit and CCE Scores* 114
Figure 25 *Relationship between Explained Medical Terms by Synonym per AS-unit and CCE Scores* 115
Figure 26 *Relationship between Explained Medical Terms by Context per AS-unit and CCE Scores* 115
Figure 27 *Relationship between Participants' Self-Assessment Categories and CCE Scores* 119
Figure 28 *Structure Model for the Extrapolation Inference of the Speaking Section of the Spanish Module of the CCE* 121
Figure 29 *Predicted Probabilities of Achieving a Specific CCE Score Category based on SRLT Scores* 126
Figure 30 *Predicted Probabilities of Achieving a Specific CCE Score Category based on Production of Phoneme Pronunciation Errors per AS-unit* 128
Figure 31 *Predicted Probabilities of Achieving a Specific CCE Score Category based on Production of Pauses per AS-unit* 129
Figure 32 *Predicted Probabilities of Achieving a Specific CCE Score Category based on Production of Subjunctive Tense Forms per AS-unit* 130

Figure 33 *Predicted Probabilities of Achieving a Specific CCE Score Category based on Production of Lexical Errors per AS-unit* 131
Figure 34 *Predicted Probabilities of Achieving a Specific CCE Score Category based on Production of Explained Medical Terms by Synonym per AS-unit* 133

Chapter 1: Introduction

Multilingual needs have existed in certain professional fields for a long time (e.g., Hohenstein & Lévy-Tödter, 2020), have appeared in new ones due to demographic shifts, and can be expected to expand and increase as a product of globalization (ACTFL, 2019). This has intensified the demand for individuals able to encode their professional expertise in additional languages (Ln) during the last decades, which prompted the development and use of language assessments for professional purposes (LAPP). Important decisions such as employment or immigration eligibility are made based on the scores that examinees receive in these assessments. These and other uses of LAPP are often grounded in a series of beliefs and assumptions that test users have about these assessments. The narratives underlying these beliefs vary among stakeholders' groups and individuals, and can influence their test scores interpretations and uses, and the amount of trust that they place on these tests.

For instance, in the assessment of speaking proficiency, previous research has shown that examinees perceive computer-delivered test tasks that use pre-recorded videos and prompts instead of face-to-face human interlocutors in real time as less representative of real life communication (Malone & Montee, 2014), and prefer the latter format over the former one (Nguyen et al., 2024; Qian, 2009). However, as Jung Youn (2023) reports, emerging technologies are increasingly replacing human interlocutors in this type of assessments and, as Ross (2017) predicts, this trend is likely to continue growing. To assist test users with test scores interpretations and uses, and examine the extent to which the scores from computer-based exams are representative of real life communication with human interlocutors, it is necessary to conduct test validity studies that use instruments including the features that are being replaced by emerging technologies.

In language test validity research, the argument-based approach to test validity (Chapelle et al., 2008; Kane, 1992, 2001, 2006, 2013; Messick, 1989) specifically focuses on the investigation of test scores interpretations and uses. Within this theoretical framework, the extrapolation inference refers to the relationship between test scores and the target linguistic performance in the target context of language use. This inference is investigated through criterion-related studies that compare test scores/performance with other external evaluation criteria such as actual samples from the target context of language use (e.g., Brooks & Swain, 2014; Llosa & Malone, 2019). However, these samples can be difficult to obtain in certain professional fields like the medical or legal ones due to privacy concerns, among other reasons. Thus, in cases where computer-based test tasks could be perceived as a more distant representation of those from the target context of language use, and actual samples from the target context are unattainable, human-based simulated real-life tasks (SRLT) that are representative of those from the target context could be a suitable external evaluation criterion to examine the extrapolation inference of these exams.

SRLT have been used for learning and assessment purposes for decades. Particularly, in the medical field, SRLT have been employed since the 1970s and are nowadays one of the most commonly used instruments to evaluate and practice a variety of clinical knowledge and skills (Cuddy et al., 2022; Harden et al., 1975; Komasawa et al., 2020). Additionally, SRLT have also been used for language learning and assessment, and previous research has shown its positive influence on a variety of learning outcomes (Bryfonski & McKay, 2019). Therefore, since SRLT are a common tool in the medical field, and research has shown their usefulness for language learning and assessment, it is worthwhile to explore their suitability to investigate the extrapolation inference of the speaking section of a computer-based language assessment for the medical field. To the best of the researcher's knowledge, no studies have yet compared oral performance in a

language test for the medical domain using computer-mediated test tasks with oral performance from the same participants in a human-based SRLT. It is the lack of such research that this dissertation project aimed to address.

Following this idea, the present dissertation project used a SRLT expressly designed for this study and representing a medical follow-up phone conversation between a doctor and a patient using human interlocutors in real time, to investigate the extrapolation inference of the speaking section of the Spanish module of the *CanopyCredential* exam (CCE) using the argument-based approach to test validity. The CCE is a standardized scenario-based bilingual language assessment developed for the U.S. medical workplace to evaluate healthcare professionals' listening and speaking abilities to communicate with patients in English and Spanish. The speaking section of the Spanish module uses tasks with pre-recorded videos of simulated patients and written prompts. These are organized in a multi-turn sequence where examinees are asked to record their responses after watching each video and reading each prompt. One of the assumptions underlying these monologic, computer-mediated tasks is that examinees' performance in this section/module of the CCE generalize to actual dialogic interactions in the target context of language use. This is the core of the extrapolation inference that this research project sought to investigate through a SRLT specifically designed for this purpose.

Participants in this study ($N = 82$) were healthcare professionals/students from the U.S. and Canada, who had familiarity with the U.S. medical context. Upon filling out a bio-data survey including a self-assessment of medical Spanish listening and speaking ability, they completed the SRLT and the Spanish module of the CCE in a randomly assigned counterbalanced order, within two weeks from each other. Their performance in the speaking section of the Spanish module of the CCE was evaluated by a team of certified raters from the test provider. Participants

performance in the SRLT was evaluated by a team of three raters specifically trained to assist with this task, using a modified version of the linguistic-criteria rubric used in the CCE. After exploring the appropriateness of the SRLT to measure the target construct of medical Spanish speaking ability for the U.S. medical workplace, this study compared participants scores on the speaking section of the Spanish module of the CCE with their overall scores on the SRLT, the linguistic characteristics of their oral production on the SRLT, and their self-assessment of medical Spanish listening and speaking ability. The research questions that guided this study were the following:

- RQ1: Do the SRLT facets (i.e., participant's ability, rating criteria, rating scale, and rater severity) function appropriately?
- RQ2: Do the SRLT raters and linguistic scoring criteria reliably differentiate participants' proficiency level?
- RQ3: To what extent are participants' scores in the speaking section of the Spanish module of the CCE related to their scores on the SRLT?
- RQ4: To what extent are participants' scores in the speaking section of the Spanish module of the CCE related to their production in the SRLT on linguistic measures of intelligibility, fluency, grammar, vocabulary, coherence, and effectiveness?
- RQ5: To what extent are participants' scores in the speaking section of the Spanish module of the CCE related to their self-assessment of medical Spanish listening and speaking ability?

1.1 Significance of this study

The main goal of this study was to investigate the suitability of SRLT to investigate the extrapolation inference of LAPP, using the speaking section of a computer-based language assessment for the medical domain as an example. This research topic is particularly important as

emerging technologies are increasingly being used in test development and delivery, and are replacing human components from traditional test formats such as those using human interlocutors in real time for the assessment of speaking ability (Jung Youn, 2023). In these instances, SRLT that include the elements being replaced by emergent technologies could be used in test validity studies investigating the extrapolation inference of these exams. Additionally, SRLT could also be an alternative for situations in which collecting actual samples from the target domain of language use to investigate this inference is not feasible. This could be the case of other LAPP for the medical domain or for the legal domain, among others. There are also other advantages from using SRLT in language test validity studies investigating the extrapolation inference.

As Chapelle and Voss (2021) maintain, the validation research of a language test is an ongoing process where the more evidence that can be gathered to assist in the interpretations and uses of test scores, the better. As such, SRLT could be used not just as an alternative but as an additional evaluation criterion. That is, the extrapolation inference of the same test could be examined through different studies using different external evaluation criteria such as SRLT, linguistic corpora, self-assessments, or even actual samples from the target context. An advantage of SRLT over actual samples from the target context is that they can be designed to represent cases that might not occur as often in the target context and would be therefore difficult to obtain. For example, in the U.S. medical workplace, certain Spanish speaking varieties might be less common than others. Still, test users would need to know the relationship between examinees' scores in an exam like the CCE and their ability to communicate with patients who speak a variety of Spanish that is less common in the U.S. medical workplace. In these cases, SRLT could be helpful additional external evaluation criteria.

Furthermore, beyond the influence that this research project could have for the field of language test validity research, this project could also be beneficial for the field of second language acquisition (SLA). The development process and analyses that were used for this study's SRLT could also be used for the design and evaluation of SRLT intended for language learning and assessment. Moreover, the way participants' performances were evaluated in this study's SRLT could also be helpful for practitioners interested in the relationship between task completion and the linguistic features associated with it.

Given that this study used a language assessment developed for the U.S. medical workplace where language barriers are an issue that have impacted negatively many people, the results from this study could also be helpful to assist in the ongoing efforts to reduce these barriers and the impact they have on individuals.

Chapter 2: Literature review

The investigation of the extrapolation inference of a LAPP requires a review of the beliefs and perceptions that test users' have about these assessments, and how they could influence this inference. Next, the adoption of an appropriate framework following best practices within the field of language test validity research is needed, based on similar studies investigating the same inference. Finally, the selection or development of a suitable external evaluation criterion that fits the specific characteristics and context of the LAPP under investigation is necessary. The next sections of this chapter describe these aspects and constitute the argument that underlie the methodology that was used in this research project.

2.1 Test users' perceptions about language assessments for professional purposes

As Macqueen et al. (2021) highlight, language assessments have become familiar and trusted tools that offer welcome solutions to the difficult task of predicting the dynamic behavior of individuals moving from one professional jurisdiction to another, or seeking employment in multilingual contexts (p. 8). This trust is based on a series of beliefs and perceptions that might differ across test users but that are important to consider when investigating the extrapolation inference and all other inferences made about a LAPP.

For example, Macqueen et al. (2021) examined data from semi-structured interviews with 11 representatives of six Australian professional bodies that used different standardized LAPP with their candidates, and found that these representatives and their organizations largely believed the test scores' meanings supplied by the test providers, and developed professional registration policies based on them. This was done in response to the need "to manage large numbers of applicants, to assure a level of English proficiency for high-risk professional communication, to

provide an objective assessment that is separate from any assessment of professional competence and to maintain consistency of standards across bodies" (Macqueen et al., 2021, p. 12). The professional bodies represented were accounting, engineering, medical, and nursing.

In a similar study, Macqueen et al. (2016) used semi-structured interviews to investigate the perspectives of three different groups of test users about the Occupational English Test (OET), a language assessment designed to assess the English communication skills of healthcare professionals seeking to register and practice in English-speaking environments (see https://oet.com/test/test-overview). They focused on the relevance of the test content with the workplace, and the language demands of the workplace for users of English as an additional language (EAL). The participants were senior representatives from two medical professional bodies (n = 2), supervisors of EAL health professionals (n = 11), and successful test takers who were working in their health professions at the time of the study (n = 18).

The findings of Macqueen et al. suggest that whereas senior representatives and EAL supervisors viewed the test as a measure of general English proficiency, successful test takers also saw it as a measure of clinical communication skills. Moreover, members of this participants' group preferred the OET over the International English Language Testing System (IELTS)—designed to measure general English communication skills—because they felt that the test content was more relevant to their profession and complemented that of other professional exams that they had to take for professional registration purposes. Additionally, the test takers expressed that preparing for the exam helped them develop knowledge and skills such as note-taking, cultural awareness, professional expectations, and use of lay terms that were useful later on to address the language demands that they faced in their professional practice (Macqueen et al., 2016, p. 285).

In the academic area, the Test of English as a Foreign Language (TOEFL) is used by many English-medium universities with international students coming from institutions where English is not the main language of instruction. This test is designed to measure the English communication skills for the classroom, and the most prominent current version is the TOEFL iBT which was introduced in 2005 (Educational Testing Service [ETS], n.d.). About ten years after this version of the test became operational, Malone and Montee (2014) conducted a study to explore stakeholders' beliefs about the exam as a measure of academic language ability. They used focus groups, surveys, and stimulated recalls with three different groups of participants (i.e., university administrators [focus groups n = 7, surveys n = 246], instructors [focus groups n = 22, surveys n = 289], and students [focus groups n = 29, surveys n = 788, stimulated recalls n = 12). All the participants had demonstrated familiarity with the exam.

The results from Malone and Montee's study revealed that the instructors considered that the test was mainly a good measure of English communication skills, and not an accurate predictor of how well examinees would perform at an English-speaking university. In addition, this participants' group felt that preparing to take the TOEFL iBT did not prepare students for life in an English-speaking university. However, the instructors' group agreed with the students' group in that a series of academic tasks and skills were represented and used in the exam. Furthermore, instructors and students also agreed that the test's integrated tasks requiring the application of multiple skills (e.g., reading a passage and listening to a short lecture, and then produce an oral or written response) were a good measure of academic language ability. The university administrators' group agreed that the test was a good measure and a predictor of students' ability to effectively use English in academic courses, and test performance reflected performance in academic classes.

An interesting finding from Malone and Montee's (2014) study that was of relevance for this research project refers to the students' views about how the TOEFL iBT measures speaking ability. This participants' group believed that the TOEFL iBT did not measure well their ability to speak English, and attributed this to the computer-based format. Specifically, many students mentioned that although the type of speaking tasks included in the test reflected those that they encountered at the university level, speaking to a computer rather than to a human, and the timed planning and responses were sources of task difficulty. Several students expressed that one of the main difficulties was the computer lack of response or feedback, and some of them believed that the time pressure negatively affected their performance and scores (Malone & Montee, 2014, pp. 30-32). This suggests that test format could influence stakeholders' beliefs and perceptions about a test.

The use of computer-based formats in language tests is particularly important to consider as emerging technologies such as artificial intelligence (AI) are being used for language assessment. These new technologies are increasingly automatizing tasks that were traditionally performed by humans. For example, in the case of interactive speaking assessments, Jung Youn (2023) reports that spoken dialog systems (SDS) and intelligent personal assistants (IPA) are replacing human interlocutors in commonly used language test formats such as role-plays, oral proficiency interviews (OPI), and paired/group speaking tasks. Jung Youn claims that whereas this offers many new opportunities, it is important to identify the features of interactive speaking that are elicited, and conduct validity research to assist in decisions related to the intended uses of these new test formats (Jung Youn, 2023, p. 54). Thus, investigating the relationship between computer-mediated and human-based versions of the same tasks—which is part of the extrapolation inference in this research project—is a useful area of research for language test validity studies.

Besides researchers' emphasis on the importance of investigating test takers' production on automated language test formats, the issue of test users' perceptions of these formats has also been raised in the literature. Keeping the focus on the assessment of speaking ability, Ross (2017) reviewed four different types of test formats used to measure speaking proficiency in LAPP and other kinds of language assessments. These were face-to-face direct speaking tests with human interlocutors in real time; semi-direct speaking tests with pre-recorded questions; indirect speaking tests with computer-delivered tasks based on reading passages, short lectures, or visual prompts; and semi-indirect speaking tests such as the elicited imitation task which requires candidates to listen to progressively more complex utterances and immediately imitate each one. Ross claimed that the human interlocutors in direct speaking test tasks made them more closely related to real-life tasks, and would be the best way to assess speaking proficiency. Nevertheless, he also recognized that factors such as technological advances, cost-effectiveness issues, or the increased number and diversity of examinees influence the development of language assessments and the type of test formats used, and encouraged the development and use of test validity studies to investigate this.

Qian (2009) explored the affective effects on test takers of the direct and semi-direct modes of speaking assessment with a group of 186 participants. For the direct mode, participants completed the speaking component of the IELTS which was conducted in the form of a one-to-one interview with a live IELTS examiner. The semi-direct mode was represented by the speaking component of the Graduating Students' Language Proficiency Assessment-English (GSLPA)—a performance-based English proficiency test developed at the Hong Kong Polytechnic University—and was administered in multi-media language laboratories. After completion of both tests,

participants responded to a semi-structured questionnaire designed to elicit participants' views on the two testing modes under study.

The results from the questionnaire used in Qian's study showed that the majority of the participants (57.6%) did not have a particular preference in terms of testing mode. However, among those who did have a preference, the number of those favoring the direct mode (32.8%) greatly exceeded that of those in favor of the semi-direct mode (9.7%). Arguments in favor of the direct mode reflected the belief that the students could produce a better English using this format, and the feeling that the semi-direct mode tested skills unrelated to English speaking ability. On the other hand, some views favoring the semi-direct mode stated that this method was more objective and fairer to all students, and the demand for an immediate response could be a good measure of test takers' speaking ability (Qian, 2009, pp. 121-122).

In a more recent study, Nguyen et al. (2024) compared test takers' scores and perceptions on the face-to-face and computer-delivered modes of two English speaking proficiency tests for Vietnamese English as a foreign language (EFL) learners. The tests used were the level 2 and level 3-5 of the Vietnamese Standardized Test of English Proficiency (VSTEP.2 and VSTEP.3-5, respectively). Two different participants' groups completed the face-to-face and computer-delivered formats of the VSTEP.2 (n = 75), and VSTEP.3-5 (n = 82) speaking tests in a counterbalanced order, and their performances were rated by a group of 16 certified VSTEP raters. Immediately after test completion, 15 participants from each group took part in focus group interviews with 3-4 students per group.

The results from Nguyen et al.'s study showed that test scores were comparable, with only one measure of speaking proficiency on each test being statistically significantly different (i.e., content development was better in the computer-delivered mode in the VSTEP.2, and

pronunciation was better in the face-to-face mode in the VSTEP.3-5). However, data from the interviews suggested that the majority of participants preferred the face-to-face mode based mainly on the idea that speaking in a computer-delivered context does not characterize real-life communication (Nguyen et al., 2024, p. 18). Mixed perceptions were found for each testing mode though. Whereas some participants felt that human interlocutors guided them through task completion and their non-verbal feedback was enabling, others felt that this was constraining and preferred the possibility of reading the computer-delivered task prompt several times, which allowed them to better prepare their responses. Some participants also perceived the semi-direct mode as more objective and fairer (Nguyen et al., 2024, pp. 18-23).

The articles described above depict some of the many different perceptions and assumptions that test users can have about a LAPP. These can be rooted in the information supplied by test providers, as was the case in Macqueen et al. (2021), or in personal experiences with the test itself, as mentioned in Macqueen et al. (2016), and Malone and Montee (2014). Moreover, test content and format can also have an influence. As was shown in Macqueen et al. (2016), the relevance of the test content to the target domain made test takers prefer one test over the other (i.e., OET vs. IELTS). Regarding test format, real-life representativeness—which could be affected by the use of emerging technologies—seems to be perceived more positively. This appears to be the case in the assessment of speaking ability with face-to-face formats being preferred over computer-delivered formats even when test results are comparable, as was shown in Nguyen et al. (2024).

Because of the influence that these perceptions and assumptions can have on the amount of trust placed on a LAPP and most importantly, the consequences that decisions based on test scores interpretations and uses can have, it is important that test users' perceptions and assumptions

are taken into consideration by researchers and test developers. This can be done through the design of specific language test validity studies and instruments that address these matters. The next section includes a review of how these issues have been previously approached in the field of language test validity research, and how the field has evolved towards the argument-based methodology. This methodology allows the investigation of all the inferences and assumptions made about a test through multiple linked studies. The evidence or lack thereof found in these studies can assist in the interpretations and uses that people make of LAPP test scores.

2.2 Validity research in language testing

In the field of language testing, validity research has evolved from an initial orientation towards the test itself and the investigation of multiple validities (e.g., Lado, 1961), to a unitary concept that focuses on test score interpretations and uses. This transition was influenced by Messick's (1989) definition of test validity, and the ensuing work of researchers like Toulmin (2003), Kane (1992, 2001, 2006, 2013), and Chapelle et al. (2008), which resulted in the development of the argument-based approach to test validity. This method has become the most used approach in language test validity research in recent years. The next paragraphs summarize the evolution of the initial approach to language test validity, and the development of the argument-based methodology.

As Chapelle and Voss (2013) illustrate, the original approach to test validity in language testing considered validity an inherent property of the test. It sought to answer if the test was measuring what it was supposed to through the investigation of different types of validity. Lado (1961) described the steps and types of validity that researchers should follow when investigating and reporting the validity of a test. He focused on assembling evidence that would support each validity type. As such, content validity would investigate if the test was measuring the language

element or skill that it was purported to measure through linguistic analyses. These analyses would be conducted to find out the linguistic elements that would be more difficult to a particular group of students. Then, tests designed to measure these elements would be examined to see if they were actually measuring them. This process would include the investigation of extraneous factors that could influence examinees' test performance, and the appropriateness of the techniques used to test the content. Criterion-related validity would be used for empirical validation, and it involved comparing scores on the test with other valid criteria. Lado suggested reporting the data obtained from these analyses using correlation coefficients and/or expectancy tables.

To some degree, it could be said that Lado's guideline served as a benchmark to the subsequent approaches to test validity in language testing as traces of it can be seen in them. The idea of evidence gathering was particularly influential and has been continued in approaches that followed. However, the focus of Lado's initial approach on how well the test was doing the job that it was supposed to do neglected the social consequences of test scores interpretations and uses, as Messick (1989) pointed out. In his chapter in *Educational Measurement,* Messick switched the focus to the score interpretation aspect of tests, which was embraced by the field of language testing.

Messick defined validity as "an integrated judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" (Messick, 1989, p. 13, emphasis in original). He considered that the enumeration of different types of validity could perpetuate the temptation to rely on one only or any one of them as sufficient evidence for the validity of a particular test use. His proposed validity framework was thus based on the combination of validity

evidence without over-reliance on any single type of validity, as in Lado's approach, and the inclusion of the implications and social consequences of test scores interpretations and uses.

Following Messick's proposal, several authors in the field of language testing began introducing these concepts and guiding its application in language testing validity research. For example, Bachman (1990) was one of the first looking at validity as a unitary concept and an ongoing process of evidence gathering. He justified this approach through a review of Messick's work and the different types of validity previously investigated in the field. He also presented statistical procedures that could be used for this research such as factor analysis and multitrait-multimethod (MTMM) designs. Through this work, Bachman introduced to the field the theories that would serve as the base to the argument-based approach. However, it did not provide a sufficiently detailed guideline that researchers could follow to conduct language test validity research.

In the development of the test usefulness approach, Bachman and Palmer (1996) built on Bachman's (1990) previous work and presented a guideline that would assist researchers in the process of evidence gathering in test validity studies. For Bachman and Palmer, the most important consideration in the design and development of a language test was its usefulness. This notion—which relates to test scores interpretations and uses—was comprised of a series of qualities such as reliability, construct validity, authenticity, interactiveness, impact, and practicality. To investigate the usefulness of a test, researchers should seek answers to a series of questions for each one of these qualities, indicating the extent to which the quality was satisfied and how. This gave an idea of the types and amount of evidence that would be needed to support each test usefulness quality, and is one of the main concepts of the argument-based framework.

The principles found in Bachman (1990) and Bachman and Palmer (1996) are reorganized in the argument-based approach to test validity (Kane, 1992, 2001, 2006, 2013). This method begins with the development of an interpretive argument for a test that identifies all claims made about the test scores. Then, the inferences and assumptions underlying each claim are stated and investigated through multiple studies using the interpretive argument as guide so a validity argument for the test can be developed. By doing this, the test validity process becomes a unitary and multi-faceted one requiring the integration of multiple sources of evidence within the same argument. The next section summarizes the main details of this approach, as it is the approach that was used in this study.

2.3 The argument-based approach to test validity

As mentioned above, the argument-based approach to test validity is based on the theories stated by Messick (1989), and the work by Toulmin (2003), and Kane (1992, 2001, 2006, 2013). It uses a basic model structure to investigate each one of the claims identified in the test's interpretive argument. This argument serves as the sequential structure to investigate the claims in a consecutive order, and build the validity argument. Since 2008, multiple studies have used this model to investigate a variety of language assessments, and it has been the method of choice in the majority of most recently published language test validity studies (Dursun & Li, 2021).

The basic structure of the argument-based model follows the sequence data → inference → conclusion. This is based on the philosophical principles about the uses of argument developed by Toulmin (2003). According to this author and as Will (1960) illustrates, in everyday life, we reach conclusions or make claims based on specific data. For example, one might assert that it will start raining soon (i.e., conclusion) because there are dark clouds in the sky (i.e., data). This transition from the data to the conclusion is authorized implicitly by the principle or warrant (i.e.,

inference) that dark clouds may bring rain. To support this warrant, one would need further information or backings for the warrant. This can come from a variety of sources (e.g., memories from previous days when the sky looked similar, changes in the temperature, wind gusts, and so on).

Whereas arguments like the ones used in mathematical sciences where the data and the backing entail the conclusion have been more prevalent (e.g., 4 times 2 is 8), Toulmin argues that these are rarely applicable to other domains of life. Instead, he advocates for the use of arguments analogous to those employed in the field of law, and that follow the sequence previously described: data → inference → conclusion. These can be applied to a wide variety of situations and domains such as test validity research, and involve the collection of evidence that would provide information to support or refute the inference and assist in the interpretation of the conclusion.

Kane followed Messick's view on test validity and its focus on test scores interpretations and uses (Messick, 1989), and Toulmin's argument model (Toulmin, 2003) to develop the argument-based approach to test validity (Kane, 1992, 2001, 2006, 2013). This approach is based on the idea that to validate the interpretations and uses of test scores is to evaluate the plausibility of the claims based on the scores. Thus, the model proposes that the claims based on the test scores should be outlined in an interpretive argument where the inferences and supporting assumptions for each claim are specified. The validation process would involve the evaluation of the coherence and completeness of this argument, and the plausibility of its inferences and assumptions (Kane, 2013, p. 1). To investigate the plausibility of the inferences and assumptions, Toulmin's argument model structure should be employed.

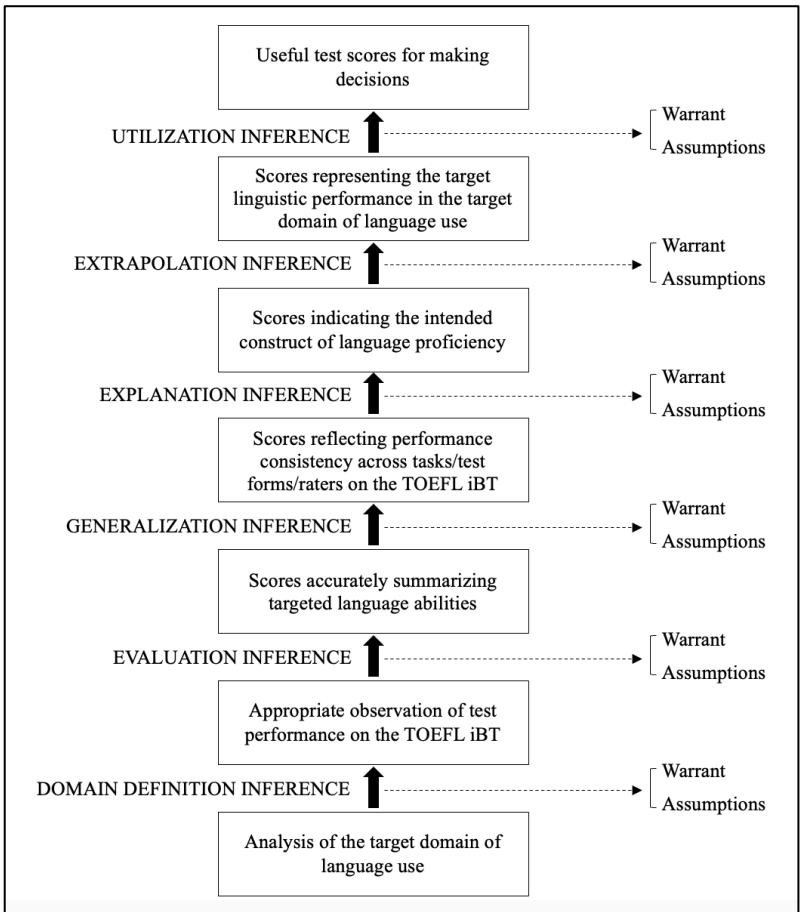
Chapelle et al. (2008) laid out the use of the argument-based framework in the development of a validity argument for the TOEFL iBT which, as mentioned before, is a test designed to

evaluate test takers' language readiness for academic study at English-medium universities. They first identified six inferences in the interpretive argument shown in Figure 1. These were the domain definition inference, the evaluation inference, the generalization inference, the explanation inference, the extrapolation inference, and the utilization inference. Chapelle et al. argue that a robust validity argument for the TOEFL iBT would need to provide evidence supporting each of these inferences.

The *domain definition inference* refers to the relationship between test performance observations with the knowledge, skills, and abilities needed for performance in the target domain. The *evaluation inference* refers to how test performance observations are evaluated. The *generalization inference* refers to the relationship between the scores of the observed test performance with the expected scores from parallel tasks and/or raters. The *explanation inference* refers to the relationship between test scores and theoretical constructs. The *extrapolation inference* refers to test scores being representative of real-life performance in the target domain. The *utilization inference* refers to the relationship between test scores and their uses.

Figure 1

Interpretive Argument for the TOEFL iBT



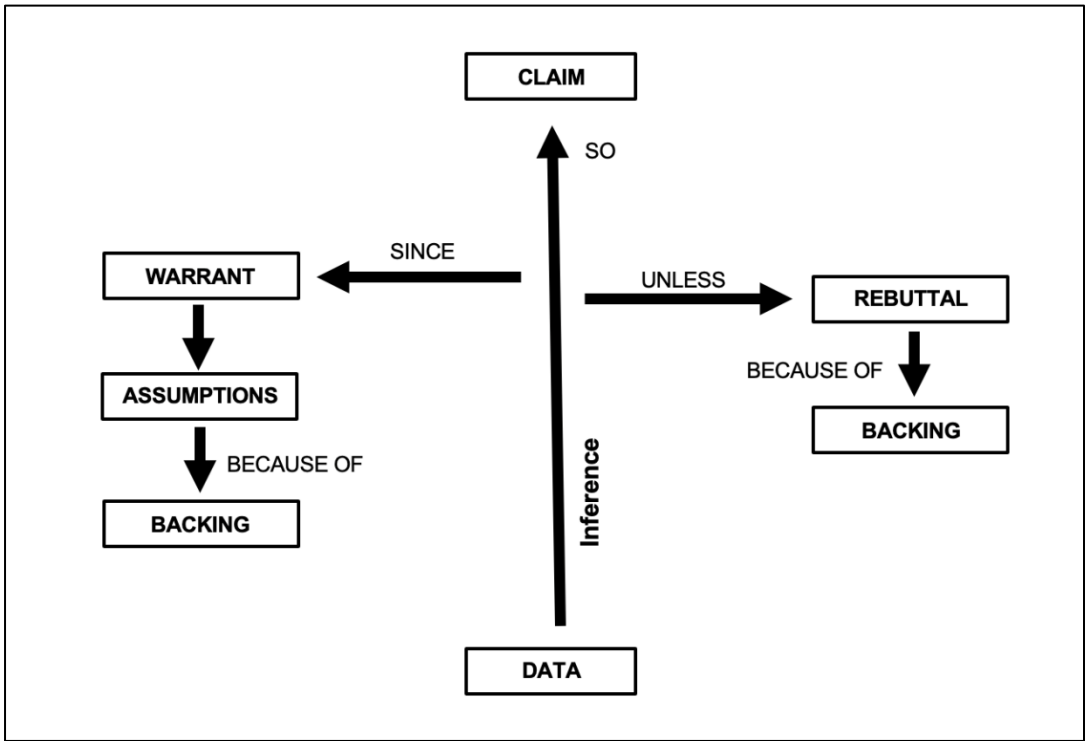
Note. (adapted from Chapelle et al., 2008; Chapelle & Voss, 2021)

Next, Chapelle et al. adapted Toulmin's (2003) argument model as shown in Figure 2, and employed it to investigate the assumptions underlying all the inferences identified in the interpretive argument. In this adaptation, the claims were the statements made about the TOEFL iBT (e.g., TOEFL iBT scores represent target linguistic performance in the target domain of language use). The inference was the logical link between the test data (i.e., test scores) and the claim or conclusion, and it was an assertion that needed justification. Inferences were justified when their warrants were judged to be plausible. A warrant could be a scientific theory, a body of

knowledge or evidence, or a precedent. Their plausibility was evaluated through the identification and investigation of the assumptions that underlay them, and the results were referred as backings. Exceptions or other relevant circumstances that might undermine the inference were referred as rebuttals and their backings were also identified and investigated (Chapelle et al., 2008, pp. 6-8).

Figure 2

Argument-Based Validity Structure Model



Note. Adapted from Chapelle et al., (2008, p. 7)

Multiple studies were conducted to examine the validity of the TOEFL iBT following this model. These studies were reported in Chapelle et al. (2008). For the domain description inference, teams of researchers identified academic domain tasks and the language abilities required for them, and studies were conducted to investigate their relevance (e.g., Enright & Schedl, 1999; Hale et al., 1996; Rosenfeld et al., 2001). For the evaluation inference, rubrics and multiple task

administration conditions were developed, trialed, and revised based on expert consensus and the results from statistical analyses (e.g., Brown et al., 2005; Cumming et al., 2001, 2002). For the generalization inference, reliability and generalizability studies were used (e.g., Zhang, 2006). For the explanation inference, studies using factor structure analyses and correlation analyses among the TOEFL iBT measures and other tests were conducted (e.g., Cumming et al., 2006). For the extrapolation inference, the relationships between test performance and students' academic placement tests, test takers' self-assessments, and instructors' judgements of students' English language proficiency were investigated (Chapelle et al., 2008, pp. 341-344). Finally, washback studies were used for the utilization inference along with information sessions to produce test materials and assist test users set cut scores (Chapelle et al., 2008, pp. 344-346).

Based on the results from all the theoretical analyses and the empirical data, Chapelle et al. argue that the conclusion for the validity argument was that the TOEFL iBT scores were valid for making decisions about test takers' language readiness for academic study at English-medium universities. This project served as an example for multiple studies published afterwards and investigating the validity of different language assessments, as Dursun and Li (2021) illustrated in a systematic review. In this review the authors found that the popularity of the argument-based approach in language test validity research increased steadily from 2005 to 2018. Additionally, since the publication of Chapelle et al. (2008) and until 2018, over 60 studies followed their model to investigate a variety of language assessments using quantitative and/or qualitative analyses. This trend accelerated after 2011 and included standardized language assessments and newly developed institutional or domain specific testing projects from North America and other continents (Dursun & Li, 2021, pp. 50-53).

Although a comprehensive validity argument in LAPP would include multiple inferences, this study will focus on the investigation of the extrapolation inference of the speaking section of the Spanish module of the *CanopyCredential* exam using the argument-based approach. The next section summarizes how this inference has been investigated previously in language testing contexts.

2.4 Use of the argument-based approach to investigate the extrapolation inference of language assessments

The extrapolation inference links test scores with examinees' performance in the target domain of language use and can be sustained if a relationship between these is found. In argument-based validity, this is investigated through criterion-related studies that compare performance in the test with other external evaluation criteria using quantitative, qualitative, or a combination of both analyses. This inference has been one of the most investigated in language test validity studies utilizing the argument-based framework and a variety of external evaluation criteria (Dursun & Li, 2021, pp. 54-58).

Some of the external evaluation criteria that have been employed so far include actual samples from the domain of interest (Brooks & Swain, 2014; Llosa & Malone, 2019; Riazi, 2016), learners' self-assessments (Li, 2015; Kumazawa et al., 2016; Santos, 2017), overall academic performance/grade (Isaacs et al., 2023; Harsch et al., 2017 ; Kelly-Riley & Elliot, 2014), linguistic corpora (LaFlair & Staples, 2017), other standardized language proficiency tests (Gaillard, 2014; Oh, 2018; Youn, 2013), and stakeholders evaluations (Isbell et al., 2023). Brooks and Swain (2014), Llosa and Malone (2019), and Li (2015) are of particular interest because some of the methods and ideas that they employed were used in this research project.

Brooks and Swain (2014) compared students' performances in a research version of the speaking section of the TOEFL iBT and two real-life academic speaking activities. The research version of the TOEFL iBT included two independent tasks (e.g., speaking about a familiar topic), and four integrated tasks (e.g., reading, listening, and then speaking). The participants were 30 international graduate students at a university in Canada who were studying Sciences and Social Sciences disciplines (e.g., Engineering and Psychology, respectively). They completed the research version of the TOEFL iBT and recorded themselves, without the researchers' presence, in an in-class and out-of-class academic speaking activity. The type of activities that the participants recorded were presentations and paired or small group discussions with their peers and/or professors. These activities were recorded within a month of doing the test. Participants production in the exam and the recordings were compared in terms of grammatical features (i.e., complexity and accuracy), discourse features (i.e., cohesion and register characteristics), and vocabulary.

The results of Brooks and Swain's study suggested that the speaking section of the TOEFL iBT aligned more closely with in-class spoken language use than out-of-class spoken language use, especially with monologues such as classroom presentations of research. This was observed through the overlap in the use of some discourse features (i.e., connectives, passivization, and nominalization) and vocabulary types between the test and the in-class performances. Grammatical complexity and inaccuracy were significantly greater in the participants' productions in the test than in their in-class and out-of-class productions. Additionally, the use of some discourse features (i.e., questions and informal language) was significantly greater in the out-of-class productions than in the in-class and test productions. Thus, since there were some similarities and differences in participants' productions across the three contexts (i.e., TOEFL iBT, in-class, and out-of-class),

Brooks and Swain concluded that their findings could pose a rebuttal to the extrapolation inference of the speaking section of the TOEFL iBT (Brooks & Swain, 2014, p. 370).

A novelty in Brooks and Swain's (2014) study was the use of actual samples from the target context (i.e., in-class and out-of-class recordings) to investigate the extrapolation inference of the speaking section of the TOEFL iBT, using the same participants for both the exam and the actual samples. This methodological procedure could be considered one of the most appropriate to investigate the extrapolation inference of an assessment like the TOEFL iBT. Moreover, another methodological procedure from Brooks and Swain's that would be appropriate to employ in similar studies is the data-driven selection of the linguistic features to be analyzed. That is, after the data was collected, Brooks and Swain's team of researchers compiled all the linguistic features of interest and selected those that were more relevant for their study based on a review of the literature, and the specific characteristics of the data that they gathered for their study.

Besides the participants' productions in the research version of the TOEFL iBT and the recorded real-life speaking academic activities, Brooks and Swain (2014) also obtained data from semi-structured interviews with all the participants. In these interviews, the participants were asked to reflect on their perceptions of the test and their real-life academic speaking, their reaction to the test tasks, and their speaking in their academic studies. However, the results from these data analyses were not reported in much detail in the article. The inclusion of this type of data in these research projects is important though because it allows the comparison of different data types (e.g., objective vs. subjective) and the triangulation of the results. Llosa and Malone (2019) is a good example of how this procedure can contribute to this type of research.

Llosa and Malone (2019) investigated the comparability of students' performance on the TOEFL iBT writing tasks and actual academic writing tasks with a group of 103 international non-

native-English-speaking undergraduates. They looked at participants' performance on two TOEFL iBT writing tasks and their performance in required writing courses in U.S. universities. They focused on five dimensions of writing quality (i.e., grammatical, cohesive, rhetorical, sociopragmatic, and content control) in the test tasks and two course assignments; instructors' ratings of students' writing and overall English proficiency; and instructor-assigned grades on the two course assignments.

The results from Llosa and Malone's study indicated that participants' performance in the test tasks was comparable to their performance in the course assignments, as shown through the association of all dimensions of writing quality between the test tasks and the academic writing tasks. Additionally, test scores were correlated with instructors' ratings of students' writing and overall English proficiency, but only the test task requiring the use of multiple skills (i.e., reading and listening and then producing a written response) was correlated with instructor-assigned grades on the course assignments. Llosa and Malone concluded that these results provided backing to support the TOEFL iBT extrapolation inference, suggesting that the construct of academic writing proficiency as assessed by the TOEFL iBT accounts for the quality of linguistic performance in English-medium universities (Llosa & Malone, 2019, p. 254).

An interesting feature of this study that was partially missing in Brooks and Swain (2014) is the combination and detailed analysis and report of multiple external evaluation criteria (i.e., linguistic measures, instructors' ratings, and instructor-assigned grades). Specially, some that could be considered more objective (i.e., linguistic measures) along with others that could be considered more subjective (i.e., instructors' ratings and instructor-assigned grades). Furthermore, as was the case in Brooks and Swain (2014), the use of actual samples from the target domain (i.e., the course assignments) is appropriate for investigating the extrapolation inference of language

assessments. However, using actual samples of performance is not always feasible for a variety of reasons such as privacy/confidentiality issues, among others. This could be the case for studies attempting to investigate this inference in LAPP designed for lawyers in the legal domain or doctors in the medical domain, as is the case with the CCE in this study.

As an alternative to actual samples from the target context, the development of specific evaluation tools to examine the extrapolation inference of language assessments has been approached in previous studies as well. For example, Li (2015) developed and used a contextualized self-assessment of English proficiency (i.e., reading, listening, speaking, and writing) to examine the extrapolation inference of an English Placement Test (MEPT) at a large Midwestern U.S. university. The participants were 181 newly admitted English as a second language (ESL) students from the same university where the MEPT was being used. Upon examining the quality of the self-assessment tool, its relationship with the MEPT and the TOEFL iBT was explored. The results showed weak to moderate relationships between the self-assessment and the two tests, although there were some significant correlations between the self-assessment and some sections of the MEPT (i.e., the writing sections of the MEPT and the self-assessment), and the TOEFL iBT (i.e., the TOEFL iBT speaking and the four subskills in the self-assessment). However, the results from the analyses investigating the self-assessment tool showed that it had an acceptable psychometric quality and a clear factor structure. This suggested that it could be potentially useful in future studies investigating the same inference and test while controlling for some of the features that seemed problematic in this study (e.g., it appears that the can-do statements had low item measure of difficulty, and the time gap between the self-assessment and the MEPT might have been too large).

In Li's (2015) study, developing an alternative to using actual language samples involved a review of the literature on self-assessment, the review of the self-assessment descriptors in the European Language Portfolio (ELP) and the ACTFL descriptors, and informal interviews with ESL students. Then, the self-assessment items were written as 'can-do' statements about the four skills (i.e., listening, reading, speaking, and writing), using a six-point Likert scale. Its quality was investigated using Rasch models and confirmatory factor analyses. Similar studies could follow this idea using the same procedure. Namely, a review of different sources of information, the design of the instrument based on the information gathered, and the examination of the instrument's quality and its suitability for the intended purpose.

In lieu of actual performance samples, simulations have also been shown to be another useful external evaluation criterion. An example of this is found in Woodward-Kron and Elder (2016) study where they compared performance from the speaking component of the OET with performance from Objective Structured Clinical Examinations (OSCEs) consisting of simulated health professional-patient communications from international medical graduates preparing for the Australian Medical Council clinical examination. OSCEs are simulations commonly used in medical schools and training worldwide to assess candidates' ability to apply content knowledge and clinical skills. In their study, Woodward-Kron and Elder used two different groups, one for the OET and another one for the OSCEs. Even though the results provided support for the OET tasks, these results could have been influenced by the fact that the groups were different. Nevertheless, the study shows that simulations are a useful external evaluation criterion, especially when collecting actual samples from the target context of language use is not possible for a variety of reasons. This is particularly true in the medical context where real data is difficult to obtain due to privacy concerns.

In such cases, simulations can be designed following procedures similar to those used by Li (2015) to better match the research questions, and to elicit data that is comparable to production from the test and the one that could be obtained from the target context of language use. The next section summarizes the use of simulations in learning and assessment and reviews their effectiveness.

2.5 Use of simulations in learning and assessment

According to Sokolowski and Banks (2009), the concept of simulations or SRLT encompasses the notions of models and simulations. A model is an approximation of the real-world and it is followed by one or many simulations which allow the repeated observation of the model. After this, analyses are performed to draw conclusions, verify and validate research, and/or make recommendations (Sokolowski and Banks, 2009, p. 3). The use of SRLT following these notions is common in many different disciplines including learning and assessment.

Educators and assessment specialists have been using SRLT for decades. Research focusing on their use and outcomes has shown that these instruments have a positive influence on learning and, of special interest for this project, are suitable to predict examinees and learners' performance in the target context. A good example of this can be found in the medical field, which was the context of this study.

A wide variety of SRLT have been designed to be included in simulation-based assessments (SBA) used in medical schools worldwide to evaluate students and examinees' clinical and professional skills (e.g., OSCEs) (Harden et al., 1975), and standardized licensure examinations such as the United States Medical Licensure Examination (USMLE) (see https://www.usmle.org/). The results from multiple studies suggest that there is a correlation between examinees' scores in the simulations included in these types of assessments and their

subsequent practice in real-life (e.g., Cuddy, Liu, Ouyang, et al., 2022; Dong et al., 2017; Komasawa et al., 2020). This research literature provided a warrant for using simulations to prove the extrapolation inference in a medical LAPP, and it is described in the next paragraphs.

The OSCEs were introduced by Harden et al. (1975) to assess the clinical competence of medical students, as an alternative to the traditional multiple-choice exams and clinical examinations with real patients. Originally, they consisted of several stations with simulated patients located in hospital wards where the students were first given a written instruction and had to carry out a medical procedure. After five minutes, they had to move to the second station where they had to answer questions on their findings and their interpretation of them. At some of the procedure stations the candidates' performances were observed and scored. The students' final scores were based on the number of correct and incorrect answers in the second station, and the score they received in the graded procedure stations. The total number of stations could vary as needed. Since their introduction, the OSCEs have become a standard method of assessment in medical schools worldwide, and have been modified to suit specific circumstances and needs. Results from recent studies have shown and association between OSCEs grades and subsequent performance in real-life.

For example, Dong et al. (2017) compared medical students scores in OSCEs with the evaluations they received from their program directors during their first year of residency (i.e., while working as medical trainees with real patients assigned to their supervising doctors). The participating medical students were from the Obstetrics/Gynecology (Ob/Gyn) clerkship ($n = 720$), and the Family Medicine clerkship ($n = 693$) at the U.S. Uniformed Services University of the Health Sciences. The results showed that poor performance on the OSCEs was significantly associated with poor ratings on medical expertise and professionalism on the first year program

directors' evaluations. Specifically, the odds of receiving a poor evaluation on medical expertise and professionalism for those at the bottom 10% of the OSCEs in Family Medicine were four times higher. For those in Ob/Gyn, the odds were three times higher on medical expertise, and almost six times higher on professionalism.

In a similar study, Komasawa et al. (2020) investigated the relationship between the OSCEs and the clinical clerkship evaluations of 94 Japanese medical students attending Osaka Medical College. In the clinical clerkships the students participated as members of a medical team to perform actual medical procedures and care on real patients. The results showed that overall, the OSCEs total scores were significantly correlated with the clinical clerkships scores. Nevertheless, when analyzed by OSCEs components, the results suggested that only the medical interview and chest examination scores were significantly correlated with clinical clerkship scores. Komasawa et al. argued that these results from the components' analyses could be due to the fact that the remaining components assessed by the OSCEs did not happen as often during the clinical clerkships. This, as the authors of the study also explained, lends support to the use of simulations for training and assessment since important competencies that candidates must have but do not happen as often during their real-life training periods can be included in simulations.

Regarding the use of simulations in standardized medical licensure examinations, the USMLE provides a good example. The USMLE is a three-step examination for medical licensure to practice in the United States. It assesses a physician's ability to apply knowledge, concepts, and principles, and demonstrate patient-centered skills (see https://www.usmle.org/). The Step 3 of the exam includes multiple-choice questions and computer-based case simulations. The computer-based case simulations are interactive simulations of patient-care situations that are designed to evaluate examinees' approach to clinical management, including diagnosis, treatment, and

monitoring, and are fully delivered via computer. They include a variety of patient-care situations and settings, over varying periods of simulated time. As simulated time passes, the patients' conditions change based on the course of the medical conditions, examinees' management, or both. Examinees' performances are scored using algorithms that represent codified expert physician-defined criteria (United States Medical Licensing Examination [USMLE], 2025).

Cuddy et al. (2022) conducted a study to examine the associations between the USMLE Step 3 scores and the likelihood of disciplinary action in practice. The sample (N = 275,392) included board-certified physicians representing 50 practice jurisdictions within the United States, including Washington, D.C.—which reflects a national sample—, and 17 major medical specialty areas, which reflects a range of training and expertise. Disciplinary actions can be received for behavior that threatens patients' safety and can include sanctions and/or license revocations. The results showed that physicians with higher Step 3 total scores tended to have lower chances of receiving a disciplinary action. Additionally, it was suggested that the computer-based case simulations scores may contribute important discrete information, above and beyond that offered from the multiple-choice questions scores, for understanding practice patterns that could receive disciplinary actions for unprofessional or incompetent behavior.

Although not often referred to as *simulations* in the literature, in the field of SLA, SRLT (known as *tasks*) have also been used for decades. In fact, two of the research areas within the field that have garnered more attention from researchers are task-based language teaching (TBLT) (Long, 1985, 2015), and task-based language assessment (TBLA) (Norris, 2009, 2016, 2018).

TBLT is a pedagogical methodology for second language (L2) teaching that uses tasks as the unit of instruction. These tasks are developed based on the needs of the language learners and are organized on the syllabus according to their relative complexity. The focus of this method is

on authentic communication, and it offers an alternative to other traditional pedagogies such as grammar-based syllabus or present-practice-produce. To give an example, in an analysis of the needs of learners of medical Spanish as an additional language it could be found that one of these needs is being able to explain laboratory results to a patient in Spanish (i.e., target task). Based on this, a series of pedagogical tasks that focus on authentic communication instead of specific grammar forms are developed. These tasks incorporate the language, procedures, and content knowledge that learners need to know to complete the target task (i.e., explain laboratory results to a patient), and are organized on the syllabus according to their relative complexity. Learners engage with these tasks while teachers monitor this engagement and offer opportunities for the learners to understand and acquire the authentic language use that is needed to complete each task and ultimately, the target task inside and outside the instructional setting (Bryfonski & McKay, 2019; Norris, 2009).

The effectiveness of the TBLT methodology has been investigated in a meta-analysis conducted by Bryfonski and McKay (2019). They looked at the long-term effects of TBLT programs implemented in authentic language classrooms with a focus on L2 learning outcomes. A total of 52 studies published between 1998 and 2016 were included in the analyses. Twenty-nine of these studies compared TBLT with a comparison or control group (i.e., between-group studies). Ten studies looked at the effect of TBLT on a single group over time (within-group studies), and 13 studies reported stakeholders' perceptions of TBLT programs upon program implementation (perception studies). Most of these studies were conducted in K-12 ($k = 18$) and university settings ($k = 25$). The results from the between-group studies showed a medium to large effect ($d = 0.93$) which suggested that TBLT was "effective for promoting L2 learning above and beyond the learning found in programs with other, traditional or non-task-based pedagogies" (Bryfonski &

McKay, 2019, p. 19). Moreover, the results from the perception studies showed an average of 79% positive rating of TBLT programs which indicated that these programs were perceived positively by stakeholders after being implemented.

Regarding TBLA, Norris (2016) defines this testing methodology as "the elicitation and evaluation of language use (across all modalities) for expressing and interpreting meaning, within a well-defined communicative context (and audience), for a clear purpose, toward a valued goal or outcome" (p. 232). As in TBLT, in TBLA tasks are the "basic unit of analysis in test design, interpretation, and intended consequence" (Norris, 2016, p. 230), and the focus is on what examinees can *do* with the language instead of what they *know* about the language (Norris, 2009, p. 578, emphasis added). The OET is an example of a language assessment that has adopted a task-based design. Its development process is described in McNamara (1996), and it included consultations with expert informants, literature search, and direct observations and collection of texts from the workplace. This allowed the elaboration of a list of prototypical tasks for each section of the test (i.e., speaking, listening, reading, and writing) along with their specifications. As an example, a task from the writing section of the OET for medical practitioners might ask examinees to read case notes from a patient and then write a letter of referral to another doctor (McNamara, 1996, pp. 97-116).

As Norris (2018) explains, a task-based assessment like the OET allows a:

> "(a) sufficient sampling of the knowledge/skill competencies identified to be critical by the profession; (b) high-fidelity replication or simulation of performance demands associated with the profession; and (c) determination and description of levels of ability deemed adequate for competent participation in the profession" (pp. 7-8).

In addition, Norris (2018) also sustains that TBLA aligns with the argument-based framework to test validity since it enables accurate interpretations about test takers that can lead to appropriate decisions and actions, and have a high likelihood of resulting in positive impacts and mitigating potential negative consequences (p. 5).

Besides the use of simulations or tasks for language teaching and assessment, these instruments have also been used in the field of SLA to review the assessment criteria employed in some current LAPP. For example, Elder and McNamara (2016) conducted a study to investigate if indigenous criteria aligned with the linguistic criteria utilized to assess speaking performance in the OET. Indigenous criteria referred to how domain experts viewed communication in real world settings. In their study, Elder and McNamara focused on the physiotherapy workplace and used two different settings. One of the settings was a workshop and the other a hospital. The participants in the workshop were physiotherapy educators (n = 12). These educators watched video recordings of students (n = 3) interacting with patients (n = 3) in a clinical setting, and were asked to provide feedback as if they were in a post-observation session with each student, which can be considered a simulation. The participants in the hospital setting were physiotherapy clinical supervisors (n = 14). These supervisors were asked to record themselves while giving feedback to individual students (n = 16) during or immediately after students' interactions with patients (n = 11), or at the end of the students' four-week clinical practicum.

The results from Elder and McNamara's study showed that the workshop setting was more useful and informative than the hospital setting. Namely, the researchers were able to gather minimal information about the kind of communication skills that were considered important for the physiotherapy workplace from the real context of language use (i.e., the hospital setting). In contrast, in the feedback gathered from the simulated post-observation sessions (i.e., the workshop

setting) the researchers were able to find the features of communication that were deemed important in clinical doctor-patient interactions. This provides support for the use of simulations for various purposes in language testing.

Besides the research-based positive results for learning and assessment obtained through the use of SRLT, stakeholders such as learners, examinees, and instructors also perceive them as useful and prefer them over more traditional methods of teaching and assessment like lectures or multiple-choice exams. For instance, Krusen and Martino (2020) examined occupational therapy students' perceptions of OSCEs using a qualitative descriptive analysis. Participants ($N = 40$) were second-year entry-level occupational therapy doctoral students enrolled at a U.S. university in the Pacific Northwest. They completed 17 OSCEs which were a culminating activity within a competencies course. They had five minutes to complete each OSCE and one minute to transition between them. Their performance was rated by faculty and practitioner experts who stayed at the same station throughout all the rotations. The students rated each OSCE immediately after completing each one of them using a paper-based Likert-type survey. Thirty-two students also completed an online survey including Likert-type and open-ended survey items within a week of finishing the OSCEs.

The results from the online Likert-type survey items in Krusen and Martino's study showed that 70% of the students felt the OSCEs helped them confirm their competence, and 80% of the students reported that they were able to identify areas for growth upon completion of the OSCEs. Furthermore, 53% of the students found the OSCEs as less stressful than written or clinical exams. The results from the online open-ended survey items suggested that the students perceived the OSCEs as an assessment *of* learning but also as an experience *for* learning (Krusen & Martino,

2020, p. 13, emphasis in original). This aligned with faculty observations about the OSCEs as formative and summative assessments.

Since SRLT have been used for teaching and assessment purposes for a long time obtaining positive outcomes, are well perceived by stakeholders, are a close approximation of real-life tasks, and can be designed to fit specific teaching/assessment needs, they can also be suitable external evaluation criteria for research studies investigating the extrapolation inference of LAPP. Specially, they are adequate for language assessments for the medical domain where actual samples are difficult to gather due to privacy concerns. That was the focus of this study.

2.6 Context of this study

The purpose of this study was to examine the suitability of SRLT as external evaluation criteria to investigate the extrapolation inference of LAPP. To this end, this research project used a SRLT specifically designed for this study to investigate the extrapolation inference of the speaking section of the Spanish module of the CCE. This SRLT represents a medical follow-up phone call between a doctor (role played by the participant) and a patient (role played by the researcher), and was developed following the theoretical frameworks of task-based language teaching (TBLT) (Long, 1985, 2015), task-based language assessment (Norris, 2009, 2016, 2018), and the language proficiency interview (Ross, 2017). The SRLT has a dialogue format and was conducted in real-time via Zoom. The SRLT features and its design process are described in the instruments section of the methodology chapter.

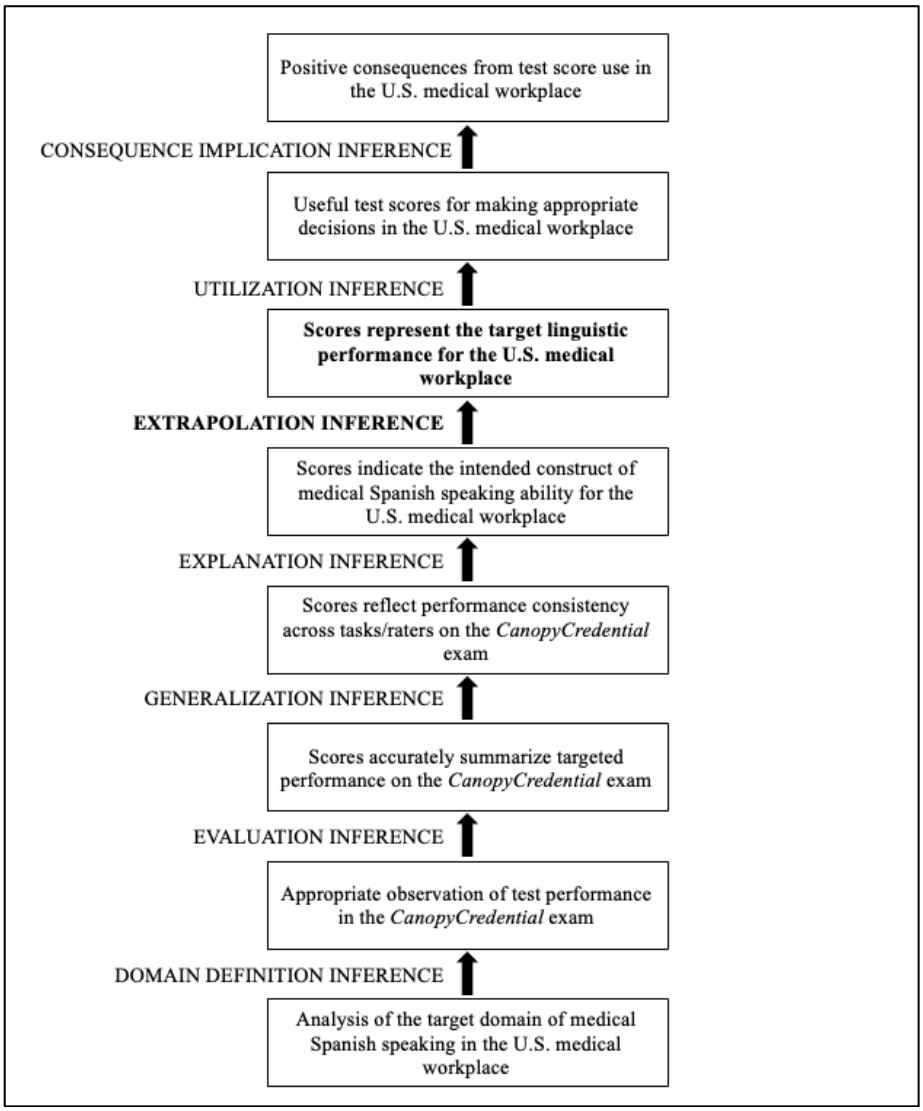
The CCE is a language assessment developed for the U.S. medical workplace with an English and a Spanish module. The Spanish module measures the medical Spanish listening and speaking abilities of healthcare professionals working or seeking employment in the U.S. medical workplace. This exam is fully delivered online, uses a scenario-based format, and all the test tasks

represent doctor-patient interactions. The medical Spanish speaking tasks ask participants to watch a pre-recorded video, read a prompt, and give a monologic spoken response that is recorded by the computer and subsequently rated by a team of certified raters. A description of the exam and the module and section that were investigated in this study are included in the instruments section of the methodology chapter.

As previously said, one assumption underlying the monologic, computer-mediated tasks that comprise the CCE is that examinees' performance will generalize to an actual interaction in the target language. This is the core of the extrapolation inference that this study was designed to investigate. As such, following the argument-based approach to test validity described before, an interpretation/use argument for the speaking section of the Spanish module of the CCE was developed. This interpretation/use argument serves as a guide to conduct validation research to build a validity argument for the speaking section of the Spanish module of the CCE. This interpretation/use argument is composed of a chain of seven inferences: the domain definition, evaluation, generalization, explanation, extrapolation, utilization, and consequence implication. The present study focused solely on the investigation of the extrapolation inference of the speaking section of the Spanish module of the CCE. The domain definition, evaluation, generalization, explanation, and utilization inferences have been investigated on previous studies (i.e., Needs analysis. The Canopy test of medical Spanish, n.d.; Validation study of Canopy Credential, n.d.). Figure 3 shows how each inference is positioned with respect to the others. Table 1 illustrates the warrants supporting each inference. The number of assumptions underlying each warrant are also described in Table 1 along with sources of backing to support these assumptions. The inference that was investigated in this study (i.e., the extrapolation inference) is marked in bold on Figure 3 and Table 1.

Figure 3

Interpretation/Use Argument Structure for the CanopyCredential Exam



Note. This interpretation/use argument is for the speaking section of the Spanish module of the *CanopyCredential* exam.

Table 1

Interpretation/Use Argument for the CanopyCredential Exam

Inference	Warrant	Assumptions	Sources for backing
Domain definition	Observations of performance on the *CanopyCredential* exam reveal relevant knowledge, skills, and abilities representative of those required for medical Spanish speaking in the U.S. medical workplace	Assessment tasks that are representative of the medical Spanish speaking in the U.S. medical workplace are identified	Domain tasks analysis
		Critical knowledge, skills, and abilities needed for the medical Spanish speaking in the U.S. medical workplace are identified	Domain language analysis; domain criteria analysis
Evaluation	Medical Spanish speaking performance is evaluated to provide observed scores that reflect the quality of performance on the *CanopyCredential* exam speaking tasks	The criteria and procedures for evaluating examinees' performance on the *CanopyCredential* exam speaking tasks are appropriate for providing evidence of targeted language abilities	Rubric development and use analyses
		Test administration conditions are appropriate for providing evidence of targeted language abilities	Examinees' survey analysis

(continued)

Table 1 (continued)

Interpretation/Use Argument for the CanopyCredential Exam

Inference	Warrant	Assumptions	Sources for backing
Generalization	Observed scores are estimates of expected scores over the parallel versions of tasks and across raters	Scores from parallel versions of the tasks are consistent	Generalizability and reliability analyses
		Ratings of different raters are consistent	Interrater reliability analysis
		Different ratings by the same rater are consistent	Intra-rater reliability analysis
Explanation	Expected scores reflect a construct defined as the ability to speak medical Spanish effectively for the U.S. medical workplace	The linguistic knowledge, speaking process and strategies required to successfully complete tasks are in line with theoretical expectations	Exploratory factor analysis
		Task difficulty is systematically influenced by task characteristics	Rasch measurement models

(continued)

Table 1 (continued)

Interpretation/Use Argument for the CanopyCredential Exam

Inference	Warrant	Assumptions	Sources for backing
Extrapolation	**The *CanopyCredential* exam target scores represent the kind of medical Spanish speaking ability required in the U.S. medical workplace**	**Participants' scores in the *CanopyCredential* exam are related to their scores in the simulated real-life task**	**Ordinal regression analyses**
		Participants' scores in the *CanopyCredential* exam are related to linguistic characteristics of participants' production in the simulated real-life task	
		Participants' scores in the *CanopyCredential* exam are related to their self-assessment of medical Spanish listening and speaking ability	**Spearman's correlation**

(continued)

Table 1 (continued)

Interpretation/Use Argument for the CanopyCredential Exam

Inference	Warrant	Assumptions	Sources for backing
Utilization	Estimates of the quality of medical Spanish speaking ability for the U.S. medical workplace obtained from the *CanopyCredential* exam are useful for making decisions	The meaning of test scores is clearly interpretable by examinees	Frequency analysis
Consequence implication	Decisions made based on the test score are beneficial	Positive washback effects of test use on learning	Stakeholders' interviews/surveys analyses

Note. This interpretation/use argument is for the speaking section of the Spanish module of the *CanopyCredential* exam.

As Figure 3 depicts, the domain definition inference links the target domain with observations of performance in the exam. The warrant for this inference for the medical Spanish speaking section of the CCE is that the observed speaking performance on the exam represents the medical Spanish speaking performance in the target context of language use (i.e., the U.S. medical workplace). This warrant is based on the assumptions that (1) the CCE speaking test tasks simulate those from the target context of language use, and (2) require the necessary medical Spanish speaking abilities. In a previous study (i.e., Needs analysis. The Canopy test of medical Spanish, n.d.), literature reviews, stakeholders' interviews, surveys, and workplace observations were conducted to examine the tasks, language, and criteria needed. The results provided backing for the mentioned assumptions. Therefore, the domain inference for this module/section of the exam is fully supported.

The evaluation inference connects observed test performance with observed score. The warrant underlying the evaluation inference for the CCE is that the observation of the medical Spanish speaking performance on the exam is properly evaluated, so that the observed score reflects the quality of medical Spanish speaking performance on the exam. This warrant depends on the assumptions that (1) the criteria and procedures used to evaluate examinees' medical Spanish speaking performance on the exam are appropriate for providing evidence of the targeted language abilities, and (2) the test administration conditions are appropriate for providing evidence of the targeted language abilities.

Results from a previous study (i.e., Validation study of Canopy Credential, n.d.) suggest that the speaking tasks in the Spanish module of the CCE have high levels of psychometric quality. In addition, the linguistic competence scale from the scoring rubric "performed optimally in terms of internal consistency, inter-rater reliability, rater agreement, and reliability of the separation" among different levels of ability (p. 45). Similar results were obtained for the effectiveness scale within the socio-cultural communicative competence rubric, but the impartiality scale was less useful to differentiate different levels of performance. (see Appendix A, Tables A1 and A2 for the CCE linguistic competence and socio-cultural communicative competence rubrics, respectively). Regarding test takers' perceptions about how the exam assessed their proficiency level, 80.41% of them fully or somewhat agreed that the exam measured their proficiency level appropriately. Thus, the evaluation inference of the speaking section of the Spanish module of the CCE is partially supported by these sources of backing since they point towards some issues with the impartiality scale from the socio-cultural communicative competence rubric. All other criteria, procedures, and test administration conditions are fully supported by the results from the validation study of Canopy Credential (n.d.).

The generalization inference links the observed scores to the expected scores over parallel versions of tasks and across raters. For the speaking section of the Spanish module of the CCE, this inference is based on the warrant that the observed scores are stable estimates of expected scores over parallel versions of tasks and raters. The warrant is based on three underlying assumptions: (1) scores from parallel versions of the tasks are consistent, (2) ratings of different raters are consistent, and (3) different ratings by the same rater are consistent. Results from the validation study of Canopy Credential (n.d.) showed that the linguistic competence and effectiveness scales within the linguistic competence and socio-cultural communicative competence rubrics reached dependency indices above 0.90. This indicates that the exam has an acceptable power to provide an estimate of how examinees would do in a similar test-taking context. This, along with the inter-rater reliability and rater agreement results mentioned in the description of the evaluation inference above, provides backings for these assumptions. However, the generalization inference of the section and module of the CCE under study is partially supported since results for the impartiality scale from the socio-cultural communicative competence rubric did not show enough variability to provide high dependability estimates.

The explanation inference connects expected scores with the tested construct. The construct for the speaking section of the Spanish module of the CCE is defined as the ability to speak medical Spanish effectively for the U.S. medical workplace. The warrant underlying this inference for the CCE's section and module under investigation is that the expected scores are reflective of the construct as previously defined. This warrant is based on the assumptions that (1) the linguistic knowledge, speaking process and strategies required to successfully complete the exam's Spanish speaking tasks are in line with theoretical expectations, and (2) the Spanish speaking tasks difficulty is systematically influenced by task characteristics. Results from an

exploratory factor analysis (EFA) conducted as part of the validation study of Canopy Credential (n.d.) yielded evidence backing these assumptions. Therefore, the explanation inference of the speaking section of the Spanish module of the CCE is fully supported.

The extrapolation inference links the test construct with target scores that are representative of target performance in the target context of language use. The warrant supporting this inference for the speaking section of the Spanish module of the CCE is that the exam target scores represent the kind of medical Spanish speaking performance required in the target context of language use (i.e., the U.S. medical workplace). This warrant requires support for the assumptions that (1) participants' scores on this section and module of the CCE are related to their scores on the SRLT, (2) participants' scores in the CCE are related to linguistic characteristics of their production in the SRLT, and (3) participants' scores in the CCE are related to their self-assessment of medical Spanish listening and speaking ability. Backing for these assumptions were sought in this study by examining (1) the relationships between participants' scores on the speaking section of the Spanish module of the CCE and the SRLT, (2) participants' scores on the speaking section of the Spanish module of the CCE and the linguistic characteristics of their production on the SRLT, and (3) participants' scores on the speaking section of the Spanish module of the CCE and their self-assessment of medical Spanish listening and speaking ability. A complete explanation of the analyses that were conducted is included in the data analyses section of the methodology chapter.

The utilization inference connects target scores to test uses. In the CCE's section and module addressed in this study, the warrant underlying this inference is that the estimates of the quality of medical Spanish speaking ability for the U.S. medical workplace obtained from the CCE are useful for making decisions. This warrant is based on the assumption that the meaning of test scores is clearly interpretable by examinees. Results from the validation study of Canopy

Credential (n.d.) suggest that examinees perceive this exam as an authentic instrument to assess their medical Spanish proficiency. This provides backing for the assumption, and support for the utilization inference of the CCE. However, more studies should be conducted to investigate other potential uses of the exam's scores.

The consequence implication inference links test uses to intended positive consequences of test scores uses. The warrant underlying this inference for the CCE is that decisions made based on the test score are beneficial. This warrant is based on the assumption that test use has positive washback effects on learning. Backing for this assumption could be obtained through stakeholders' interviews and surveys. To the best of the researcher's knowledge, no study has yet addressed this inference for the speaking section of the Spanish module of the CCE.

This interpretation/use argument outlines the validation needs for the CCE's module and section that were the focus of this study. As mentioned before, previous validation studies have investigated four of these inferences (i.e., domain definition, evaluation, generalization, explanation, and utilization). This study concentrated on the extrapolation inference. The next chapter describes the instruments and methodology that were used to conduct this research project.

Chapter 3: Methodology

To build a validity argument for the speaking section of the Spanish module of the CCE, it would be necessary to conduct multiple studies to investigate all the inferences stated above in the interpretation/use argument for this section/module of the CCE. As mentioned before, some of these inferences have already been investigated in previous studies. The present study focused on exploring the extrapolation inference of the speaking section of the Spanish module of the CCE, guided by the following research questions:

- RQ1: Do the SRLT facets (i.e., participant's ability, rating criteria, rating scale, and rater severity) function appropriately? This is a prerequisite to using scores from the SRLT as the criterion for the extrapolation inference.
- RQ2: Do the SRLT raters and linguistic scoring criteria reliably differentiate participants' proficiency level? To be useful, the SRLT should differentiate a range of abilities similar to the CCE.
- RQ3: To what extent are participants' scores in the speaking section of the Spanish module of the CCE related to their scores in the SRLT? This is the core of the extrapolation inference.
- RQ4: To what extent are participants' scores in the speaking section of the Spanish module of the CCE related to their production in the SRLT on linguistic measures of intelligibility, fluency, grammar, vocabulary, coherence, and effectiveness?
- RQ5: To what extent are participants' scores in the speaking section of the Spanish module of the CCE related to their self-assessment of medical Spanish listening and speaking ability?

The goal of RQ1 and RQ2 was to explore the suitability of the SRLT as external evaluation criterion by examining the functionality of its measurement components (RQ1), and the reliability of the raters and scoring instruments (RQ2). Once sufficient evidence supporting the appropriateness of the SRLT as external evaluation criterion was found, RQ3-RQ5 explored the extrapolation inference of the speaking section of the Spanish module of the CCE by analyzing the relationships between participants' medical Spanish speaking scores in the CCE and (1) their overall scores in the SRLT (RQ3), (2) the linguistic characteristics of their production on the SRLT (RQ4), and (3) their self-assessment of medical Spanish listening and speaking ability (RQ5). The next sections describe each one of the methods used in this study.

3.1 Participants

The data for this study was obtained from 82 participants with experience in the U.S. medical workplace as student, healthcare professional, or both, and who had previous knowledge of medical Spanish. They were all able to read, listen, speak and write fluently in English, and their ages ranged from 18 to 65 years old (M = 24.91, SD = 7.46). Sixty-seven of them were students in the healthcare field, nine were healthcare professionals, two were studying and working in the healthcare field simultaneously, and four were not studying or working in the healthcare field at the time of the study. Prior to completing the study tasks, they were asked to complete a self-assessment of their medical Spanish listening and speaking ability by choosing among four different options, the one that better described them. These four options aimed at differentiating among four different levels of medical Spanish listening and speaking ability, with option one being the lowest level and four the highest. The four options, their descriptions, and the number of participants that chose each one of them are displayed on Table 2. As displayed on Table 2, most

participants were at levels one through three, according to their self-assessment of medical Spanish listening and speaking ability.

Table 2

Participants by Self-Assessment of Medical Spanish Listening and Speaking Ability Options

Option number	Option description	*n*
1	I can understand and communicate simple medical information and sustain very simple medical interactions.	18
2	I can communicate medical information and sustain medical interactions with some fluency. I can understand familiar language normally encountered in medical interactions. Sometimes I might ask for clarification.	30
3	I can communicate medical information and sustain medical interactions fluently, with few hesitations. I can understand most language encountered in medical interactions.	29
4	I can communicate medical information and sustain medical interactions fluently and effectively. I can completely understand language encountered in medical interactions.	5

Participants also self-reported their use of Spanish language in four different contexts (i.e., at home, with friends, at school, and at work), and the age at which they started using this language in these environments. Forty-two participants self-reported using Spanish at home. The ages at which they started using Spanish in this environment ranged from 5 to 57 years old (M = 7.64, SD = 8.44). Fifty-nine participants self-reported using Spanish with friends, and the age range at which they started doing so was from 5 to 25 years old (M = 11.98, SD = 6.23). Regarding those using Spanish at school, 74 participants self-reported doing so, and the age range at which they started

was from 5 to 28 years old (M = 12.07, SD = 5.39). Fifty-seven participants self-reported using Spanish at work with a starting age range of 13 to 30 years old (M = 19.88, SD = 3.43).

English was the primary language of instruction for the majority of participants' medical education (i.e., medical school, nursing school, etc.) (n = 70), followed by Spanish (n = 4), and both English and Spanish (n = 1). This information was missing from seven participants. In addition, 43 participants reported having completed medical Spanish courses. The number of medical Spanish courses completed by participants in this group ranged from 1 to 16 (M = 2, SD = 2.48). Participants also self-reported their use of medical Spanish at school or at work while taking part in this study. According to their responses, 17 participants reported never using Spanish at school, 22 used it rarely, 28 used it sometimes, 14 used it regularly, and one used it always. As for their medical Spanish use at work, nine participants reported never using Spanish in this context, 11 used it rarely, 40 used it sometimes, 15 used it regularly, and seven used it always.

Overall, the characteristics of the participants who took part in this study are closely related to those of the individuals who normally complete the CCE for their personal or professional use (Needs analysis. The Canopy test of medical Spanish, n.d.; Validation study of Canopy Credential, n.d.).

3.2 Instruments

3.2.1 The *CanopyCredential* exam

The CCE evaluates bilingual English-Spanish medical speaking and listening skills and socio-cultural communicative competence using an interactive scenario-based format and is fully delivered online. Examinees can complete both, the English and Spanish modules, or just one, depending on their needs. Listening tasks are scored automatically and speaking tasks are scored

by a group of certified raters. Once all tasks have been scored, examinees receive a scorecard with a description of their proficiency level. The scores reflected on the scorecard are aligned with the Common European Framework of Reference for Languages' (CEFR) proficiency scale and range from the A2 to the C1 levels. Participants in this study completed the Spanish module of the CCE and only their CCE speaking score was used for this project. Thus, the next paragraphs focus on the speaking section of the Spanish module of the CCE.

The speaking section of the Spanish module of the CCE consists of two parts: Part A and Part B. Part A includes four tasks and Part B includes two tasks, making a total of six tasks. In all tasks examinees are asked to play the role of a healthcare professional. Each task—except Task 4 in Part A—requires examinees to watch a pre-recorded video and record their response following instructions from a prompt and after receiving some preparation time. Depending on the task, examinees have a maximum of four to five minutes to prepare, and the same amount of time to record their spoken response. The average length of the pre-recorded videos in Part A tasks is two minutes, and five minutes for those in Part B. Task 4 in Part A is a sight translation where examinees are prompted to explain to the patient a brochure about polysomnography.

The subjects in the pre-recorded videos are native speakers (NS) of Spanish actors playing the role of a patient. They are all directly looking at the camera and there is only one actor per video. All the videos in Part A are played by the same actor and form a multi-turn sequence in which the actor/patient and the examinee take turns speaking. However, since all the videos are pre-recorded, the production of both the actor/patient and the examinees have a monologic format. The two videos in part B constitute a single turn each and are played by two different actors with actor/patient and examinees spoken productions having the same format (i.e., monologic) as those from Part A.

Examinees' spoken responses are rated for linguistic and socio-cultural communicative competence on each test task. The rubric for linguistic competence (see Appendix A, Table A1) is aligned with the CEFR proficiency scale and the levels that are assessed range from A2 to C1. The target CEFR level for different tasks varies and so does the range of scores that can be assigned to examinees on each task. For all the tasks in Part A, the target CEFR level is High-B1 (HB1) and the score range is A2-HB1. For Part B, the target CEFR level for Task 1 is High-B2 (HB2) and the score range is A2-HB2, and for Task 2 the target CEFR level score is C1 and the score range is A2-C1. The rubric for socio-cultural communicative competence (see Appendix A, Table A2) includes four levels ranging from 1 (Ineffective), 2 (Effective), 3 (Competent), and 4 (Adept), and all four levels are used to assess examinees' performance on each task. The final scores for linguistic and socio-cultural communicative competence that examinees receive in their score cards is calculated using a computerized algorithm based on the scores they got for each competence (i.e., linguistic and socio-cultural) on each test task, and the different weights that have been assigned to every test task. The scores that were used in this study were those that participants received for linguistic competence. Table 3 summarizes the structure, topics, and score ranges of the speaking section of the Spanish module of the CCE.

Table 3

Characteristics of Tasks in the Speaking Section of the Spanish Module of the CCE

Speaking section	Task number	Task topic	Target level	Linguistic competence score range	Socio-cultural communicative competence score range
Part A	1	Describe/ask about symptoms.	HB1	A2-HB1	1 (Ineffective)-4 (Adept)
	2	Explain results, diagnosis, and treatment; provide multi-steps instructions.	HB1	A2-HB1	1 (Ineffective)-4 (Adept)
	3	Address patient's questions and concerns; provide multi-steps instructions and recommendations.	HB1	A2-HB1	1 (Ineffective)-4 (Adept)
	4	Sight translation (polysomnography brochure).	HB1	A2-HB1	1 (Ineffective)-4 (Adept)
Part B	1	Explain results, diagnosis, and treatment; provide recommendations.	HB2	A2-HB2	1 (Ineffective)-4 (Adept)
	2	Explain treatment; provide recommendations.	C1	A2-C1	1 (Ineffective)-4 (Adept)

3.2.2. The simulated real-life task

The development of the SRLT followed a series of stages. First, the literature on TBLT and TBLA for task design, development, and use was reviewed. Equally, various publications on language proficiency interviews (LPIs) and other LAPP such as the OET were also reviewed. The reports from the need analyses and the theoretical framework for the CCE were examined as well. Then,

the task topic was chosen. The task topic was a simulated medical follow-up phone call in which the participant (playing the role of the healthcare professional) had to call a patient (role played by the researcher) to follow-up on a previous in-person office visit. The next step consisted in designing the task materials.

The first version of the task materials was a prompt, results sheets from a medical laboratory, doctor's notes from a patient's in-person visit to a primary care medical office, and a task scoring rubric. The laboratory results and doctor's notes were designed using authentic materials as models and all personal information was modified or eliminated. These materials were sent in different sets to three different expert groups for their review. Each group had a specific area of expertise. Namely, TBLT/TBLA, healthcare, and medical Spanish. There were two individuals per expert group. Those in the TBLT/TBLA group were researchers with decades of experience in this domain. The experts in the healthcare group were an emergency room doctor and a nurse practitioner, both with years of experience in primary care. The experts in the medical Spanish group were instructors of medical Spanish courses also with years of experience teaching these courses.

The healthcare expert group reviewed the results sheets from the medical laboratory, and the doctor's notes for authenticity, and the other two groups reviewed the CCE exam scoring rubrics, and the SRLT prompt and scoring rubric focusing on the design and content. All experts reviewed the materials individually in their own time and met with the researcher afterwards individually via Zoom for a one-hour discussion. The prompt and medical documents were modified based on their feedback to make them more close-to-real. They also advised to use the CCE rubric instead of the one designed by the researcher to make the results more comparable.

The updated materials were piloted with a group of five volunteers from the same population group that was sampled for this project. The procedures were also the same than the ones that were used in the final project: each volunteer met with the researcher individually via Zoom and after a brief introduction and description of the task, the researcher shared with them the materials through a link to a Google Doc pasted on the chat (with access removed once the task was completed), gave them seven minutes to prepare, and asked them to complete the task within 10 minutes. The videos were turned off during the task and only the audios were recorded. All the audios were sent via Qualtrics to the same expert groups afterwards along with a brief questionnaire where they were asked to provide feedback on the task being close-to-real enough, its appropriateness to assess medical Spanish speaking ability, and the researcher's performance playing the patient's role.

Several conclusions were reached after analyzing the audios and the experts' feedback. First, the materials were too long (seven pages in total), and all volunteers needed more than seven minutes to prepare for the task which would extend too much the duration of the experiment. Thus, all the materials content was summarized in a one-page single document, and it was decided that the preparation time would be reduced to five minutes, but the task completion time would remain the same (i.e., 10 minutes). The experts noted that the volunteers talked too much during the task and the simulated patient talked too little. Whereas they agreed that the volunteers' behavior represented a real situation quite well, the simulated patient's behavior did not and recommended the inclusion of more interruptions in the form of questions, reactions, or complaints. Therefore, a guideline was designed for the researcher's performance as a simulated patient in the SRLT to achieve a better turn balance and make it more close-to-real, while allowing enough opportunities for the examinees to produce a sample that would allow an accurate evaluation of their medical

Spanish speaking ability. This version of the task was piloted with a different group of six volunteers and enhanced until the production from both interlocutors was considered satisfactory for the purpose of this research project. The final version of the task prompt is shown in Figure 4, and the guideline for playing the simulated patient's role is shown in Figure 5.

Figure 4

Simulated Real-Life Task Prompt

Task description

SETTING: Follow-up medical phone call – Suburban Healthcare Clinic

CONTEXT:
You work as a healthcare provider in the primary care unit of a suburban healthcare clinic. This 39-year-old female patient (señora Kahlo) came in-person to your office a week ago because she was experiencing new symptoms (excessive fatigue, thirst, hunger, weight loss, frequent urination, tingling in her fingers, and blurry vision). This was her first visit to your office and clinic. You took her medical history, conducted a physical exam, and ordered some labs. You just received her lab results which include the following abnormal values:

Name	Patient's Value	Reference Range
Cholesterol	270 **(H)**	100-199 mg/dL
Triglycerides	200 **(H)**	20-149 mg/dL
HDL Cholesterol	30 **(L)**	>=40 mg/dL
LDL Cholesterol	160 **(H)**	<130 mg/dL
Glucose	187 **(H)**	70-100 mg/dL

TASK:
After reviewing her symptoms and the lab results, call the patient over the phone and:

- Using a pseudonym instead of your real name, explain reason for phone call (follow-up following office visit). Find out how the patient has been feeling since visit (fatigue level, thirst and hunger, frequency of urination, changes in vision).
- Ask patient about her lifestyle (work/life balance, diet, exercise). Find out further relevant details (any family history of cholesterol and diabetes).
- Explain the lab results and deliver diagnosis (high levels of cholesterol and triglycerides indicate that patient has hyperlipidemia; glucose level also high--further tests needed to discard diabetes).
- Give information about hyperlipidemia (blood has too many lipids/fats which increases fatty deposits in arteries and the risk of blockages). Outline treatment of hyperlipidemia (reduce fat intake in diet; at least 30 minutes of moderate-intensity aerobic exercise 5 days per week).
- Express your opinion about plant-based diet.
- Convince patient that lifestyle changes such as diet and exercise can improve her cholesterol levels and there is no need to prescribe any medication at this moment. If cholesterol levels don't improve after making these lifestyle changes, then you would consider prescribing medication.
- Tell patient to come back to the clinic in one month.

PREPARATION TIME: 5 minutes.

- You can take notes and ask me questions about the task during the preparation time.
- You can use your notes and look at this document during the phone call.

TASK TIME: 10 minutes.

- In your response you should incorporate information from this prompt and the patient's speech, show empathy and understanding, and be socio-culturally appropriate.
- Remember that we are interested in your medical Spanish skills, NOT your medical knowledge.

Figure 5

Guideline for the SRLT's Simulated Patient

Task description

SETTING: Follow-up medical phone call – Suburban Healthcare Clinic

CONTEXT: You are a 39-year-old elementary school teacher. A week ago you went to the doctor because you have been experiencing some new symptoms for a few months (excessive fatigue, thirst, hunger, weight loss, frequent urination, tingling in your fingers, and blurry vision). This was your first visit with this doctor and clinic. They took your medical history, conducted a physical exam, and ordered some labs. Today the doctor who saw you is calling you over the phone.

TASK:

- When asked, say you are still having the same symptoms. You feel very tired all the time, hungry and thirsty. You need to use the bathroom more than usual and your vision is blurry.
- When asked, say that you are an elementary school teacher and work too many hours. You don't have any free time and don't exercise at all. You don't have any time to cook either, so you eat at the school cafeteria which serves pizza, hot dogs, chicken nuggets and the like.
- When asked, say that both of your parents have high cholesterol and diabetes and your grandmother had diabetes too and needed to inject herself with insulin.
- Say that you don't understand very well what hyperlipidemia/high cholesterol is and how it can affect your health.
- Say that you think what they serve at the school cafeteria can't be too bad since that is what they cook for the children. You can't exercise because you are very tired when you get home from work and you only want to sit on the couch and watch TV.
- Say that you have heard about plant-based diet and ask doctor what they think about it.
- Ask doctor to prescribe a medication to lower your cholesterol level. Say you have heard of people whose cholesterol level only improved after taking medication.
- Say that you will try to change your diet and exercise but if your cholesterol level doesn't improve, insist you would like to take some medication.

As can be seen in Figure 4, the SRLT prompt included multiple elements and topics from the CCE (see Table 3 above). Moreover, the proficiency criteria index for LPIs (Ross, 2017, p. 46) and the OET speaking tasks' structure and interlocutors' cards were used as models as well (see medicine sample tests in https://oet.com/ready/speaking). The idea was to combine a variety of features for the task to represent a close-to-real situation from the medical context and elicit a sample that would allow an overall accurate evaluation of examinees' medical Spanish speaking ability. To this end, all the medical information (i.e., context, laboratory results, and sub-tasks topics) were verified to match the profile of a patient with high cholesterol levels. Furthermore, the sub-tasks (i.e., bullet points list in the TASK section in Figure 4) were organized in a predicted

range of difficulty with the easier ones at the beginning and the more difficult at the end, based on the literature and the proficiency criteria index for LPIs (Ross, 2017, pp. 43-99). For example, one of the tasks used at level 1 of the Interagency Language Roundtable (ILR) in LPIs requires the examinee to ask questions (Ross, 2017, p. 46). So, the first two sub-tasks in this study's SRLT require the examinee to ask questions to the patient regarding symptoms, lifestyle, and medical family history. These and all the previously described steps were taken to ensure that the SRLT was an appropriate instrument to assess the target construct of medical Spanish speaking proficiency for the U.S. medical workplace. Table 4 outlines the correspondence between the ILR scale levels and the proficiency criteria index for LPIs, and the sub-tasks order for this project's SRLT along with its proficiency criteria index and predicted range of difficulty.

Table 4

Correspondence between the LPI and the Study's SRLT

LPI		Simulated real-life task		
ILR scale levels	Proficiency criteria index	Sub-tasks order	Proficiency criteria index	Predicted difficulty
1	Candidate asks questions.	1-2	Examinee asks questions to the patient about symptoms, lifestyle, and medical family history.	Easier
2	Description/Report facts	3	Examinee explains laboratory results to the patient and delivers diagnosis.	
2	Description/Report facts	4	Examinee gives information to the patient about hyperlipidemia and outlines its treatment.	
3	Supported opinion	5	Examinee expresses their opinion about plant-based diet.	
4	Influencing tasks (convince, persuade, advise)	6	Examinee advices patient to follow the proposed treatment option.	More difficult

Note. Adapted from Ross (2017)

Following the advice from the expert groups and after evaluating different drafts, it was decided to use a modified version of the CCE rubrics to rate participants' performance in the SRLT. This was done to better suit the characteristics of the SRLT, and the purpose of this study. As illustrated on Appendix A, Tables A1 and A2, the linguistic competence and socio-cultural competence rubrics used to rate examinees' medical Spanish speaking performance in the CCE

are holistic rubrics where each score represents a combination of criteria. The criteria for linguistic competence are intelligibility, fluency, grammar, vocabulary and coherence, and the criteria for socio-cultural communicative competence are effectiveness and impartiality. As mentioned before, CCE examinees receive a score for linguistic competence and another one for socio-cultural communicative competence on each test task, and these scores are used to calculate their final scores for linguistic and socio-cultural competence. Thus, the scores examinees receive on each test task along with their final score represent all these criteria combined. For example, if an examinee achieves a final score of A2 for linguistic competence in the medical Spanish speaking section of the CCE, this score represents the overall degree of intelligibility, fluency, grammar, vocabulary, and coherence of their performance in the exam, as described on the CCE rubric for linguistic competence for this particular score (see Appendix A, Table A1, A2 score).

To obtain a better understanding of the psychometric properties of the SRLT, and because this was a single task with a series of embedded sub-tasks with different predicted levels of difficulty instead of multiple test tasks at different levels of difficulty, as in the CCE, it was decided to evaluate participants' performance in the SRLT by each individual linguistic criterion included in the CCE rubrics. As such, participants' performances in the SRLT were rated on intelligibility, fluency, grammar, vocabulary, coherence, and effectiveness by assigning a score on a scale from one to four, one being the lowest, to each one of these linguistic criteria. By doing so, each participant's production on the SRLT could receive different scores on each one of these linguistic criteria instead of an overall score representing all the linguistic criteria combined. Figure 6 displays an example of the score rater A assigned to a participant's performance on the SRLT on each linguistic scoring criterion. Table 5 illustrates the rubric that was employed to evaluate participants' performance on the SRLT.

Figure 6

Example of a Scored SRLT Performance



Note. The linguistic scoring criteria are listed in the left column, and the rating scale is displayed in the table's top cell. The blue dots indicate the score rater A assigned to this participant on each linguistic scoring criterion.

Table 5

Scoring Rubric for the SRLT

Domain	Score	Descriptor
Intelligibility	1 Inadequate	Sometimes difficult to understand with severe strain for the listener. Very strong influence of L1 often causes pronunciation or prosodic errors.
	2 Minimal	At times difficult to understand with some strain for the listener. The noticeable influence of L1 sometimes causes pronunciation or prosodic errors.
	3 Good	Most of the time easy to understand but with occasional strain for the listener. The influence of L1 may cause minor pronunciation or prosodic errors.
	4 Very good	Always easy to understand with minimal or no strain for the listener. The minimal influence of the L1 does not cause any misunderstanding.

(continued)

Table 5 (continued)

Scoring Rubric for the SRLT

Criteria	Score	Descriptor
Fluency	1 Inadequate	Utterances beyond short and isolated phrases and sentences, but speech is constantly interrupted through many pauses, false starts, and reformulations.
	2 Minimal	No lengthy utterances, but oftentimes effortless speech, especially when talking about basic ideas. Rare long pauses are noticeable.
	3 Good	Some relatively lengthy utterances with some effort and hesitation, but long pauses are not common.
	4 Very good	Mostly lengthy utterances that are almost effortless, with a few occasional hesitations and unnatural pauses, when the topic is difficult.
Grammar	1 Inadequate	Noticeable inaccuracies with obvious and basic errors that sometimes cause misunderstanding.
	2 Minimal	Noticeable inaccuracy but the errors rarely cause misunderstanding.
	3 Good	Mostly accurate grammar. Errors are not very common and oftentimes corrected on the spot.
	4 Very good	Consistently accurate grammar, but with a few occasional difficult-to-spot errors, especially while talking about complex and unfamiliar topics.

(continued)

Table 5 (continued)

Scoring Rubric for the SRLT

Criteria	Score	Descriptor
Vocabulary	1 Inadequate	Very limited range of vocabulary for both everyday situations and medical terms and expressions. Gaps in vocabulary knowledge causes lots of frequent hesitations and repetitions. May rely on L1 when unable to retrieve a word.
	2 Minimal	Working range of vocabulary for both everyday situations and medical terms and expressions. While gaps in vocabulary are noticeable and hesitations are frequent, the speaker is able to communicate a message, mostly without resorting to L1.
	3 Good	A good range of vocabulary for a wide range of general topics, but no use of low frequency words and colloquialisms. Competent use of medical terms and expressions.
	4 Very good	A broad range of vocabulary including the occasional use of relatively low frequency words and colloquialisms. Expert-level use of medical terms and expressions.
Coherence	1 Inadequate	No clear organization. Mainly relies on the use of listing cohesive devices. Clear connections between ideas are often missing.
	2 Minimal	Disorganized and jumpy, especially when discussing unfamiliar topics and elaborating on complex ideas with longer utterances.
	3 Good	Mostly organized, but at times disorganized and jumpy, especially when elaborating on complex ideas with longer utterances.
	4 Very good	Well organized and smoothly flowing narratives with sufficient organizational patterns.

(continued)

Table 5 (continued)

Scoring Rubric for the SRLT

Criteria	Score	Descriptor
Effectiveness	1 Inadequate	Incomplete communicative-task goal achievement. Limited delivery of somewhat accurate information. Cannot provide information in a language accessible to patients. Obvious issues transitioning between formal/technical and informal/non-technical language.
	2 Minimal	Partial communicative-task goal achievement. Partial but still acceptable delivery of mostly accurate information. With effort provides information in a language accessible to patients. Occasional issues transitioning between formal/technical and informal/non-technical language.
	3 Good	Slightly less than full communicative-task goal achievement. Slightly less than full delivery of almost completely accurate information. Almost always provides information in a language accessible to patients. Mostly smooth transitions between formal/technical and informal/non-technical language.
	4 Very good	Full communicative-task goal achievement. Full delivery of completely accurate information. Always provides information in a language accessible to patients. Smooth transitions between formal/technical and informal/non-technical language.

As can be seen, the impartiality criterion from the CCE socio-cultural communicative competence rubric was not included in the SRLT scoring rubric. This was based on the results from the validation study of the CCE which suggested that all the criteria within the linguistic competence rubric along with the effectiveness criterion within the socio-cultural communicative competence rubric were the ones that "performed exceptionally well" and reached "high internal consistency, inter-rater reliability, rater agreement, and reliability" whereas the impartiality criterion "was not as helpful for differentiating between proficiency levels" (Validation study of Canopy Credential, n.d., p. 1).

For this study, having the linguistic criteria scored individually allowed the researcher to conduct specific analyses to better investigate the suitability of the SRLT as external evaluation criterion. It also allowed her to perform better comparisons between participants' performances in the exam and the SRLT. A description of the analyses that were used in this study is included in the data analyses section of this chapter.

3.3 Procedures

Participants were recruited from healthcare schools and institutions and medical Spanish courses across the United States and Canada. The recruitment methods included email messages, flyers (see Appendix B, Figure B1), and newsletters announcements. Email messages were sent to department heads of healthcare schools and institutions, and instructors of medical Spanish courses for distribution. They were also sent to individuals interested in participating in the study and who had voluntarily shared their email addresses with the researcher. Flyers were posted on public bulletin boards located in healthcare schools and institutions, and locations where medical Spanish courses took place. They were also distributed in medical Spanish classrooms during researcher's visits to verbally explain the details and purpose of the study. Newsletters announcements were sent to various editors for publication.

All participants completed a pre-screening bio-data survey including their age, professional status, a self-assessment of their medical Spanish listening and speaking ability, questions about their use of Spanish in various environments, and medical Spanish courses completed (see Appendix B, Figure B2). This survey was used to ensure that all participants met the eligibility criteria, and to have participants at different levels of medical Spanish proficiency. It was delivered individually via Qualtrics and participants were able to access it through a shared QR code or link. An email was sent to those who fully completed this survey informing them if they were eligible

to participate in the study or not. Those who were eligible received further instructions in the same email to complete the consent form.

Once participants completed the consent form, they received instructions to complete the Spanish module of the CCE, and the SRLT. The order in which participants completed the CCE and the SRLT was randomly counterbalanced. Thirty-nine participants completed the CCE first and then the SRLT, and 43 participants followed the opposite order. This was done to minimize the potential influence of a practice effect. Furthermore, to minimize the potential influence of proficiency variations while participating in the study, all participants completed the second instrument (i.e., either the CCE or the SRLT) within two weeks of completing the first instrument except for two participants who completed the second instrument 24 and 25 days after completing the first one, respectively.

The CCE was fully delivered online. Participants were able to take it at their own location of choice, using a computer or laptop with Internet connection and a speaker and microphone. They received instructions in English to sign-up, access, and complete the assessment. The platform tested their equipment and Internet connection automatically to ensure they were working appropriately. The duration of the assessment was two hours approximately with a 10-minute break between Parts A and B. All responses were recorded by the platform.

The SRLT was completed via Zoom. Each participant met individually with the researcher at a date/time previously agreed by both parts. After a brief greeting and introduction, the researcher explained the task and allowed the participant to ask questions. Then the researcher and the participant did a quick 1–2-minute warm-up in Spanish to get used to each other's tones and voices using questions such as why they chose to study/work in the healthcare context, how a normal day looked like for them at work/school, what their favorite part was of being a healthcare

professional/student. Following the warm-up, the researcher shared the task prompt through a link to a Google Doc via the chat feature and gave the participant five minutes to prepare for the task. During this time, the participant was allowed to ask questions about the task and take notes. The participant could use their notes and look at the task prompt while completing the SRLT. These task-preparation procedures resembled those used in the CCE and in similar LAPP for the medical domain like the OET (see medicine sample videos in https://oet.com/en-us/ready/speaking). Additionally, research findings regarding the relationship between task preparation time/type and its influence on linguistic measures of complexity, accuracy, lexis, and fluency (CALF) in participants' oral production were also taken into account (e.g., Brown Nielson & DeKeyser, 2019; Bui & Huang, 2018; Gavin, 2014).

Once the SRLT preparation time was over, the researcher announced it, asked the participant to turn the video off if it was on, started the audio recording, and played a sound effect for phone ring tones. The task began once the researcher simulated answering the phone playing the role of the simulated patient. The SRLT lasted around 10 minutes or 600 seconds ($M = 697.23$, $SD = 163.41$) and participants were able to see a timer on their screen. After the task was completed, participants were dismissed and logged out from Zoom. The total duration of this section of the experiment did not exceed 30 minutes and only the SRLT was audio recorded, per the approved IRB protocol.

Upon full completion of all the study instruments within the specified time period, all participants received an official *CanopyCredential* score card at no monetary cost to them (the standard rate is $295 per attempt), and their names were entered in a raffle of one $50 Amazon gift card. The *CanopyCredential* scorecard was offered by Canopy Innovations, Inc., which provides the CCE and offers other products for healthcare institutions and professionals.

Each participants' performance in the CCE and the SRLT was rated by two or three raters using the rubrics described in the instruments section. Participants' performance in the CCE were rated by certified raters from Canopy Innovations, Inc. Following procedures from the test provider, a third rater was used in the CCE performances when the scores assigned by the two first raters were different. Participants' performance in the SRLT were rated by a team of three raters specifically trained by the researcher to assist with this task.

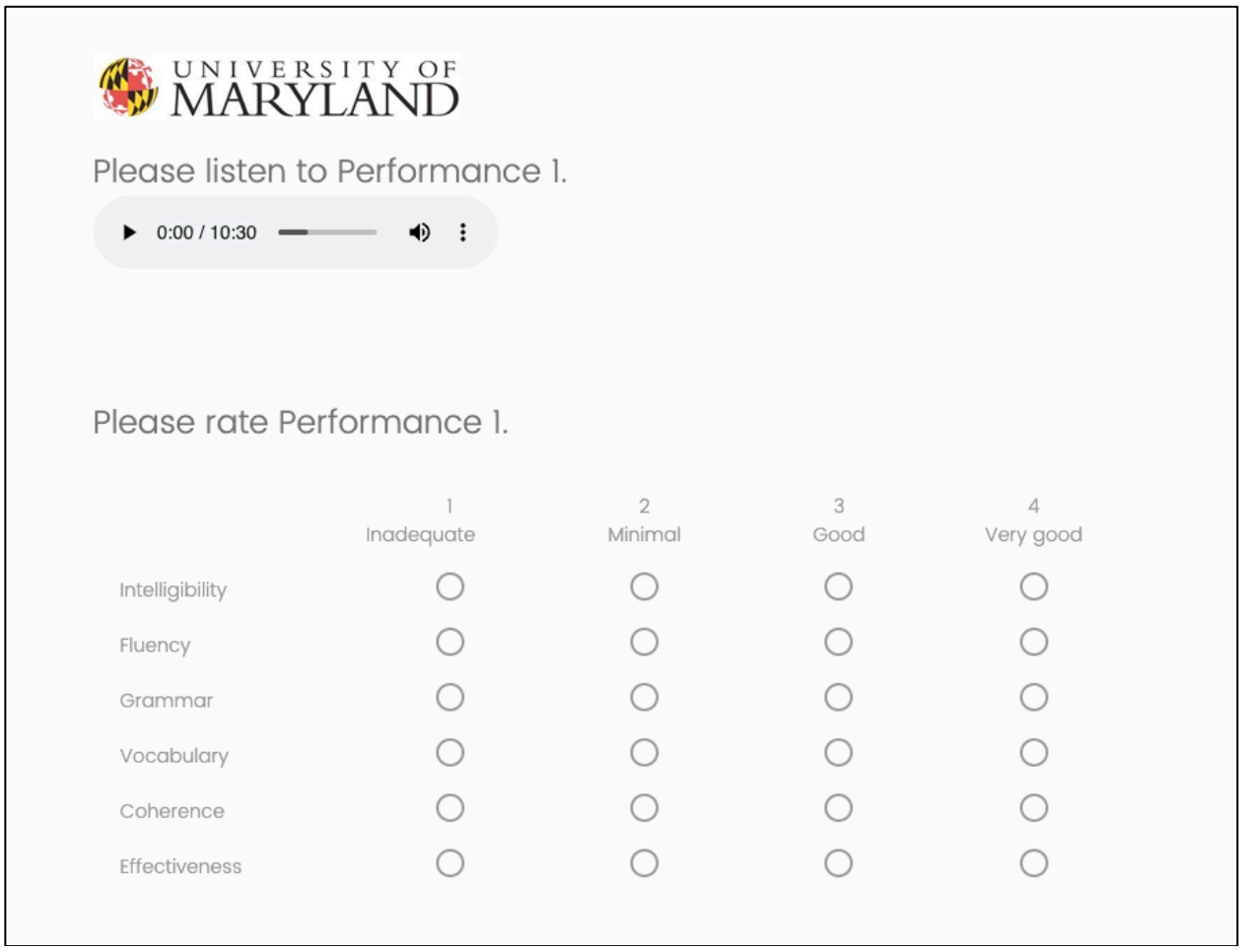
The SRLT raters were all Spanish professors with previous experience rating similar tasks. The training they received to complete this task involved a two-hour group meeting with the researcher via Zoom, and an individual rating practice exercise. Before the Zoom group meeting, the raters had reviewed the task description (Figure 4 above), the guideline for the interlocutor playing the simulated patient role (Figure 5 above), and the SRLT scoring rubric (Table 5 above). These items were explained during the Zoom meeting, followed by a rating exercise where all the raters scored five different SRLT performances. After this rating exercise, a group discussion took place to go over the ratings and answer the raters' questions. The individual rating practice exercise consisted of rating 10 different SRLT performances. Raters completed this exercise individually during their own time, within a week from the Zoom group meeting. This exercise was delivered and completed via Qualtrics, using the same format that was used for the study participants' ratings. Five of the SRLT productions that the raters scored in these two rating exercises were from participants in this study. Therefore, there were five study participants whose SRLT performance was rated by three raters whereas the rest of the participants SRLT performance was rated by two raters.

All participants' audio recordings from the SRLT and the scoring rubric were shared with the SRLT raters' team via Qualtrics. Raters were able to listen and rate each production in

Qualtrics. No personal information from the participants was shared with the raters. They were not able to download participants' audio recordings and were asked to sign a confidentiality agreement form (see Appendix B, Figure B3). Figure 7 displays the template that was created and used to rate participants' performances in Qualtrics. The SRLT scoring rubric and the task description that participants received (Table 5 and Figure 4 above, respectively) were also included in this Qualtrics template and displayed underneath the content shown in Figure 7.

Figure 7

Qualtrics Template used to Rate Participants' SRLT Performance



As mentioned before, each participants' performance in the SRLT was rated by two different raters from the SRLT raters' team, with the exception of five participants whose performance was rated by all three raters. Participants' performances were assigned to each rater following a judging plan to ensure that each rater had several participants' performances shared with each one of the other raters in the team. This judging plan was divided in four different medical Spanish speaking proficiency sections, using the medical Spanish speaking scores that participants received in the CCE. As such, section one included all participants receiving an A2 score in the CCE, section two included participants receiving a B1 score (i.e., low-B1 and high-B1 combined), section three was composed of all participants receiving a B2 score (i.e., low-B2 and high-B2 combined), and section four included all participants receiving a C1 score. Then, a letter was assigned to each rater (i.e., A, B, C), and pairs of raters were created (i.e., A & B; A & C; B & C). Participants' performances were assigned to these pairs of raters. When assigning performances to raters, the researcher made sure that each rater had a minimum number of participants' performances from each proficiency level. By following this judging plan, the requirements for the Rasch model that was used in this study was fulfilled. This requirement indicates that there needs to be enough linkage among the parameters included in the analyses. The judging plan employed in this study is illustrated on Appendix B, Table B1.

Each participant's production in the SRLT was transcribed and a range of linguistic measures were coded for data analyses purposes. These linguistic measures were selected based on the linguistic scoring criteria included in the CCE and the SRLT scoring rubrics (i.e., intelligibility, fluency, grammar, vocabulary, coherence, and effectiveness), the target construct of medical Spanish speaking ability for the U.S. medical workplace, a review of the literature, and the characteristics of the data that was collected for this study. These linguistic measures included

pronunciation errors, pauses, explained medical terms, grammar clauses, grammar errors, clitic pronouns, subjunctive tense forms, medical terms, and conjunctions.

The pronunciation errors measure involved phoneme errors (e.g., hiperlipid*i*mia instead of hiperlipid*e*mia), syllable errors (e.g., hiperlidemia instead of hiperli*pi*demia), and word stress errors (e.g., arte*ri*as instead of ar*te*rias). These types of pronunciation errors were selected following recent studies investigating the extrapolation inference of the speaking section of standardized language assessments for the academic domain (Isbell et al., 2024). Following Bui and Huang (2018), the pause measure selected for this study refers to any silence of 0.40 second or longer (p. 99). The explained medical terms measure followed the criteria by Ortega et al. (2023). As such, a medical term was considered as explained when the participants (1) provided a clear and accurate definition or explanation (e.g., el *colesterol alto* es como una grasa que se forma en las venas de su cuerpo [English translation: *High cholesterol* is like fat that forms in the veins of your body]), (2) provided a word/phrase synonym in plain language (e.g., la *diabetes* o azúcar en la sangre [English translation: *Diabetes* or blood sugar]), and/or (3) provided sufficient context during the dialogue so that the meaning of the term became clear (e.g., participant clearly and accurately defines a term after delivering diagnosis and then re-uses the term, without re-defining it, when explaining treatment) (Ortega et al., 2023, p. 4).

Grammar clauses were selected based on the literature investigating the relationship between speaking proficiency and CALF measures (Norris & Ortega, 2009). Regarding grammar errors, these were also selected following similar studies for language assessments in the academic domain (Isbell et al., 2024). The grammar errors selected for the present study were lexical errors, morphological errors, and syntactic errors. Additionally, the use of clitic pronouns and subjunctive tense forms was also selected based on the literature on processability theory which suggests a

relationship between Spanish speaking proficiency development and the use of these forms (Bonilla, 2015).

The use of medical terms measure was selected based on the literature on medical Spanish. According to the results from Ortega et al.'s (2023) study, there might be a relationship between medical Spanish speaking proficiency level and the use of medical terms. Therefore, following Ortega et al.'s (2023) criteria, in this study a word was considered a medical term if it was listed in a Spanish medical dictionary (i.e., *Diccionario de términos médicos*) (Real Academia Nacional de Medicina de España, 2012).

Finally, the use of conjunctions was selected as a measure based on previous studies investigating the relationship between speaking proficiency test scores and real-life speaking performance in the academic domain (Brooks & Swain, 2014). In the present study, the conjunctions measure included coordinating and subordinating conjunctions.

The unit of analysis used in this study for all the linguistic measures selected was the analysis of speech unit (AS-unit). The AS-unit is considered in the literature as one of the most appropriate unit of analyses for speaking performance data (Norris & Ortega, 2009). In addition, it seemed particularly well suited for the type of data that was collected for this study. According to Foster et al. (2000), "an AS-unit is a single speaker's utterance consisting of *an independent clause, or sub-clausal unit,* together with any *subordinate clause*(s) associated with either" (p. 365, emphasis in original). As Foster et al. argue, this unit of analysis is a robust measure for spoken performance data because it can be used to evaluate finished and unfinished statements that can appear across turns (pp. 365-372). This property of AS-units matched the characteristics of the data that was collected for this study.

Table 6 summarizes the linguistic measures selected for this study and their relationship with the linguistic criteria included in the CCE and the SRLT scoring rubrics. As can be seen in this table, there is one or more linguistic measures associated with each one of the linguistic criteria included in the CCE and the SRLT scoring rubrics. This was done to allow a thorough examination of the relationship between participants' scores in the CCE and the linguistic characteristics of their performance in the SRLT since participants' scores in the CCE are intended to be representative of all these linguistic criteria. The linguistic measures for the grammar criterion were separated in two different sets, one focusing on grammatical complexity and the other on grammatical accuracy.

Table 6

Study's Linguistic Criteria/Measures

Linguistic criteria	Linguistic measures
Intelligibility	Pronunciation errors (phoneme) per AS-unit Pronunciation errors (syllable) per AS-unit Pronunciation errors (word stress) per AS-unit
Fluency	Pauses per AS-unit
Grammar – Complexity	Clauses per AS-unit Clitic pronouns per AS-unit Subjunctive tense forms per AS-unit
Grammar – Accuracy	Lexical errors per AS-unit Morphological errors per AS-unit Syntactic errors per AS-unit
Vocabulary	Medical terms per AS-unit
Coherence	Coordinating conjunctions per AS-unit Subordinating conjunctions per AS-unit
Effectiveness	Explained medical term (definition) per AS-unit Explained medical term (synonym) per AS-unit Explained medical term (context) per AS-unit

Before conducting the statistical analyses, participants' indexes for each one of the linguistic measures selected were obtained as follows. Each participant's transcription was segmented in AS-units. Then, the total number of occurrences of each linguistic measure was counted and divided by the total number of AS-units on each participant's production.

3.4 Data analyses

3.4.1 Analyses conducted to examine the simulated real-life task's functionality

As mentioned before, research questions one and two addressed the appropriateness of the SRLT as an external evaluation criterion:

- RQ1: Do the SRLT facets (i.e., participant's ability, rating criteria, rating scale, and rater severity) function appropriately?
- RQ2: Do the SRLT raters and linguistic scoring criteria reliably differentiate participants' proficiency level?

Following Knoch and Chapelle's (2018) report to validate rating processes within an argument-based framework, the goal of RQ1 and RQ2 was to examine (1) if the rating scale/ linguistic scoring criteria were used as intended in the rubric's development process to score participants' performance in the SRLT, (2) the reliability of the raters, (3) the rating consistency of the raters, and (4) the relationship between the criteria descriptors in the rating rubric and linguistic features related to the target construct of medical Spanish speaking ability for the U.S. medical workplace. To investigate these, the MFRM (Linacre, 1994) was used.

The data used for the MFRM was obtained from participants' production in the SRLT (N = 82), and the raters' team (N = 3). The descriptive statistics were calculated in R statistical software, version 4.4.2 (R Core Team, 2024), using the psych R package, version 2.4.12 (Revelle,

2024). The many-facet Rasch model (MFRM) (Linacre, 1994) and the *FACETS* software, version 4.2.4 (Linacre, 2024a) were used for the MFRM analysis. The steps and guidelines used by Aryadoust (2016, pp. 7-19) and Linacre (2025) were followed and adapted to report the results from the MFRM analysis.

The MFRM follows the mathematical model developed by Rasch (1960) (i.e., the Rasch model). As Linacre (2024b) explains, the Rasch methodology assumes that the latent variable is the truth. When this latent variable is expressed in linear terms, the Rasch model is necessary and sufficient to describe it. The Rasch model uses a logarithmic transformation to convert ordinal data into a real number (logit) scale. Unlike ordinal data scales, the intervals between real numbers in the logit scale have a consistent value or meaning. This allows researchers investigating specific assessment types to better interpret test items and test takers' performance in the test (Bond et al., 2021, pp. 43-49). However, no data perfectly fits the Rasch model. The coefficients obtained through Rasch model analyses are used by researchers to determine whether the results can be used for their specific research purposes.

The MFRM is an extension of the Rasch model and allows the introduction of one or more additional components of the measurement situation such as rater severity or rating scale. These components or facets are analyzed simultaneously, so they can be examined separately and compared to one another (Bond et al., 2021, pp. 145-162; Linacre, 2024b, pp. 17-18). In this study, the latent variable of interest was medical Spanish speaking ability for the U.S. medical workplace. All the instruments/components that were used in this study and that have been described in the previous sections had been designed to measure this latent variable. The MFRM transformed the ordinal data that was obtained through the data collection process and expressed them in linear terms following this formulation:

$$\ln\left[\frac{P_{njik}}{P_{njik-1}}\right] = B_n - C_j - D_i - E_k$$

where

P_{njik} is the probability of participant *n* rated by rater *j* on linguistic criterion *i* receiving a rating of *k*;

P_{njik-1} is the probability of participant *n* rated by rater *j* on linguistic criterion *i* receiving a rating of *k-1;*

B_n is the ability (proficiency) of participant *n*;

C_j is the severity of rater *j*;

D_i is the difficulty level of scoring criterion *i*;

E_k is the difficulty level of the threshold from category *k-1* to category *k* of the scale to score criterion *i* (adapted from Aryadoust, 2016; Wind & Engelhard, 2013).

According to this formula, the probability of a participant receiving a certain score on a certain linguistic scoring criterion is a function of the ability of the participant to speak medical Spanish, the severity of the rater, and the difficulty of the linguistic scoring criterion (Linacre, 1994). Upon conducting the MFRM analysis, the elements of these facets were represented by independent parameters on a logit interval scale (log-odd units) along with their standard error of measurement. The standard error of measurement was estimated to determine the precision of each parameter. Fair average measures (FAM) were also produced by the MFRM analysis. These are generated when "each element's FAM in logits is converted to a raw score and then adjusted for the impact of other facets" (Aryadoust, 2016, p. 8; Aryadoust & Liu, 2015).

In addition, the MFRM analysis also generated reliability statistics for each facet. Reliability statistics are calculated as the ratio of true variance over observed variance (Linacre,

2025). These coefficients were used to interpret how different the elements in each facet were. In the MFRM, the reliability index ranges from zero to one and higher index values indicate a wider spread of the elements in each facet. Thus, in this study, higher reliability indexes for participants and linguistic scoring criteria (i.e., intelligibility, fluency, grammar, vocabulary, coherence, and effectiveness) were preferred, whereas lower reliability coefficients for raters were desirable. Furthermore, separation coefficients were also produced by the MFRM analysis as part of the reliability statistics. These coefficients assisted in the interpretation of the number of statistically distinct levels of performance in the data sample (Aryadoust, 2016; Bond et al., 2021, p.59; Linacre, 2025).

Chi-square tests of statistical significance were also produced by the MFRM analyses. These coefficients were used to interpret the differences between distinct levels of participants' medical Spanish speaking ability, raters' severity, and scoring linguist criteria's difficulty. Significant *p* values rejected the hypothesis that the levels had equal measures (Aryadoust, 2016).

Infit and outfit mean square (MNSQ) statistics were also computed. These statistics facilitated the diagnosis of aberrant observations and idiosyncratic elements (Linacre, 2025). Infit statistics place more weight on expected observations whereas outfit statistics are more sensitive to outliers (Bond et al., 2021, p. 56). Infit and outfit values can range from 0 to infinity but the expected value is 1.00 (Bond et al., 2021, pp. 54-56). For rating scale measurement like the one that was used in this study, infit and outfit values between 0.60 and 1.40 are considered adequate (Bond et al., 2021, p. 109).

Additionally, the MFRM analysis produced standardized (ZSTD) infit and outfit values as well. These values are a conversion of MNSQ values to normally distributed standardized statistics, and range between $-\infty$ and $+\infty$. Whereas MNSQ fit values provide useful fit information,

ZSTD values indicate how closely the data fit the model (Aryadoust, 2016). Following Linacre (2025) recommendations to interpret how closely the data fit the model, in this study the infit and outfit values of the elements in each facet were explored first. When any of these values fell outside the range of 0.60-1.40 recommended by Bond et al. (2021), their ZSTD values were examined applying the ±2.00 threshold recommended by Linacre (2025).

Regarding the examination of the functionality of the scoring rating scale used in this study, the guidelines for designing rating scales explained in Bond et al. (2021, pp. 222-237), and enumerated in Aryadoust (2016, p. 9) were employed and applied as follows:

Guideline 1: Each rating category needed to be applied a minimum of 10 times to ensure an accurate estimation of each rating category threshold. These thresholds are the boundaries between adjacent scoring categories.

Guideline 2: Raters needed to choose the scoring categories in a uniform manner.

Guideline 3: Average measures in the scoring categories needed to be monotonic. That is, the observations in higher categories should be related to high-proficient participants, whereas those in lower categories should be related to low-proficiency participants. According to Bond et al. (2021, p. 226) and Aryadoust (2016, p. 9), monotonicity exists when the difference between adjacent categories increases between 1.5 and 5 logits.

Guideline 4: The MNSQ fit coefficients needed to fall in the range between 0.60 and 1.40.

Guideline 5: Scoring categories needed to show an increasing trend with no disordering. As such, the measurements of the lower categories needed to be smaller and the measurements of the higher categories needed to be larger.

Guideline 6: The relationship between raw scores and MFRM linearized scores needed to be strong. A weak relationship could be caused by construct-irrelevant factors.

3.4.2 Analyses conducted to examine the extrapolation inference of the *CanopyCredential* exam

RQ3-RQ5 focused on investigating the extrapolation inference of the speaking section of the Spanish module of the CCE:

- RQ3: To what extent are participants' scores in the speaking section of the Spanish module of the CCE related to their scores in the SRLT?
- RQ4: To what extent are participants' scores in the speaking section of the Spanish module of the CCE related to their production in the SRLT on linguistic measures of intelligibility, fluency, grammar, vocabulary, coherence, and effectiveness?
- RQ5: To what extent are participants' scores in the speaking section of the Spanish module of the CCE related to their self-assessment of medical Spanish listening and speaking ability?

Ordinal regression analyses were used for RQ3 and RQ4, and Spearman's correlation was used for RQ5. The data used for these analyses were participants' final speaking scores in the CCE and their logit scores in the SRLT (RQ3); participants' final speaking scores in the CCE and the indexes for the linguistic measures selected from their production in the SRLT (RQ4); and participants' final speaking scores in the CCE and their responses to the self-assessment of medical Spanish listening and speaking ability (RQ5).

Since participants' speaking scores in the CCE were a categorical variable, and their SRLT scores and indexes for the linguistic measures were a numerical/continuous variable, ordinal regression analyses were considered as the most suitable analyses for RQ3 and RQ4. This type of analysis allows researchers to predict the behavior of a categorical dependent variable (i.e., CCE scores in this study), employing a set of independent variables (i.e., SRLT scores and linguistic measures indexes in this study). A logit link function was used to enable model estimation, and

the fitted models were compared to a null model using a likelihood ratio test. The Pearson chi-square coefficients generated provided information about the strength of the relationship between the dependent and the independent variables. R^2 coefficients gave information about how much variance was explained by the independent variables. A likelihood ratio test was used to evaluate the assumption of proportional odds (Statistics solutions, 2025).

Because the inclusion of the linguistic measures used to address RQ4 was largely exploratory, a decision was made to investigate them in the analyses by linguistic criteria (i.e., intelligibility, fluency, grammar, vocabulary, coherence, and effectiveness). As such, an ordinal regression analysis was conducted for each one of these linguistic criteria using CCE medical Spanish speaking scores as the dependent variable, and the set of linguistic measures selected for each criterion (see Table 6 above) as the independent variables. To help protect against Type I errors arising from multiple comparisons, the alpha level was set at 0.01 and 99% confidence intervals are reported. For this same reason, R^2 coefficients were interpreted with caution as well.

Spearman's rank-order correlation was used for RQ5 because both variables (i.e., CCE medical Spanish speaking scores and self-assessment of medical Spanish listening and speaking ability) were categorical. This analysis was used to measure the direction and strength of the relationship between these two variables.

All these analyses were conducted using R statistical software, version 4.4.2 (R Core Team, 2024), and the packages ordinal, version 2023.12-4.1 (Christensen, 2023), rcompanion, version 2.4.36 (Mangiafico, 2024), and brant, version 0.3-0 (Schlegel, 2020). The boxplots, category probability plots, and heatmap were generated using the tidyverse R package, version 2.0.0 (Wickham et al., 2019), the dplyr R package, version 1.1.4 (Wickham et al., 2023), the ggplot2 R package, version 3.5.1 (Wickham, 2016), and the viridis R package, version 0.6.5 (Garnier et al.,

2024). Christiansen (2022) tutorial was followed to perform and report the results of the ordinal regression analyses.

Chapter 4: Results

This chapter is composed of two sections. The first section describes the results from the analyses conducted to explore the suitability of the SRLT as external evaluation criterion, following the steps and model used by Aryadoust (2016). The second section includes de results from the analyses performed to examine the extrapolation inference of the speaking section of the Spanish module of the CCE. All these results and the research questions that guided this study are discussed in chapter five.

4.1 Suitability of the SRLT as external evaluation criterion

Participants' production on the SRLT were rated on six linguistic scoring criteria (i.e., intelligibility, fluency, grammar, vocabulary, coherence, and effectiveness), using a scale from 1-4 with 1 being the lowest score and 4 being the highest score. Table 7 displays the mean, standard deviation, skewness, and kurtosis coefficients of each one of these linguistic scoring criteria. According to these results, the vocabulary criterion (M = 2.84, SD = 1.01) is the most difficult scoring criterion, and the intelligibility (M = 3.15, SD = 0.95) and the coherence (M = 3.15, SD = 0.87) criteria are the easiest scoring criteria. All the scoring criteria have a normal distribution, as suggested by their skewness coefficients which range from -0.89 (intelligibility criterion) to -0.39 (vocabulary criterion), and kurtosis coefficients which range from -1.00 (vocabulary criterion) to -0.19 (coherence criterion).

Table 7

Descriptive Statistics for the Linguistic Scoring Criteria

Linguistic criteria	Mean	*SD*	Skewness	Kurtosis
Intelligibility	3.15	0.95	-0.89	-0.24
Fluency	3.06	1.07	-0.79	-0.71
Grammar	2.99	0.99	-0.53	-0.89
Vocabulary	2.84	1.01	-0.39	-1.00
Coherence	3.15	0.87	-0.77	-0.19
Effectiveness	3.00	0.93	-0.57	-0.63

Three facets (i.e., participants, raters, and linguistic scoring criteria) were calibrated in the MFRM analysis. Figure 8 presents a Wright map or variable map. The first column on the left-hand side is a ruler-like scale ranging from -6 to +8. The units of this scale are logits and the anchor or central point is zero. Each participant is represented by a star (*). Participants are arranged on the second column from left to right according to their estimated medical Spanish speaking proficiency level. Each participant was rated by two or three raters on the six linguistic scoring criteria displayed in the fourth column of the map (from left to right). Fifty-seven participants are located above the anchor point (i.e., zero), one participant is located at the anchor point, and 24 participants are located below the anchor point. Highly proficient participants are at the top of the map and low proficient participants are at the bottom.

Participants' medical Spanish speaking proficiency as measured by the SRLT range from -5.79 to +8.30 logits (spread: 14.09 logits). Participants' fit values (MNSQ) range from 0.20 to 1.93, which indicates that there are participants with fit values outside the fit range of 0.60-1.40 recommended by Bond et al. (2021). The standardized (ZSTD) infit and outfit values of those participants outside the recommended MNSQ fit range of 0.60-1.40 were examined and only four participants had values slightly above or below the ±2 threshold recommended by Linacre (2025).

This suggests that the majority of the participants fit the model acceptably well. In addition, participants' separation and reliability statistics are 4.11 and 0.94 (χ^2 = 1920.40, df = 81, p < 0.001), respectively, indicating approximately four statistically distinguishable proficiency levels. This suggests that the measurement was successful in discriminating between low- and high-proficient participants, and that a four-level scale is appropriate.

The third column presents the raters' severity/leniency levels. More severe raters appear at the top and more lenient raters appear at the bottom. The raters' severity/leniency levels range from -0.68 to +0.98 logits (spread: 1.66 logits). All raters fit values fall within the fit range of 0.60-1.40 recommended by Bond et al. (2021), which suggests that the raters fit the model well. The average infit and outfit mean-squares are 1.00 and 0.98, respectively. Since these mean-square averages are near 1.0, this provides further evidence suggesting that the raters closely fit the model (Linacre, 2025). The inter-rater agreement statistics show that all raters closely agree with each other (range of agreement: 55.5%-59.40%), and with what would be the expected agreement if the data fit the model perfectly (range of agreement: 55.5% [lowest observed agreement]-64.10% [highest expected agreement]). The summary of the inter-rater agreement statistics indicates that out of 552 inter-rater agreement opportunities, the observed exact agreements are 314 (56.90%), and the expected agreements are 340.10 (61.60%). This suggests that the raters closely behaved as independent experts, which is what it is expected from the model. The raters' separation and reliability statistics are 6.95 and 0.98 (χ^2 = 100.90, df = 2, p < 0.001), respectively, indicating heterogeneity among raters. This suggests that the raters had varying degrees of severity which were account for by the model.

Figure 8

Wright Map Displaying the Location of the Elements in Each Facet

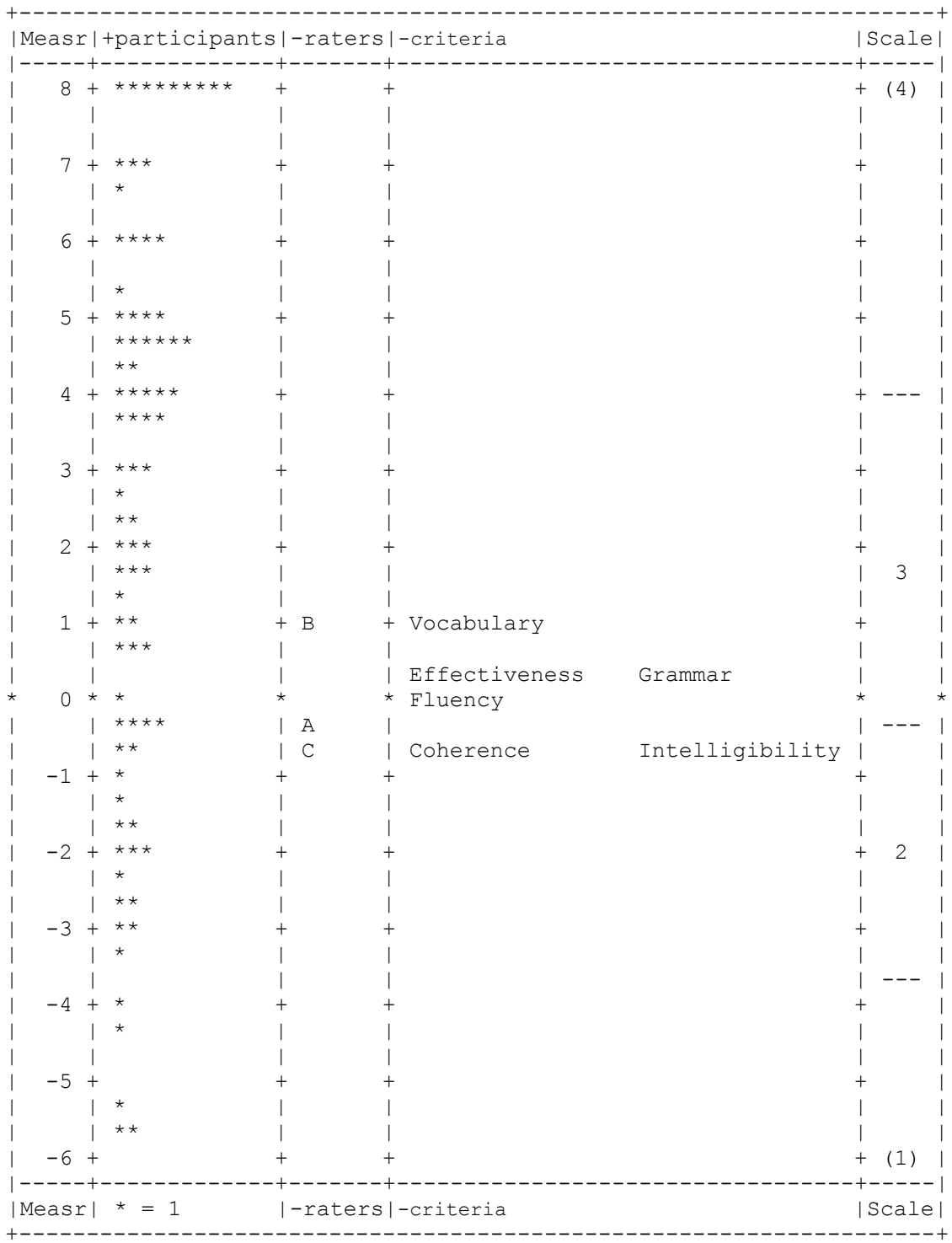


Table 8 shows the linguistic scoring criteria logit difficulty measures, fair average measures (FAM), standard error of measurement (SE), infit and outfit mean squares (MNSQ) and z-standardized (ZSTD) coefficients. The linguistic scoring criteria difficulty ranges from -0.63 to +0.95 logits with a 1.58-logit spread. This spread is 12.51 logits less than the participants' medical Spanish speaking ability spread. The separation and reliability statistics of the linguistic scoring criteria are 3.25 and 0.91 ($\chi^2 = 58.60$, $df = 5$, $p < 0.001$), respectively, suggesting that the linguistic scoring criteria formed slightly less than four statistically distinct difficulty levels.

The infit and outfit MNSQ coefficients are between 0.64 (Coherence criterion) and 1.49 (Effectiveness criterion), indicating a relatively acceptable consistency in the functionality of the linguistic scoring criteria. The standard errors of the linguistic scoring criteria's difficulty measures are low (0.17 and 0.18), which suggests good precision in the difficulty estimates. There are some differences between the observed averages and the fair average measures (FAM). This suggests that setting the effect of the other facets to zero would have an effect on linguistic scoring criteria difficulty estimates (Aryadoust, 2016; Aryadoust & Liu, 2015).

Table 8

Linguistic Scoring Criteria Difficulty Measures (logits), Fair Average Measures, and Fit Statistics Estimated Based on the Raters' Ratings

Linguistic criteria	Total score	Observed average	FAM	Measure (logits)	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
Intelligibility	533	3.15	3.12	-0.63	0.18	1.04	0.40	0.90	-0.48
Fluency	517	3.06	3.02	-0.13	0.17	1.12	1.05	1.00	0.02
Grammar	505	2.99	2.94	0.23	0.17	1.03	0.28	1.13	0.90
Vocabulary	480	2.84	2.79	0.95	0.17	0.71	-2.84	0.74	-1.87
Coherence	532	3.15	3.11	-0.60	0.18	0.70	-2.69	0.64	-2.07
Effectiveness	507	3.00	2.96	0.17	0.17	1.35	2.78	1.49	2.86
Reliability = 0.91; Separation = 3.25; ($\chi^2 = 58.60$, $df = 5$, $p < 0.001$)									

Note. FAM = Fair average measures. MNSQ = Mean square. SE = Standard error. ZSTD = z-standardized.

Table 9 shows values related to the functionality of the scoring rubric scale. As it can be seen, 11% (95) of the responses are in category 1 (Inadequate), 20% (182) are in category 2, 37% (333) are in category 3 (Good), and 32% (404) are in category 4 (Very good). An average measure that corresponds to the mean participant evaluation score was calculated for each scale category. An increase from -4.07 to +5.08 is shown in these average measures. This increase is monotonic which supports the notion that higher scale categories relate to higher medical Spanish speaking proficiency. The expected measures column displays the average measures if the data perfectly fit the Rasch model. The MNSQ coefficients of 1.10 for categories 1 and 2, and 0.90 for categories 3 and 4 are within the recommended range, indicating a good fit. In addition, the Rasch-Andrich thresholds between adjacent scale categories display a monotonic trend increasing from -3.44 between categories 1 and 2, to -0.38 between categories 2 and 3, to 3.83 between categories 3 and 4.

Figure 9 displays the MFRM's predicted probability of observing each category of the scoring rating scale. The x-axis is the latent variable (i.e., medical Spanish speaking ability), drawn relative to the difficulty of the item, in this case the ratings for the SRLT. The y-axis is the predicted probability of observing each category. As can be seen, as one moves along the x-axis from left to right in increasing levels of the latent variable, all the categories appear in the expected order and correspond to a distinct range of item difficulty. This suggests that the SRLT scoring rubric scale meets the criteria for rating scale functionality (Bond et al., 2021, pp. 222-237; Linacre, 2025).

According to these results, the MFRM suggests that the SRLT tasks and ratings functioned appropriately to create a ruler-like metric of medical Spanish speaking ability along which participants could be located.

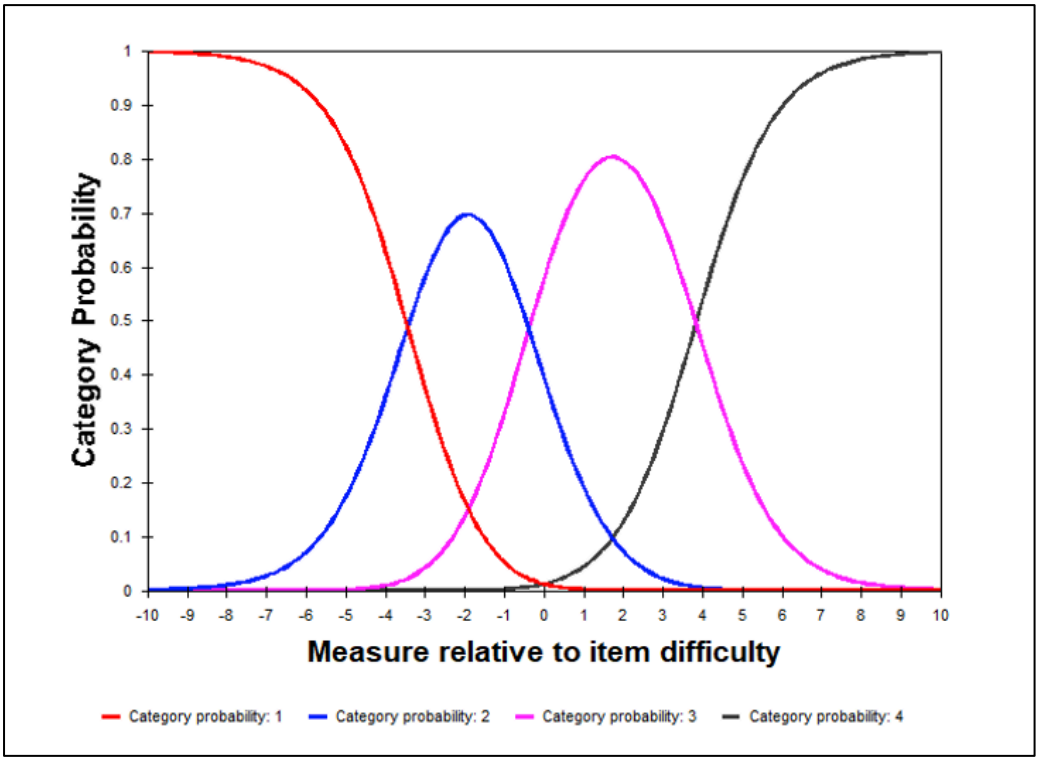
Table 9

Scoring Rubric Scale Functionality

Scoring category	Total	Percent	Average measure	Expected measure	Outfit MNSQ	Rasch-Andrich thresholds
1 (Inadequate)	95	11%	-4.07	-4.12	1.10	NA
2 (Minimal)	182	20%	-1.43	-1.43	1.10	-3.44
3 (Good)	333	37%	2.00	2.05	0.90	-0.38
4 (Very good)	404	32%	5.08	5.04	0.90	3.83

Figure 9

Category Probability Curves for the Scoring Rubric Scale



4.2 The extrapolation inference of the CanopyCredential exam

4.2.1 Relationship between participants' scores in the CCE and the SRLT

The data used in this analysis were participants' scores in the speaking section of the Spanish module of the CCE, and their overall scores in the SRLT. Participants in this study received medical Spanish speaking scores in the CCE ranging from A2 to C1, which covers the entire range of medical Spanish speaking proficiency levels that the CCE measures. However, there were fewer participants at the lowest proficiency levels (i.e., A2 and Low B1), and the C1 level was the one with more participants. Table 10 displays the number of participants at each CCE medical Spanish speaking proficiency level.

Table 10

Participants by CCE Medical Spanish Speaking Proficiency Level

CCE scores	*n*
A2	3
Low B1	5
High B1	13
Low B2	12
High B2	15
C1	34

The SRLT overall scores were the logit measures obtained from the MFRM analysis that was used to examine the SRLT and that was described in the previous section. As mentioned before, these logit measures are an estimate of participants' medical Spanish speaking ability and were generated while taking into account all the facets included in the MFRM (i.e., participants ability, raters severity/leniency, linguistic scoring criteria difficulty, and scoring rubric scale difficulty). Table 11 displays the descriptive statistics for participants scores (log-odds) in the

SRLT. As it can be seen, participants' logit scores in the SRLT ranged from -5.79 to 8.30 logits (M = 2.35, SD = 3.72).

Table 11

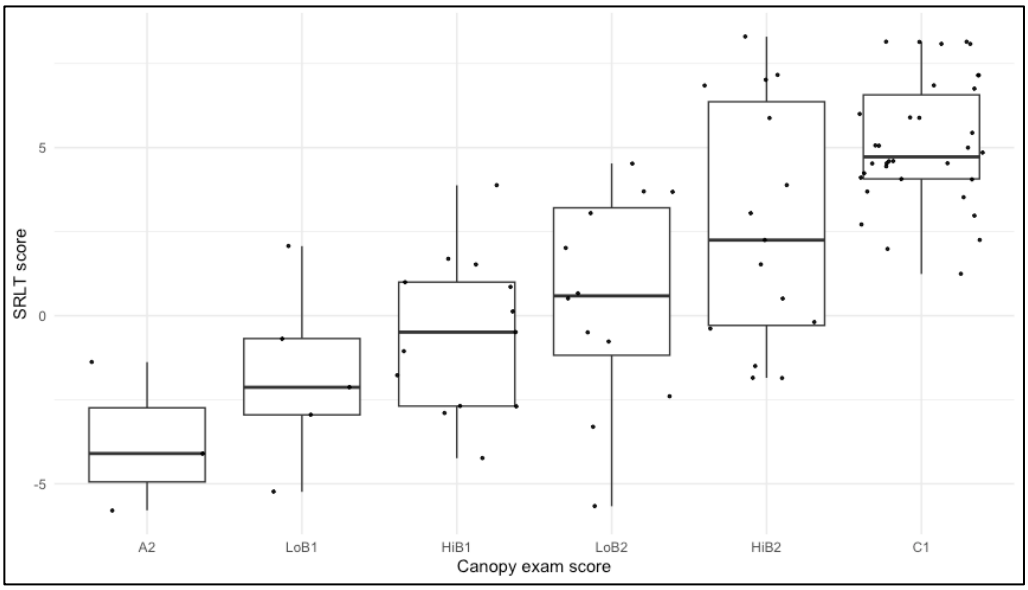
Descriptive Statistics for Participants' Scores (log-odds) in the SRLT

N	Mean	*SD*	Median	Min.	Max.
82	2.35	3.72	3.01	-5.79	8.30

Figure 10 shows a boxplot depicting the relationship between participants' medical Spanish speaking scores in the CCE and their overall scores in the SRLT. The small dots represent the number of participants at each CCE proficiency level. As shown in this figure, there seems to be a positive relationship between the variables suggesting that participants who achieved higher scores in the SRLT also attained higher CCE medical Spanish speaking scores.

Figure 10

Relationship between Participants Medical Spanish Speaking Scores in the CCE and the SRLT



To further explore this relationship, an ordinal regression model was fit using R statistical software, version 4.4.2 (R Core Team, 2024), and the package ordinal, version 2023.12-4.1 (Christensen, 2023). The dependent variable was participants' CCE medical Spanish speaking scores, and the independent variable was participants' score in the SRLT. The analysis used a logit link function, and the fitted model was compared to a null model with no predictor using a likelihood ratio test.

The comparison indicated that the model fit was statistically significantly better for the model with predictors than the null model, $X^2(1) = 54.94$, $p < 0.001$. The Nagelkerke's R^2 coefficient ($R^2 = 0.51$) suggests a moderate relationship between the predictor and CCE medical Spanish speaking scores. The proportional odds assumption was evaluated with a likelihood ratio test for the predictor using the nominal_test function from the R ordinal package. The test was non-significant, which indicates that the assumption of proportional odds was met.

Focusing on the predictor, the results indicated that SRLT scores were a significant factor, $\beta = 0.50$, $SE = 0.08$, $z = 6.33$, $p < 0.001$, *99% CI* [0.31, 0.71], with a positive relationship. This suggests that higher SRLT scores were associated with higher CCE medical Spanish speaking scores. That is, for every one logit increase in the SRLT score, the odds of being rated in a higher CCE medical Spanish speaking score category—as opposed to the same or lower category—increased by a factor of 1.65. A summary of the results is displayed in Table 12.

Table 12

Results Summary for the Relationship between Participants Medical Spanish Speaking Scores in the CCE and the SRLT

Predictor	Estimate (β)	*SE*	*z*	*p*	*99% CI* (Lower)	*99% CI* (Upper)
SRLT scores	0.50	0.08	6.33	0.001	0.31	0.71

4.2.2 Relationship between participants' scores in the CCE and linguistic measures of intelligibility

A Spearman's correlation matrix was generated to illustrate the relationship between all the variables included in the analyses for all the linguistic measures reported in this section and the next ones. This matrix was produced using R statistical software, version 4.4.2 (R Core Team, 2024), and the packages Hmisc, version 5.2.3 (Harrell, 2025), psych, version 2.4.12 (Revelle, 2024), flextable, version 0.9.7 (Gohel & Skintzos, 2024), dplyr, version 1.1.4 (Wickham et al., 2023), and officer, version 0.6.8 (Gohel et al., 2025). This correlation matrix can be seen in Appendix B, Table B2.

Prior to conducting the ordinal regression model for the intelligibility criterion, descriptive statistics were obtained for each one of the linguistic measures selected for this criterion (i.e., phoneme pronunciation errors per AS-unit, syllable pronunciation errors per AS-unit, and word stress pronunciation errors per AS-unit). These descriptive statistics were grouped by the six different speaking proficiency scores that participants could receive in the CCE, and are shown in Table 13. As can be seen, phoneme pronunciation errors and syllable pronunciation errors per AS-unit were more common across the six CCE medical Spanish speaking score categories than word stress pronunciation errors per AS-unit. However, no clear trend is observed concerning whether these pronunciation error types increase or decrease across CCE medical Spanish speaking score categories.

Figure 11 displays a boxplot of the relationship between phoneme pronunciation errors per AS-unit in the SRLT, and CCE medical Spanish speaking scores. As can be seen, there seems to be a weak negative relationship between these two variables suggesting that phoneme

pronunciation errors per AS-unit might decrease from the A2 CCE score category to the C1 CCE score category.

Table 13

Descriptive Statistics for the Intelligibility Criterion's Linguistic Measures

Linguistic measure	CCE scores	*n*	Mean	*SD*	Median	Min.	Max.
Pronunciation errors (phoneme) per AS-unit	A2	3	0.13	0.05	0.11	0.09	0.18
	Low B1	5	0.10	0.04	0.10	0.07	0.17
	High B1	13	0.13	0.10	0.09	0.03	0.35
	Low B2	12	0.08	0.10	0.06	0.00	0.35
	High B2	15	0.09	0.07	0.06	0.00	0.25
	C1	34	0.03	0.05	0.02	0.00	0.23
Pronunciation errors (syllable) per AS-unit	A2	3	0.06	0.07	0.05	0.00	0.13
	Low B1	5	0.05	0.04	0.04	0.00	0.11
	High B1	13	0.05	0.03	0.06	0.00	0.12
	Low B2	12	0.04	0.06	0.03	0.00	0.19
	High B2	15	0.01	0.02	0.00	0.00	0.06
	C1	34	0.02	0.02	0.01	0.00	0.08
Pronunciation errors (word stress) per AS-unit	A2	3	0.02	0.03	0.00	0.00	0.06
	Low B1	5	0.00	0.01	0.00	0.00	0.02
	High B1	13	0.01	0.02	0.00	0.00	0.06
	Low B2	12	0.00	0.01	0.00	0.00	0.03
	High B2	15	0.00	0.01	0.00	0.00	0.03
	C1	34	0.00	0.00	0.00	0.00	0.02

Figure 11

Relationship between Pronunciation Errors (Phoneme) per AS-Unit and CCE Scores

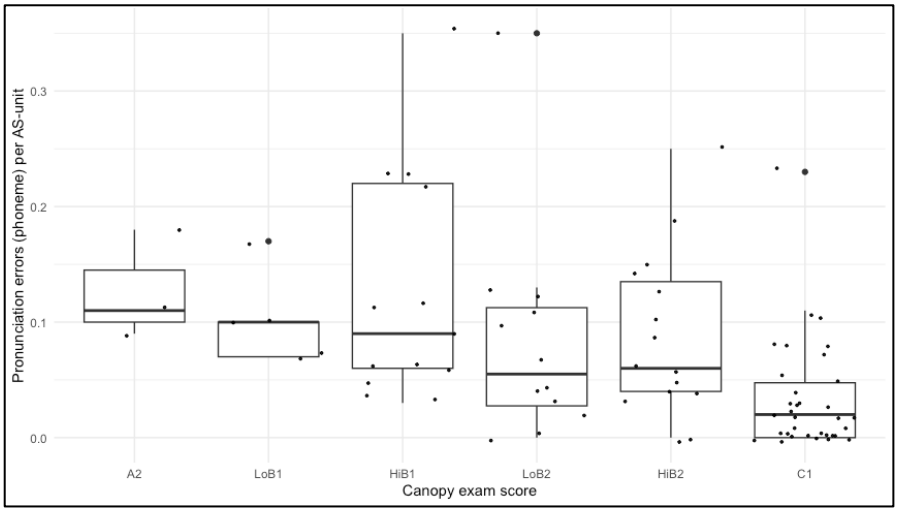


Figure 12 presents a boxplot of the relationship between syllable pronunciation errors per AS-unit in the SRLT, and CCE medical Spanish speaking scores. A similar pattern as the one observed in Figure 11 seems to be present, although there are some discrepancies between the low-B1 and high-B1 CCE score categories, and the high-B2 and C1 CCE score categories. Therefore, this trend is unclear. Regarding the relationship between word stress pronunciation errors per AS-unit in the SRLT and CCE medical Spanish speaking scores, this is displayed in Figure 13. As shown in this figure, no relationship is observed between these two variables. This is potentially due to the fact that there was a large number of participants who did not make word stress pronunciation errors.

Figure 12

Relationship between Pronunciation Errors (Syllable) per AS-unit and CCE Scores

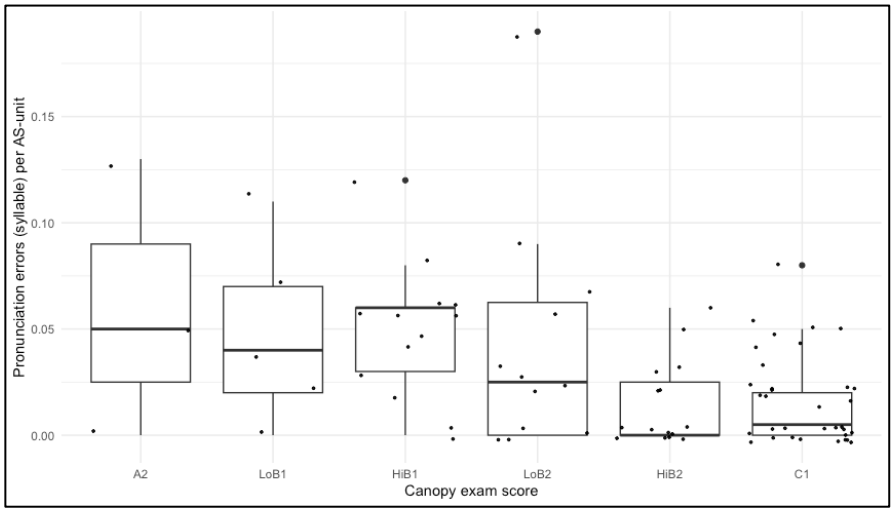
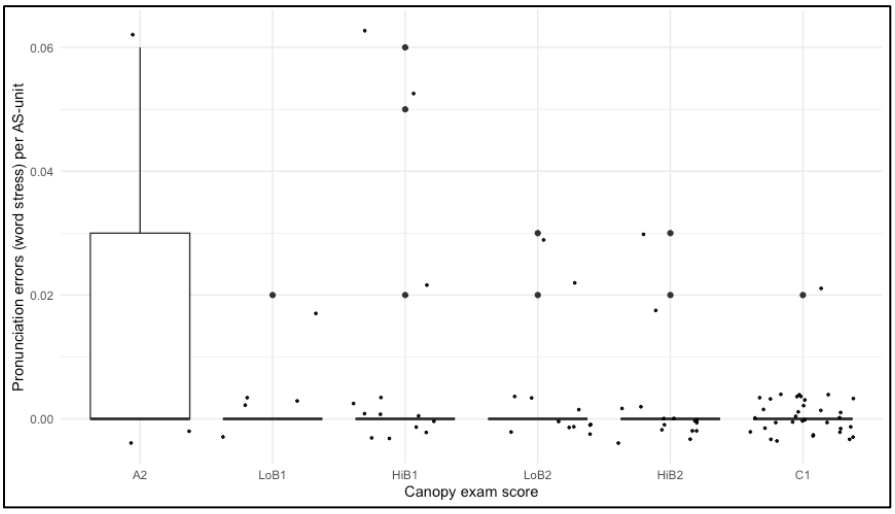


Figure 13

Relationship between Pronunciation Errors (Word Stress) per AS-unit and CCE Scores



To further explore the relationship between these linguistic measures for the intelligibility criterion from participants' production in the SRLT and their medical Spanish speaking scores on the CCE, an ordinal regression model was fit using the same instruments that were used in the previous section. The independent variables were phoneme pronunciation errors per AS-unit, syllable pronunciation errors per AS-unit, and word stress pronunciation errors per AS-unit. The dependent variable was participants' CCE medical Spanish speaking scores. The analysis used a logit link function, and the fitted model was compared to a null model with no predictors using a likelihood ratio test.

The comparison indicated that the model fit was statistically significantly better for the model with predictors than the null model, $X^2(3) = 26.87$, $p < 0.001$. The Nagelkerke's R^2 coefficient ($R^2 = 0.29$) indicates a weak relationship between the predictors and CCE medical Spanish speaking scores. The proportional odds assumption was evaluated with a likelihood ratio

test for the predictors using the nominal_test function from the R ordinal package. The tests were non-significant, which indicates that the assumption of proportional odds was met.

In terms of the individual predictors, phoneme pronunciation errors per AS-unit was a significant factor with a negative relationship, β = -7.15, *SE* = 2.83, z = -2.52, p = 0.01, *99% CI* [-14.93, -0.12]. Syllable pronunciation errors per AS-unit was also a significant factor with a negative relationship, β = -16.90, *SE* = 6.52, z = -2.59, p = 0.01, *99% CI* [-34.76, -1.02]. This suggests that higher production of phoneme pronunciation errors per AS-unit, and higher production of syllable pronunciation errors per AS-unit was associated with lower CCE medical Spanish speaking scores. Word stress pronunciation errors per AS-unit was not found to be a significant factor, β = -32.49, *SE* = 16.26, z = -2.00, p = 0.05, *99% CI* [-77.14, 10.23]. These results are summarized in Table 14.

Table 14

Results Summary for the Relationship between Linguistic Measures of Intelligibility and CCE Scores

Predictor	Estimate (β)	*SE*	*z*	*p*	*99% CI* (Lower)	*99% CI* (Upper)
Pronunciation errors (phoneme) per AS-unit	-7.15	2.83	-2.52	0.01	-14.93	-0.12
Pronunciation errors (syllable) per AS-unit	-16.90	6.52	-2.59	0.01	-34.76	-1.02
Pronunciation errors (word stress) per AS-unit	-32.49	16.26	-2.00	0.05	-77.14	10.23

4.2.3 Relationship between participants' scores in the CCE and the linguistic measure of fluency

The linguistic measure selected for the fluency criterion was pauses per AS-unit. Table 15 shows the descriptive statistics for this measure, grouped by CCE medical Spanish speaking scores. Figure 14 displays a boxplot of the relationship between pauses per AS-unit and CCE medical

Spanish speaking scores. As can be seen, it seems that there is a negative relationship between these two variables, and that pauses per AS-unit decrease from the A2 CCE score category to the C1 CCE score category.

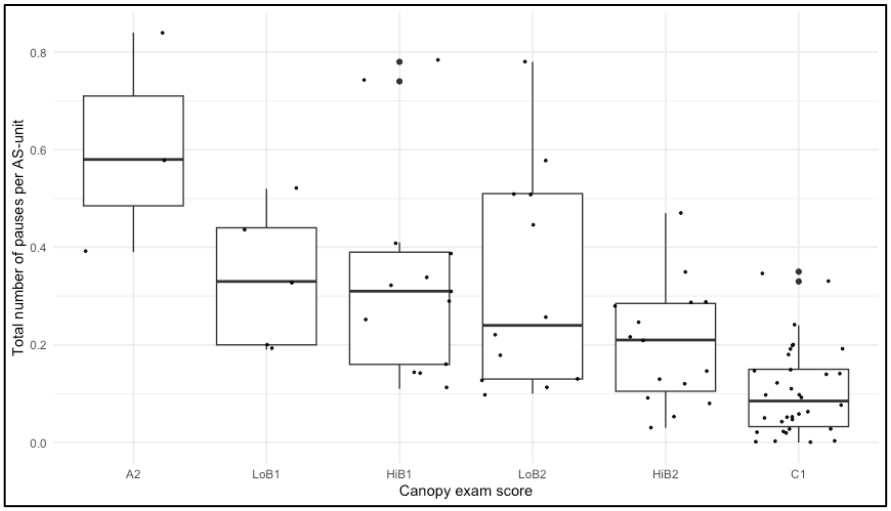
Table 15

Descriptive Statistics for the Fluency Criterion's Linguistic Measure

Linguistic measure	CCE scores	*n*	Mean	*SD*	Median	Min.	Max.
Pauses per AS-unit	A2	3	0.60	0.23	0.58	0.39	0.84
	Low B1	5	0.34	0.15	0.33	0.19	0.52
	High B1	13	0.34	0.21	0.31	0.11	0.78
	Low B2	12	0.33	0.23	0.24	0.10	0.78
	High B2	15	0.20	0.12	0.21	0.03	0.47
	C1	34	0.10	0.09	0.09	0.00	0.35

Figure 14

Relationship between Pauses per AS-unit and CCE Scores



To further explore the relationship between pauses per AS-unit and CCE medical Spanish speaking scores, an ordinal regression model was fit using the same software and package that was used for the two previous ordinal regression analyses. The dependent variable was participants'

CCE medical Spanish speaking scores and the independent variable was pauses per AS-unit. The analysis used a logit link function.

The fitted model was compared to a null model with no predictors using a likelihood ratio test. The comparison indicated that the model fit was statistically significantly better for the model with predictors than the null model, $X^2(1) = 34.43$, $p < 0.001$. The Nagelkerke's R^2 value ($R^2 = 0.36$) indicates a weak relationship between the predictor and CCE medical Spanish speaking scores. The proportional odds assumption was evaluated with a likelihood ratio test for the predictor using the nominal_test function from the R ordinal package. The test was non-significant, which indicates that the assumption of proportional odds was met. Looking at the predictor, pauses per AS-unit was found to be a significant negative factor, $\beta = -6.70$, $SE = 1.27$, $z = -5.30$, $p < 0.001$, *99% CI* [-10.22, -3.63], with a negative relationship. This suggests that higher production of pauses per AS-unit was associated with lower CCE medical Spanish speaking scores. These results are summarized in Table 16.

Table 16

Results Summary for the Relationship between the Linguistic Measure of Fluency and CCE Scores

Predictor	Estimate (β)	*SE*	*z*	*p*	*99% CI* (Lower)	*99% CI* (Upper)
Pauses per AS-unit	-6.70	1.27	-5.30	0.001	-10.22	-3.63

4.2.4 Relationship between participants' scores in the CCE and linguistic measures of grammar

For the grammar criterion, there were two different sets of linguistic measures. One set had a focus on grammatical complexity, and the linguistic measures that were selected for it were clauses, clitic pronouns, and subjunctive tense forms per AS-unit. The other set had a focus on accuracy

and the linguistic measures that were selected for it were lexical, morphological, and syntactic errors per AS-unit.

Beginning with the set focusing on grammatical complexity, Table 17 shows the descriptive statistics for each one of the linguistics measures selected (i.e., clauses per AS-unit, clitic pronouns per AS-unit, and subjunctive tense forms per AS-unit), grouped by CCE medical Spanish speaking proficiency scores. As can be seen on this table, production of clauses per AS-unit was more common across the six CCE medical Spanish speaking categories than production of clitic pronouns and subjunctive tense forms.

Table 17

Descriptive Statistics for the Grammar Criterion's Linguistic Complexity Measures

Linguistic measure	CCE scores	*n*	Mean	*SD*	Median	Min.	Max.
Clauses per AS-unit	A2	3	1.49	0.17	1.48	1.32	1.66
	Low B1	5	1.47	0.27	1.44	1.15	1.89
	High B1	13	1.68	0.46	1.59	1.18	2.94
	Low B2	12	1.72	0.45	1.71	1.21	2.81
	High B2	15	1.76	0.47	1.56	1.31	2.85
	C1	34	2.03	0.47	1.89	1.22	2.85
Clitic pronouns per AS-unit	A2	3	0.00	0.00	0.00	0.00	0.00
	Low B1	5	0.00	0.00	0.00	0.00	0.00
	High B1	13	0.04	0.06	0.00	0.00	0.16
	Low B2	12	0.05	0.05	0.04	0.00	0.13
	High B2	15	0.06	0.06	0.04	0.00	0.17
	C1	34	0.10	0.07	0.09	0.00	0.23
Subjunctive tense forms per AS-unit	A2	3	0.00	0.00	0.00	0.00	0.00
	Low B1	5	0.01	0.02	0.00	0.00	0.04
	High B1	13	0.03	0.03	0.03	0.00	0.11
	Low B2	12	0.03	0.03	0.00	0.00	0.07
	High B2	15	0.03	0.04	0.02	0.00	0.14
	C1	34	0.10	0.09	0.08	0.00	0.26

Figures 15 through 17 below show boxplots of the relationship between each one of these linguistic measures and CCE medical Spanish speaking score categories. Regarding the relationship between clauses per AS-unit and CCE scores (Figure 15), it seems that there is an

inconsistent trend although it might point towards a weak positive relationship. As per the clitic pronouns per AS-unit measure (Figure 16), it appears that there is also a weak positive relationship slightly more visible in the three highest CCE medical Spanish speaking scores categories (i.e., low-B2, high-B2, C1). This might potentially be due to participants in the A2 and low-B1 CCE medical Spanish speaking score categories not producing too many clitic pronouns. Figure 17 seems to display an inconsistent trend for the relationship between subjunctive tense forms per AS-unit and CCE medical Spanish speaking scores as well. This trend might point towards a positive relationship between the variables, and it is slightly more visible in the highest CCE medical Spanish speaking score categories. As in the previous case, this could also be potentially due to participants in the lower CCE medical Spanish speaking score categories not producing too many of these forms.

Figure 15

Relationship between Clauses per AS-unit and CCE Scores

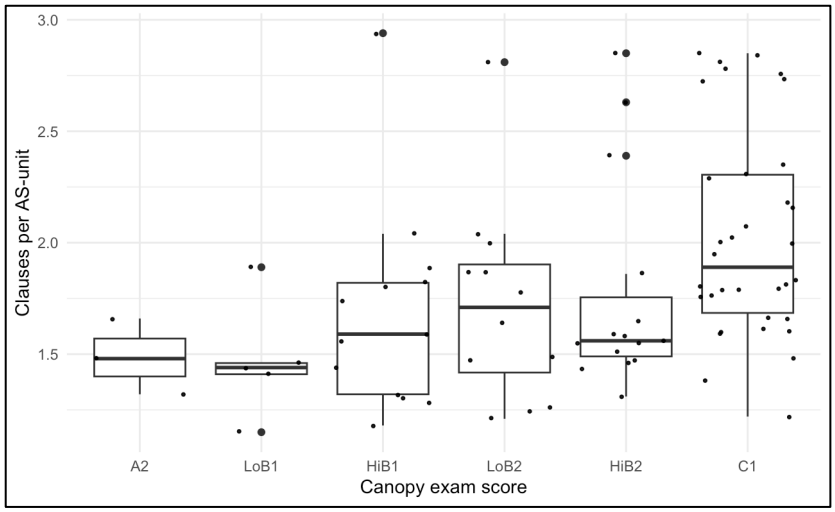


Figure 16

Relationship between Clitic Pronouns per AS-unit and CCE Scores

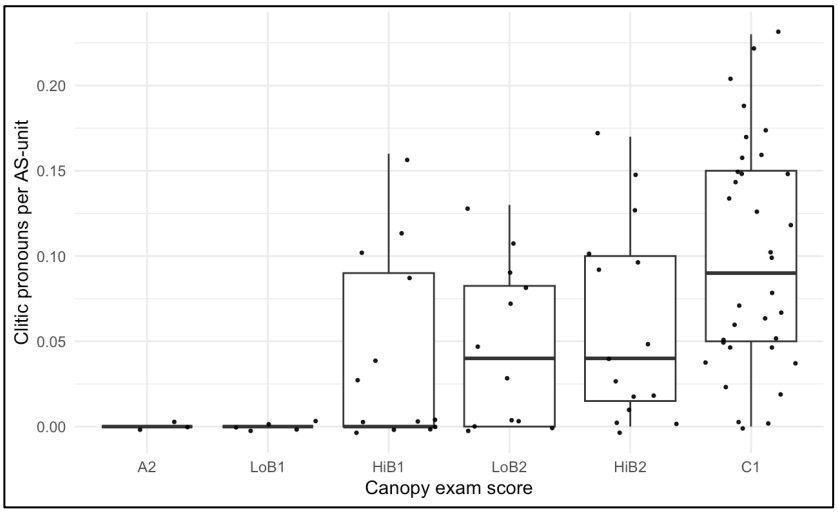
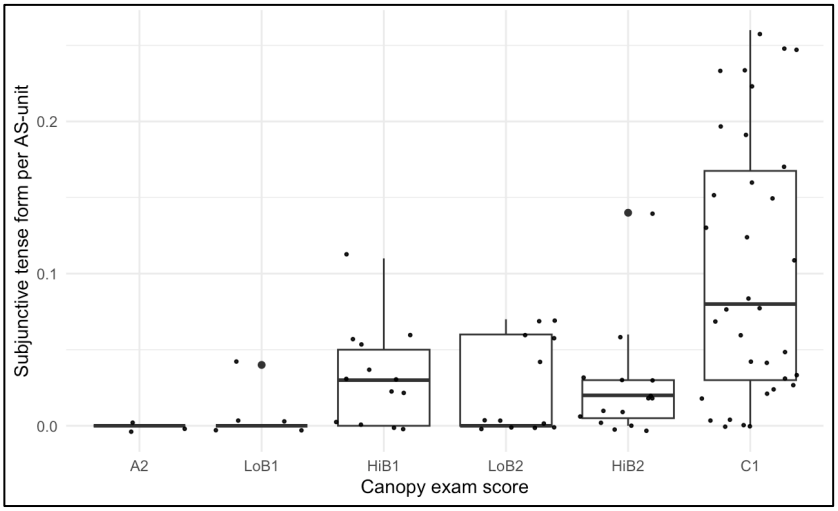


Figure 17

Relationship between Subjunctive Tense Forms per AS-unit and CCE Scores



To further explore the relationship between this set of linguistic complexity measures for the grammar criterion and participants' medical Spanish speaking scores on the CCE, an ordinal regression model was fit using the same instruments and procedures as in the previous linguistic

measures examined before. The predictors for this model were clauses per AS-unit, clitic pronouns per AS-unit, and subjunctive tense forms per AS-unit. The dependent variable was CCE medical Spanish speaking scores. This analysis also used a logit link function.

A likelihood ratio test was used to compare the fit model to a null model with no predictors. The comparison indicated that the model fit was statistically significantly better for the model with predictors than the null model, $X^2(3) = 42.89$, $p < 0.001$. The Nagelkerke's R^2 value ($R^2 = 0.43$) suggests a weak to moderate relationship between the predictors and CCE medical Spanish speaking scores. To evaluate the proportional odds assumption, likelihood ratio tests for the predictors using the nominal_test function from the R ordinal package were employed. The results from these tests were non-significant, indicating that the assumption of proportional odds was met.

Concerning the individual predictors, the results showed that subjunctive tense forms per AS-unit was the only significant factor and the relationship was positive, $\beta = 17.78$, $SE = 5.35$, $z = 3.32$, $p = 0.001$, *99% CI* [5.81, 33.88]. This suggests that higher production of subjunctive tense forms per AS-unit was associated with higher CCE medical Spanish speaking scores. Clauses per AS-unit was not found to be a significant factor, $\beta = 1.00$, $SE = 0.55$, $z = 1.82$, $p = 0.07$, *99% CI* [-0.37, 2.49]. Equally, clitic pronouns per AS-unit was not found to be a significant factor either, $\beta = 10.17$, $SE = 4.39$, $z = 2.32$, $p = 0.02$, *99% CI* [-0.86, 22.08]. Table 18 presents a summary of these results.

Table 18

Results Summary for the Relationship between Linguistics Measures of Grammar Complexity and CCE Scores

Predictor	Estimate (β)	*SE*	*z*	*p*	*99% CI* (Lower)	*99% CI* (Upper)
Clauses per AS-unit	1.00	0.55	1.82	0.07	-0.37	2.49
Clitic pronouns per AS-unit	10.17	4.39	2.32	0.02	-0.86	22.08
Subjunctive tense forms per AS-unit	17.78	5.35	3.32	0.001	5.81	33.88

The next analysis focused on the linguistic accuracy measures for the grammar criterion. These measures were lexical errors per AS-unit, morphological errors per AS-unit, and syntactic errors per AS-unit. Table 19 presents the descriptive statistics for each one of these linguistics measures, grouped by CCE medical Spanish speaking proficiency scores. As can be seen on this table, production of lexical errors per AS-unit and morphological errors per AS-unit was more common than production of syntactic errors per AS-unit.

Figures 18 through 20 below show boxplots of the relationship between each one of these grammar error types and CCE medical Spanish speaking score categories. When looking at Figure 18, there seems to be a negative relationship between production of lexical errors per AS-unit and CCE medical Spanish speaking scores although a discrepancy can be observed between the low-B1 and high-B1 CCE medical Spanish speaking scores. Figure 19 also appears to display a negative relationship between morphological errors per AS-unit and CCE medical Spanish speaking scores. Regarding the relationship between syntactic errors per AS-unit and CCE medical Spanish speaking scores (Figure 20), it seems that there is no clear trend potentially due to not too many participants producing this type of grammar error.

Table 19

Descriptive Statistics for the Grammar Criterion's Linguistic Accuracy Measures

Linguistic measure	CCE scores	*n*	Mean	*SD*	Median	Min.	Max.
Lexical errors per AS-unit	A2	3	0.48	0.18	0.58	0.27	0.59
	Low B1	5	0.19	0.13	0.13	0.07	0.37
	High B1	13	0.25	0.20	0.20	0.06	0.68
	Low B2	12	0.14	0.08	0.14	0.02	0.30
	High B2	15	0.13	0.10	0.08	0.00	0.28
	C1	34	0.03	0.05	0.01	0.00	0.28
Morphological errors per AS-unit	A2	3	0.42	0.07	0.44	0.34	0.48
	Low B1	5	0.43	0.13	0.39	0.26	0.59
	High B1	13	0.25	0.10	0.25	0.09	0.43
	Low B2	12	0.23	0.13	0.20	0.06	0.47
	High B2	15	0.22	0.14	0.15	0.08	0.52
	C1	34	0.06	0.06	0.04	0.00	0.26
Syntactic errors per AS-unit	A2	3	0.02	0.02	0.03	0.00	0.03
	Low B1	5	0.02	0.02	0.00	0.00	0.05
	High B1	13	0.01	0.01	0.00	0.00	0.04
	Low B2	12	0.01	0.02	0.00	0.00	0.05
	High B2	15	0.01	0.02	0.00	0.00	0.09
	C1	34	0.01	0.01	0.00	0.00	0.06

Figure 18

Relationship between Lexical Errors per AS-unit and CCE Scores

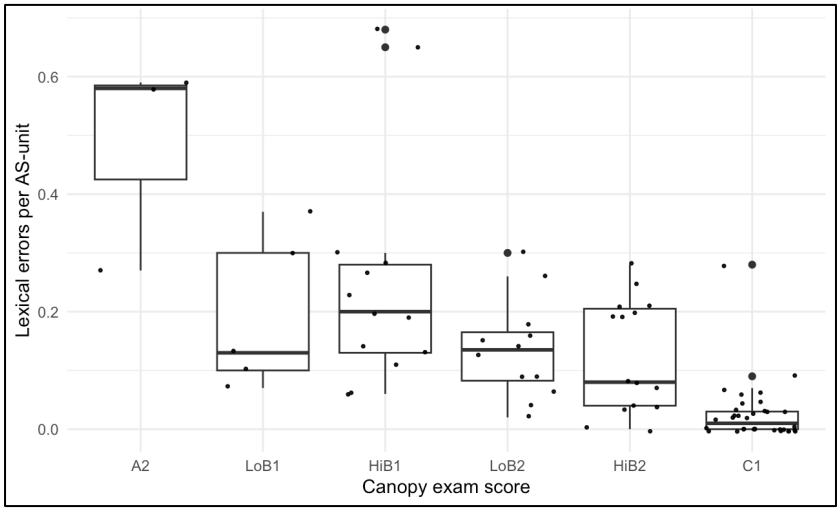


Figure 19

Relationship between Morphological Errors per AS-unit and CCE Scores

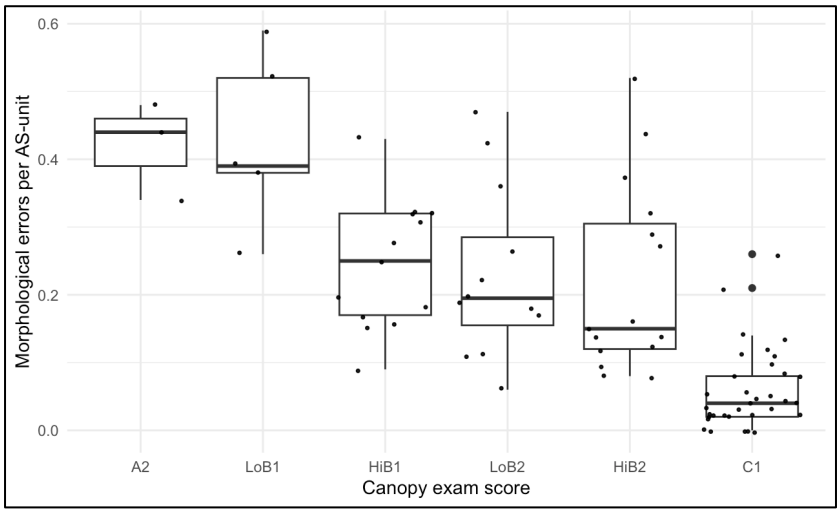
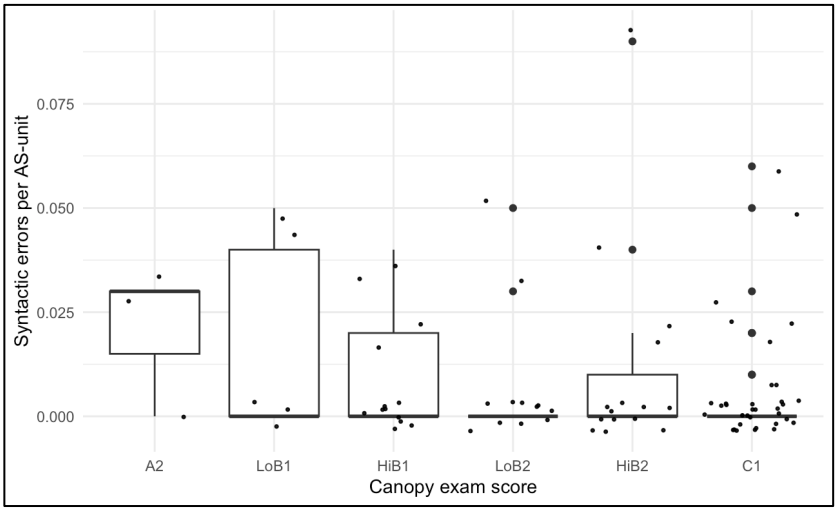


Figure 20

Relationship between Syntactic Errors per AS-unit and CCE Scores



Using the same software and procedures that were used in all the linguistic measures examined above, an ordinal regression model was fit to explore the relationship between the set of linguistic accuracy measures selected for the grammar criterion, and participants medical Spanish

speaking scores on the CCE. The predictors for this model were lexical errors per AS-unit, morphological errors per AS-unit, and syntactic errors per AS-unit. The dependent variable was CCE medical Spanish speaking scores.

The fit model was compared to a null model with no predictors using a likelihood ratio test. The comparison indicated that the model fit was statistically significantly better for the model with predictors than the null model, $X^2(3) = 69.34$, $p < 0.001$. The Nagelkerke's R^2 coefficient ($R^2 = 0.60$) indicate a moderate to considerable relationship between the predictors and CCE medical Spanish speaking scores. The proportional odds assumption was evaluated using the nominal_test function from R ordinal package. The results showed that none of the test was significant, indicating that the assumption of proportional odds was met.

In regard to the individual predictors, the results showed that lexical errors per AS-unit was found to be a significant factor, $\beta = -7.22$, $SE = 2.19$, $z = -3.30$, $p = 0.001$, *99% CI* [-13.29, -1.90], with a negative relationship. This suggests that higher production of lexical errors per AS-unit was associated with lower CCE medical Spanish speaking scores. Additionally, morphological errors per AS-unit was also found to be a significant factor with a negative relationship, $\beta = -10.05$, $SE = 2.23$, $z = -4.51$, $p < 0.001$, *99% CI* [-16.16, -4.59]. This suggests that higher production of morphological errors per AS-unit was also associated with lower CCE medical Spanish speaking scores. As for syntactic errors per AS-unit, the results showed that this measure was not found to be a significant factor, $\beta = 26.18$, $SE = 13.75$, $z = 1.91$, $p = 0.06$, *99% CI* [-9.46, 63.57]. Table 20 summarizes these results.

Table 20

Results Summary for the Relationship between Linguistic Measures of Grammar Accuracy and CCE Scores

Predictor	Estimate (β)	*SE*	*z*	*p*	*99% CI* (Lower)	*99% CI* (Upper)
Lexical errors per AS-unit	-7.22	2.19	-3.30	0.001	-13.29	-1.90
Morphological errors per AS-unit	-10.05	2.23	-4.51	0.001	-16.16	-4.59
Syntactic errors per AS-unit	26.18	13.75	1.91	0.06	-9.46	63.57

4.2.5 Relationship between participants' scores in the CCE and the linguistic measure of vocabulary

The linguistic measure selected for the vocabulary criterion was medical terms per AS-unit. Table 21 shows the descriptive statistics for this linguistic measure, grouped by participants' medical Spanish speaking proficiency scores in the CCE. As can be seen in this table, there is variability across CCE medical Spanish speaking score categories in the use of medical terms, and no clear trend is observed. Figure 21 below shows a boxplot of the relationship between medical terms per AS-unit and CCE medical Spanish speaking proficiency score categories. As shown, there is no clear trend towards a positive or negative relationship between these variables.

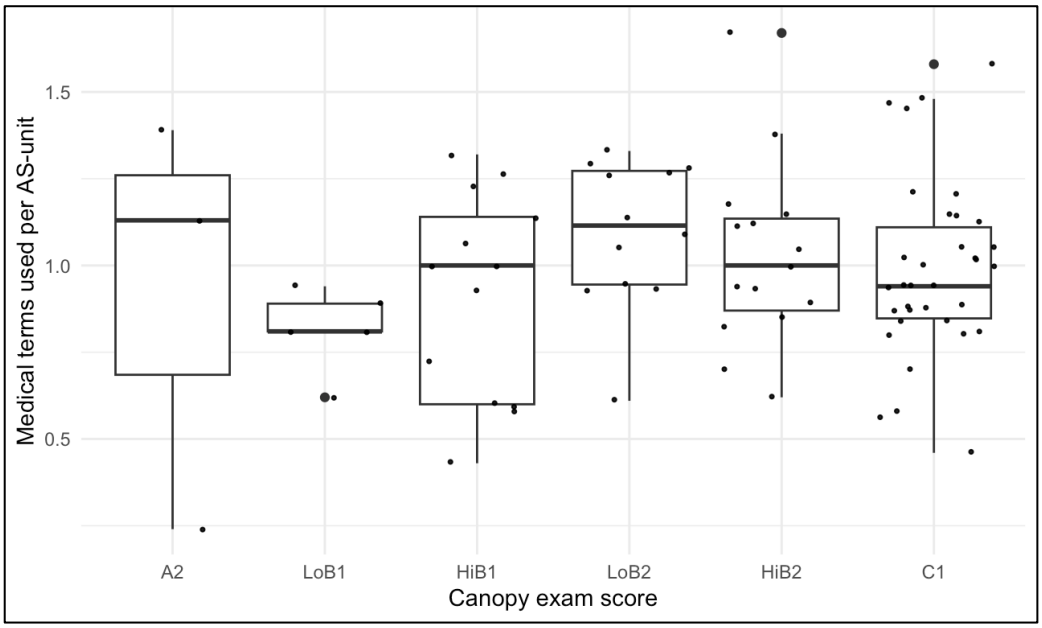
Table 21

Descriptive Statistics for the Vocabulary Criterion's Linguistic Measure

Linguistic measure	CCE scores	*n*	Mean	*SD*	Median	Min	Max
Medical terms per AS-unit	A2	3	0.92	0.60	1.13	0.24	1.39
	Low B1	5	0.81	0.12	0.81	0.62	0.94
	High B1	13	0.91	0.30	1.00	0.43	1.32
	Low B2	12	1.09	0.21	1.12	0.61	1.33
	High B2	15	1.03	0.26	1.00	0.62	1.67
	C1	34	0.99	0.26	0.94	0.46	1.58

Figure 21

Relationship between Medical Terms per AS-unit and CCE Scores



An ordinal regression model was fit to examine the relationship between participants' use of medical terms per AS-unit and their medical Spanish speaking scores on the CCE. The predictor for this model was medical terms per AS-unit, and the dependent variable was CCE medical Spanish speaking scores. This fit model was compared to a null model with no predictors using a likelihood ratio test. The comparison indicated that the model fit was not statistically significant, $X^2(1) = 0.74$, $p = 0.39$. Thus, medical terms per AS-unit was not a significant factor. Table 22 displays these results.

Table 22

Results Summary for the Relationship between the Linguistic Measure of Vocabulary and CCE Scores

Predictor	Estimate (β)	*SE*	*z*	*p*	*99% CI* (Lower)	*99% CI* (Upper)
Medical terms per AS-unit	0.67	0.78	0.86	0.39	-1.36	2.71

4.2.6 Relationship between participants' scores in the CCE and linguistic measures of coherence

The linguistic measures selected for the coherence criterion were coordinating and subordinating conjunctions per AS-unit. Table 23 shows the descriptive statistics for these linguistic measures, grouped by participants' medical Spanish speaking proficiency score categories on the CCE. As can be seen, these two types of conjunctions were produced across the six CCE medical Spanish speaking score categories, and an increasing trend from the lowest to the highest CCE score categories can be observed.

Figures 22 displays a boxplot of the relationship between coordinating conjunctions per AS-unit and CCE medical Spanish speaking score categories. As this figure shows, there is no clear trend towards a positive or negative relationship between coordinating conjunctions per AS-unit and CCE medical Spanish speaking scores. Figure 23 displays the relationship between subordinating conjunctions per AS-unit and CCE medical Spanish speaking score categories. As can be seen, the pattern is similar as the one shown in Figure 22 and no clear trend is observed between these two variables either.

Table 23

Descriptive Statistics for the Coherence Criterion's Linguistic Measures

Linguistic measure	CCE scores	*n*	Mean	*SD*	Median	Min.	Max.
Coordinating conjunctions per AS-unit	A2	3	0.15	0.06	0.12	0.11	0.22
	Low B1	5	0.14	0.07	0.15	0.04	0.21
	High B1	13	0.20	0.10	0.17	0.04	0.38
	Low B2	12	0.20	0.08	0.19	0.08	0.34
	High B2	15	0.21	0.08	0.19	0.08	0.34
	C1	34	0.25	0.13	0.24	0.00	0.64
Subordinating conjunctions per AS-unit	A2	3	0.09	0.08	0.13	0.00	0.15
	Low B1	5	0.10	0.06	0.11	0.00	0.16
	High B1	13	0.14	0.06	0.14	0.03	0.26
	Low B2	12	0.19	0.13	0.17	0.00	0.43
	High B2	15	0.19	0.10	0.16	0.06	0.43
	C1	34	0.21	0.09	0.21	0.02	0.40

Figure 22

Relationship between Coordinating Conjunctions per AS-unit and CCE Scores

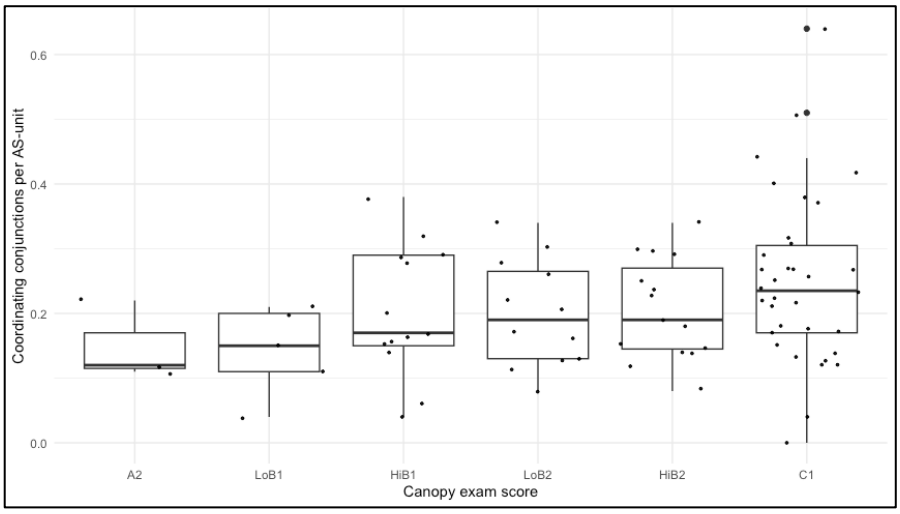
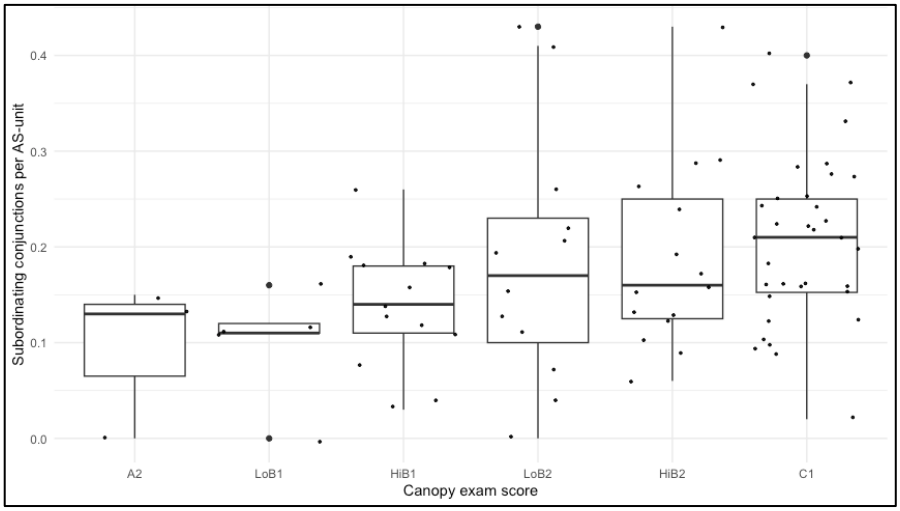


Figure 23

Relationship between Subordinating Conjunctions per AS-unit and CCE Scores



To further examine the relationship between participants' production of coordinating and subordinating conjunctions per AS-unit and their medical Spanish speaking scores on the CCE, an ordinal regression model was fit. The predictors for this model were coordinating conjunctions per

AS-unit, and subordinating conjunctions per AS-unit. The dependent variable was CCE medical Spanish speaking scores. This fit model was compared to a null model with no predictors using a likelihood ratio test. The comparison indicated that the model fit was statistically significantly better for the model with predictors than the null model, $X^2(2) = 11.87$, $p = 0.003$. The Nagelkerke's R^2 value ($R^2 = 0.14$) indicated a weak relationship between these predictors and CCE medical Spanish speaking scores. The nominal_test function from R ordinal package was used to evaluate the proportional odds assumption. The results showed that none of the tests were significant, which indicates that the assumption of proportional odds was met.

Looking at the individual predictors, the results showed that coordinating conjunctions per AS-unit was not found to be a significant factor, $\beta = 3.72$, $SE = 2.25$, $z = 1.65$, $p = 0.10$, *99% CI* [-1.86, 9.90]. Subordinating conjunctions per AS-unit was also not found to be a significant factor, $\beta = 4.90$, $SE = 2.42$, $z = 2.02$, $p = 0.04$, *99% CI* [-1.08, 11.51]. This suggests that production of coordinating conjunctions per AS-unit, and production of subordinating conjunctions per AS-unit was not associated with higher CCE medical Spanish speaking scores. These results are summarized on Table 24.

Table 24

Results Summary for the Relationship between Linguistic Measures of Coherence and CCE Scores

Predictor	Estimate (β)	*SE*	*z*	*p*	*99% CI* (Lower)	*99% CI* (Upper)
Coordinating conjunctions per AS-unit	3.72	2.25	1.65	0.10	-1.86	9.90
Subordinating conjunctions per AS-unit	4.90	2.42	2.02	0.04	-1.08	11.51

4.2.7 Relationship between participants' scores in the CCE and linguistic measures of effectiveness

The linguistic measures selected for the effectiveness criterion were explained medical terms by definition per AS-unit, explained medical terms by synonym per AS-unit, and explained medical terms by context per AS-unit. Table 25 displays the descriptive statistics for these linguistic measures, grouped by participants' medical Spanish speaking proficiency scores on the CCE. As can be seen, production of explained medical terms by definition per AS-unit occurred across the six CCE medical Spanish speaking score categories whereas the other two measures occurred mainly in the higher CCE score categories.

Figure 24 shows the relationship between explained medical terms by definition and CCE medical Spanish speaking score categories. As can be seen, no clear trend is observed between these two variables. The relationship between explained medical terms by synonym per AS-unit and CCE medical Spanish speaking score categories is shown in Figure 25. In this case, no clear trend can be observed either. Figure 26 presents the relationship between explained medical terms by context and CCE medical Spanish speaking score categories. As was the case in the two previous figures, no clear trend is observed. This could potentially be due to a low production of these measures across participants/CCE medical Spanish speaking score categories.

Table 25

Descriptive Statistics for the Effectiveness Criterion's Linguistic Measures

Linguistic measure	CCE scores	*n*	Mean	*SD*	Median	Min.	Max.
Explained medical terms (definition) per AS-unit	A2	3	0.00	0.00	0.00	0.00	0.00
	Low B1	5	0.02	0.01	0.02	0.02	0.04
	High B1	13	0.02	0.03	0.02	0.00	0.08
	Low B2	12	0.03	0.02	0.03	0.00	0.05
	High B2	15	0.02	0.01	0.02	0.00	0.05
	C1	34	0.02	0.02	0.02	0.00	0.08
Explained medical terms (synonym) per AS-unit	A2	3	0.00	0.00	0.00	0.00	0.00
	Low B1	5	0.00	0.01	0.00	0.00	0.02
	High B1	13	0.00	0.01	0.00	0.00	0.02
	Low B2	12	0.00	0.01	0.00	0.00	0.02
	High B2	15	0.01	0.02	0.00	0.00	0.06
	C1	34	0.01	0.02	0.00	0.00	0.10
Explained medical terms (context) per AS-unit	A2	3	0.01	0.02	0.00	0.00	0.03
	Low B1	5	0.00	0.01	0.00	0.00	0.02
	High B1	13	0.00	0.01	0.00	0.00	0.03
	Low B2	12	0.01	0.02	0.00	0.00	0.07
	High B2	15	0.01	0.01	0.00	0.00	0.04
	C1	34	0.01	0.02	0.00	0.00	0.09

Figure 24

Relationship between Explained Medical Terms by Definition per AS-Unit and CCE Scores

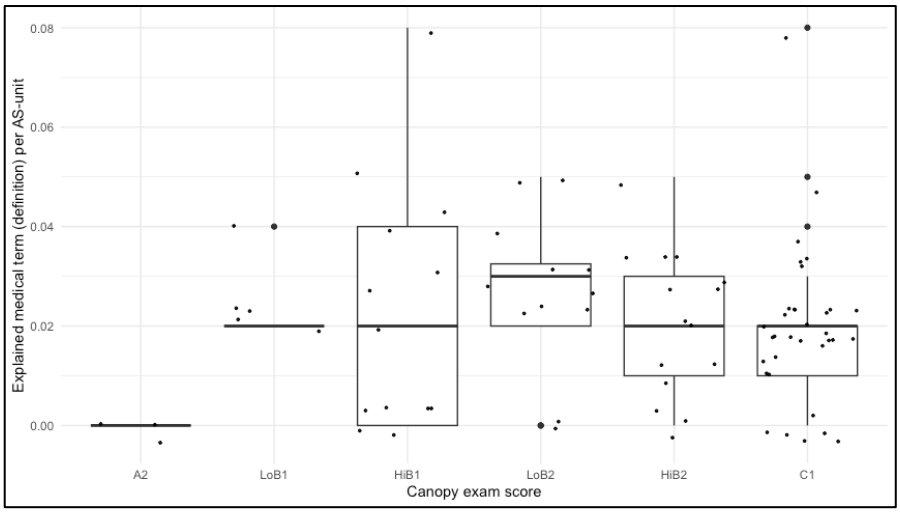


Figure 25

Relationship between Explained Medical Terms by Synonym per AS-unit and CCE Scores

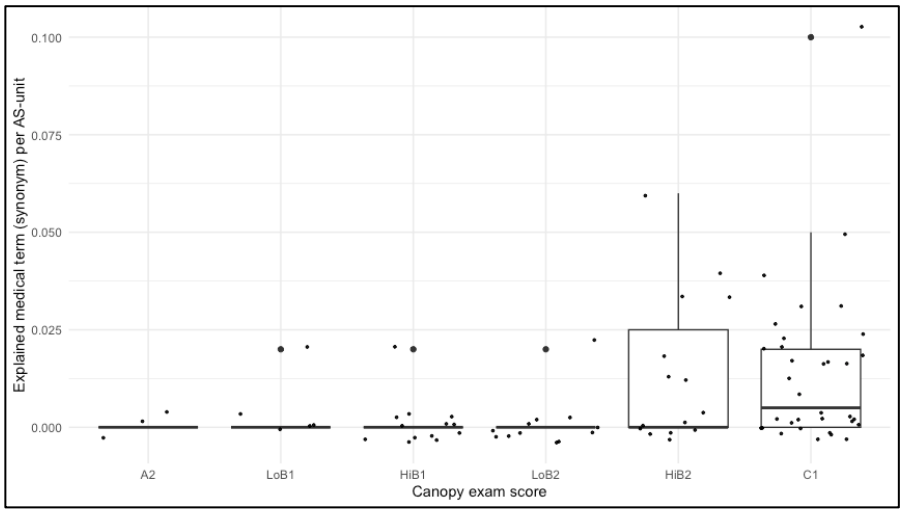
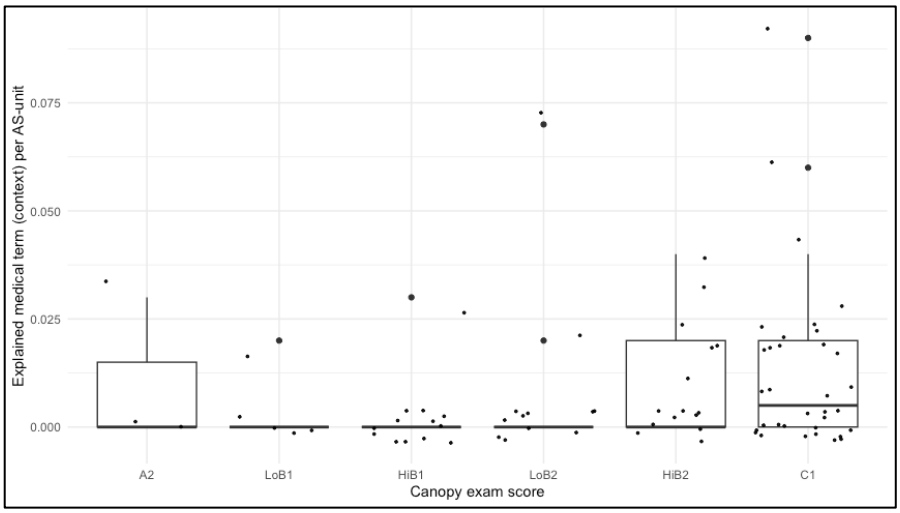


Figure 26

Relationship between Explained Medical Terms by Context per AS-unit and CCE Scores



An ordinal regression model was fit to examine the relationship between explained medical terms by definition, synonym, and context per AS-unit, and participants' medical Spanish speaking

scores on the CCE. The predictors for the fit model were explained medical terms by definition per AS-unit, explained medical terms by synonym per AS-unit, and explained medical terms by context per AS-unit. The dependent variable was CCE medical Spanish speaking scores. This fit model was compared to a null model with no predictors using a likelihood ratio test. The comparison indicated that the model fit was statistically significantly better for the model with predictors than the null model, $X^2(3) = 14.93$, $p = 0.002$. The Nagelkerke's R^2 value ($R^2 = 0.17$) suggests a weak relationship between these predictors and CCE medical Spanish speaking scores. The proportional odds assumption was evaluated using the nominal_test function from R ordinal package. The results yielded that none of the tests were significant, which indicates that the assumption of proportional odds was met.

Concerning the individual predictors, the results showed that explained medical terms by synonym per AS-unit was the only significant factor, $\beta = 47.66$, $SE = 16.80$, $z = 2.84$, $p = 0.005$, *99% CI* [9.41, 96.63], and the relationship was positive. This suggests that higher production of explained medical terms by synonym per AS-unit was associated with higher CCE medical Spanish speaking scores. Explained medical terms by definition per AS-unit was not found to be a significant factor, $\beta = 0.57$, $SE = 12.14$, $z = 0.05$, $p = 0.96$, *99% CI* [-30.66, 32.94]. Equally, explained medical terms by context was also not found to be a significant factor, $\beta = 30.74$, $SE = 14.87$, $z = 2.07$, $p = 0.04$, *99% CI* [-3.71, 75.29]. Table 26 summarizes these results.

Table 26

Results Summary for the Relationship between Linguistic Measures of Effectiveness and CCE Scores

Predictor	Estimate (β)	*SE*	*z*	*p*	*99% CI* (Lower)	*99% CI* (Upper)
Explained medical terms (definition) per AS-unit	0.57	12.14	0.05	0.96	-30.66	32.94
Explained medical terms (synonym) per AS-unit	47.66	16.80	2.84	0.005	9.41	96.63
Explained medical terms (context) per AS-unit	30.74	14.87	2.07	0.04	-3.71	75.29

4.2.8 Relationship between participants' scores in the CCE and their self-assessment of medical Spanish listening and speaking ability

As mentioned before, participants in this study achieved medical Spanish speaking scores in the CCE covering the full range of medical Spanish proficiency levels that the CCE measures (i.e., A2, low B1, high B1, low B2, high B2, and C1). They also completed a self-assessment of medical Spanish listening and speaking ability where they had to select among four different categories, the one that better described them. These four categories were organized from *a* to *d*, *a* representing the lowest proficiency level and *d* the highest. These were the statements included in the self-assessment categories:

a. I can understand and communicate simple medical information and sustain very simple medical interactions.
b. I can communicate medical information and sustain medical interactions with some fluency. I can understand familiar language normally encountered in medical interactions. Sometimes I might ask for clarification.

c. I can communicate medical information and sustain medical interactions fluently, with few hesitations. I can understand most language encountered in medical interactions.

d. I can communicate medical information and sustain medical interactions fluently and effectively. I can completely understand language encountered in medical interactions.

Table 27 presents the distribution of participants across CCE medical Spanish speaking score categories, and self-assessment categories. As can be seen, there were fewer participants in the lowest CCE proficiency score categories (i.e., A2 [*n* = 3]; low-B1 [*n* = 5]), and the self-assessment category representing the highest proficiency level of medical Spanish listening and speaking ability (i.e., category d [*n* = 5]).

Table 27

Participants Distribution across CCE Score Categories and Self-Assessment Categories

CanopyCredential exam		Self-assessment	
Scores	*n*	Categories	*n*
A2	3	a	18
Low-B1	5	b	30
High-B1	13	c	29
Low-B2	12	d	5
High-B2	15		
C1	34		

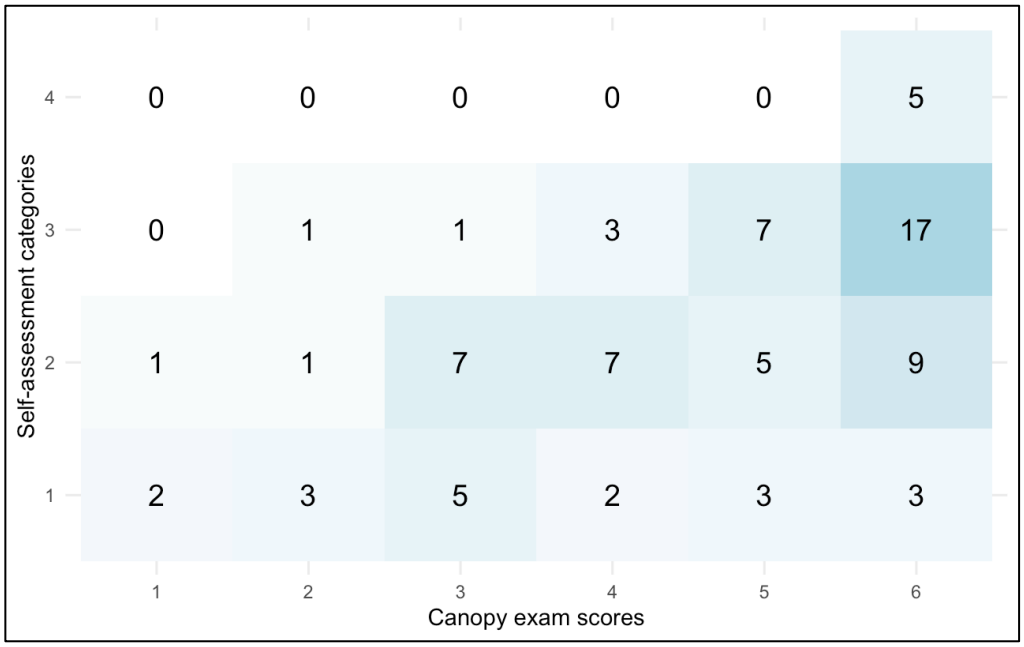
A Spearman's rank-order correlation was used to examine the relationship between participants medical Spanish speaking scores in the CCE and the categories they selected in their self-assessment of medical Spanish listening and speaking ability. This analysis was also conducted employing R statistical software, version 4.4.2 (R Core Team, 2024), and the packages

ordinal, version 2023.12-4.1 (Christensen, 2023), ggplot2, version 3.5.1 (Wickham, 2016), and vcd, version 1.4-13 (Meyer et al., 2024). The results from the Spearman's rank-order correlation indicate a significant moderate positive relationship, *rho* = 0.51, $p < 0.001$. This suggests that the CCE medical Spanish speaking scores and the self-assessment categories might be positively related.

Figure 27 displays a heatmap of the relationship between participants medical Spanish speaking scores in the CCE and the categories they selected in their self-assessment of medical Spanish listening and speaking ability. The numbers in the cells indicate the frequency of occurrence for the corresponding CCE medical Spanish speaking scores and self-assessment categories. As can be seen, participants who selected categories two and three (i.e., b and c, respectively) in the self-assessment achieved higher medical Spanish speaking scores in the CCE.

Figure 27

Relationship between Participants' Self-Assessment Categories and CCE Scores

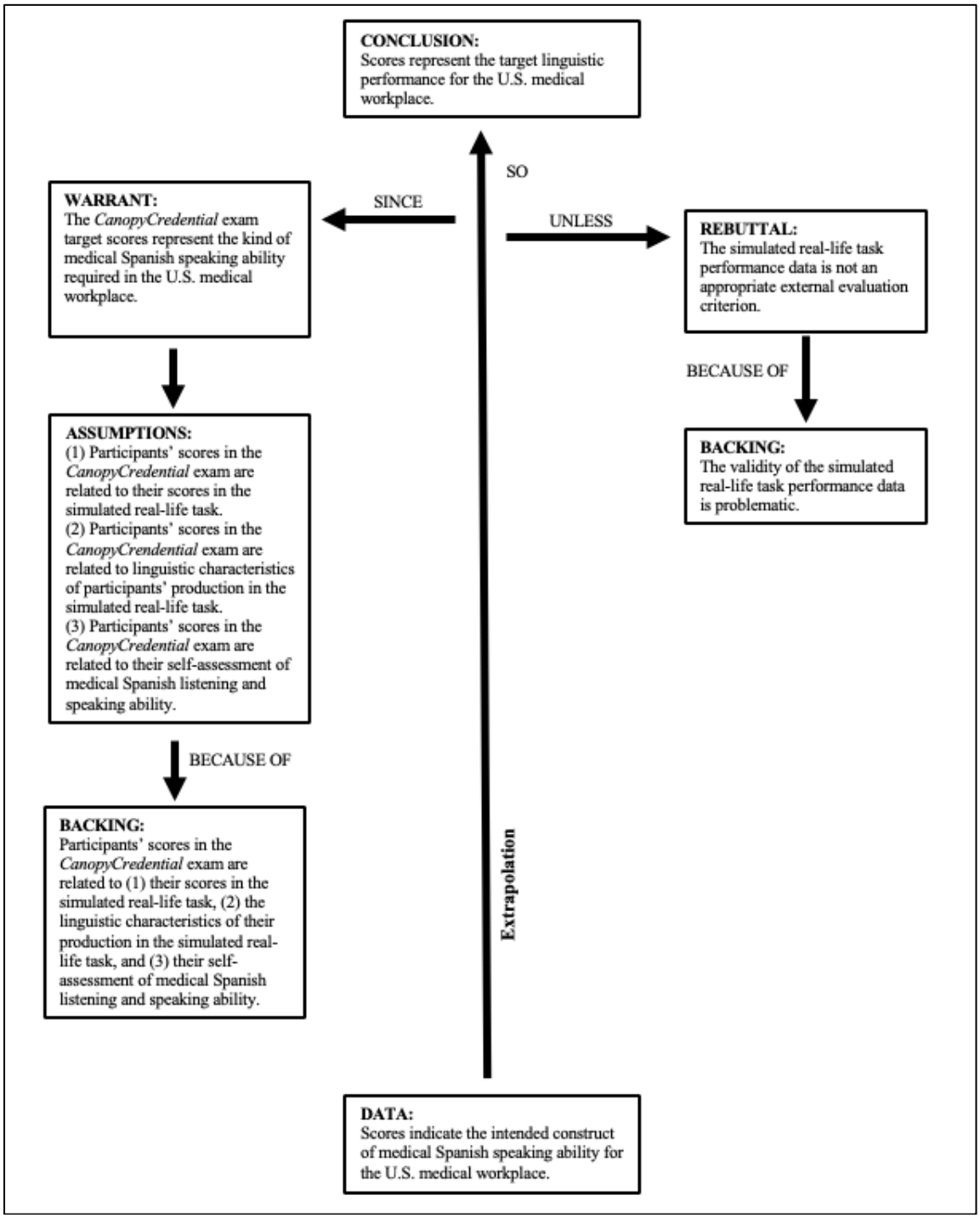


Chapter 5: Discussion

As stated before, RQ1 and RQ2 sought to examine the suitability of the SRLT as an external evaluation criterion. Following this examination, RQ3-RQ5 addressed the extrapolation inference of the speaking section of the Spanish module of the CCE. In the next sections, each RQ is discussed using the structure model displayed in Figure 28 as guide. This structure model is based in the interpretation/use argument that was developed for the speaking section of the Spanish module of the CCE described in chapter two. As such, RQ1 and RQ2 focus on the right side of this structure model, as seen on Figure 28. Specifically, the rebuttal for the extrapolation inference of the speaking section of the Spanish module of the CCE and its backing. Refuting the rebuttal requires showing that the SRLT developed for this study is an appropriate instrument for investigating the extrapolation inference of this section and module of the CCE. These RQs are discussed following the model by Aryadoust (2016, pp. 16-22). On the other hand, RQ3-RQ5 focus on the left side of the structure model shown on Figure 28. Namely, the assumptions and backing for the extrapolation inference of the speaking section of the Spanish module of the CCE. This involves detailing the relationship between the participants' SRLT performance and their scores on the CCE.

Figure 28

Structure Model for the Extrapolation Inference of the Speaking Section of the Spanish Module of the CCE



Note. Adapted from Chapelle et al., (2008, p. 7)

5.1 RQ1: Do the SRLT facets (i.e., participant's ability, rating criteria, rating scale, and rater severity) function appropriately?

The goal of RQ1 was to explore the functionality of the SRLT facets. These were participants' ability, the linguistic scoring criteria, the rating scale, and raters' severity/leniency levels. Beginning with the participants in the study, these came from the same pool of individuals who normally take the CCE for their personal/professional use (Validation study of Canopy Credential, n.d.). Thus, the sample used in this study is representative of the target population. The SRLT was designed to tap into the same underlying construct of medical Spanish speaking ability for the U.S. medical workplace as the speaking section of the Spanish module of the CCE. The results from the MFRM analysis showed that the majority of the participants had MNSQ fit values within the range of 0.60-1.40 recommended by Bond et al. (2021). This suggests that their response patterns align with the model expectations. That is, based on their estimated ability to speak medical Spanish, the difficulty of the scoring rating criteria and scale, and the severity/leniency levels of the raters, participants' performance in the SRLT displayed a pattern that is similar to what it is predicted by the Rasch model.

The fit values for the linguistic scoring criteria were also within the recommended fit range with the exception of the effectiveness criterion, which showed a slight deviation from this range. Since all the other criteria fit the model well and the effectiveness criterion deviation was small, this suggests that overall, the linguistic scoring criteria constitute a unidimensional measurement scale comprised of multiple subcomponents which aligns with model expectations (Aryadoust, 2016; Tennant & Pallant, 2006). Therefore, these criteria seemed to be effective at capturing performance on a task intended to assess medical Spanish speaking ability for the U.S. medical workplace.

As for the rating scale, the results from the MFRM analysis indicated that it also fits the model well. This suggests that the rating scale categories have a clear order where higher categories represent higher levels of medical Spanish speaking proficiency across the six linguistic scoring criteria. Furthermore, the scale seemed to function well across raters, which suggests that they used the scale categories in alignment with model expectations. This seems to be the case for participants as well, which suggests that they transitioned across categories orderly with no significant skipping or jumping between categories.

The raters' fit values were also within the recommended range. This suggests that their ratings were consistent with model expectations, and that they applied the rating scale in a uniform manner across participants and linguistic scoring criteria. Thus, none of the raters seemed to display levels of arbitrary severity/leniency that influenced participants' scores, the scoring of the linguistic criteria, or the rating scale categories.

According to these results, all the SRLT facets appeared to function in alignment with Rasch model predictions and influenced ratings in agreement with model expectations. No significant distortions or biases were observed, and each facet seemed to contribute to the overall measurement in a consistent manner without introducing significant discrepancies. Therefore, it can be inferred that the overall measurement fitted the Rasch model well and aligned with its expectations. This provides evidence supporting the use of the SRLT as a measure of medical Spanish speaking ability in a target scenario of language use.

5.2 RQ2: Do the SRLT raters and linguistic scoring criteria reliably differentiate participants' proficiency level?

RQ2 focuses on the reliability of the measurement. This refers to the probability of obtaining similar results if this participants group was evaluated by a different team of raters with the same

profile as those employed in this study, having received the same training, and using the same linguistic scoring criteria and rating scale. This can be inferred from the reliability and separation coefficients generated by the MFRM analysis.

As mentioned in the results chapter, participants' reliability and separation coefficients showed four statistically distinguishable proficiency levels. This suggests that this measurement was effective at discriminating between low- and high-proficient participants. Therefore, it could be predicted that if these participants were assessed by another team of raters following the same raters' selection and training procedures, and the instruments that were used in this study, similar results would be obtained. In addition, if the same team of raters that participated in this study assessed a different group of participants from the same population group as the one that was sampled for this study, their performance would also show a degree of consistency similar to the one observed in this study. Regarding the difficulty of the linguistic scoring criteria, the results suggest that their level of difficulty would remain constant across different participants and raters' samples, from the same population groups. Accordingly, it can be inferred that the SRLT raters and linguistic scoring criteria were reliable at differentiating participants' proficiency levels in this study's measurement, and that they probably would perform similarly with other samples from the same participants/raters' pools.

Based on the results obtained from the MFRM analysis and discussed above for RQ1 and RQ2, there is evidence suggesting that the data obtained from the SRLT designed and used in this study is suitable to explore the extrapolation inference of the speaking section of the Spanish module of the CCE. No evidence supporting the rebuttal for this inference—included in the right side of the structure model displayed on Figure 28—was found in this study's measurement. The next sections of this chapter will discuss RQ3-RQ5, which address the assumptions and backing

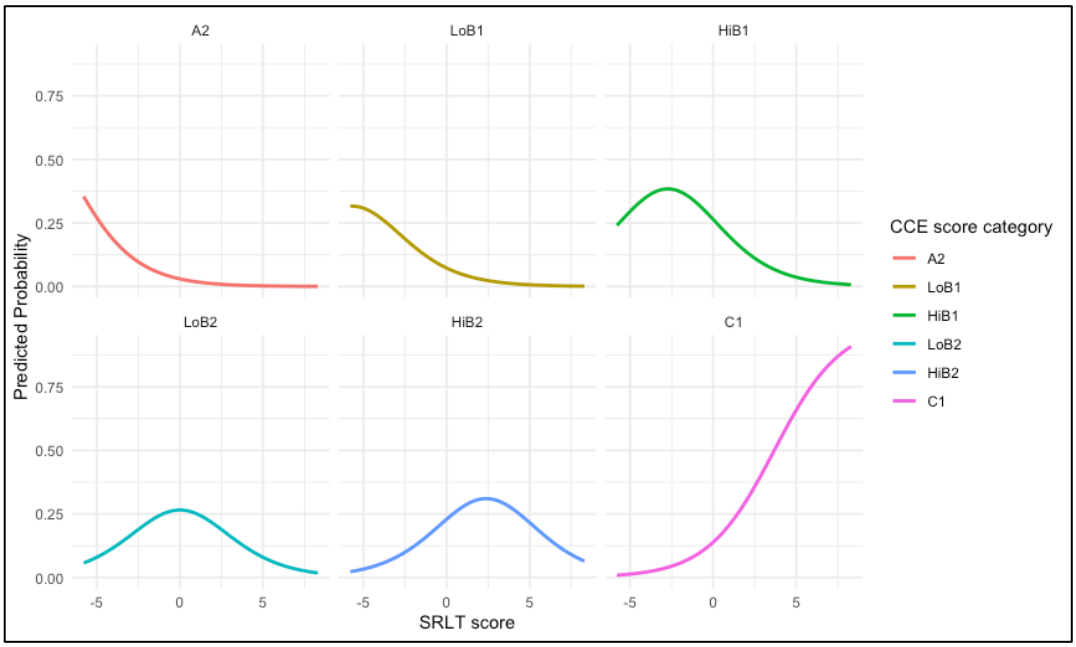
for the extrapolation inference of the speaking section of the Spanish module of the CCE (see left side of Figure 28), using the data obtained from this study's SRLT measurement as external evaluation criterion.

5.4 RQ3: To what extent are participants' scores in the speaking section of the Spanish module of the CCE related to their scores in the SRLT?

As was described in the results chapter, the results from the ordinal regression model used to address RQ3 showed a moderate association between participants' scores in the SRLT and the speaking section of the Spanish module of the CCE. This suggests that SRLT scores could be a predictor of CCE medical Spanish speaking score achievement. The ordinal regression results indicated that for every one logit increase in SRLT score, the odds of being rated at a higher category increased by a factor of 1.65. To illustrate this, the ordinal regression model predicted probabilities of CCE medical Spanish speaking score attainment for a range of SRLT logit scores are displayed in Figure 29. In this figure, each CCE medical Spanish speaking score category is represented by a separate graph where overall SRLT scores are shown in the x-axis, and the y-axis shows the predicted probability of achieving that particular score category in the speaking section of the Spanish module of the CCE. These graphs are organized from left to right, top to bottom, with the lowest CCE score (i.e., A1) at the top left corner, and the highest CCE score (i.e., C1) at the bottom right corner. As can be seen when looking from left to right, top to bottom, the model's predicted probabilities of achieving higher score categories in the speaking section of the Spanish module of the CCE progressively increase as overall SRLT scores increase.

Figure 29

Predicted Probabilities of Achieving a Specific CCE Score Category based on SRLT Scores



These results lend support to the first assumption stated in the argument-based structure model used to investigate the extrapolation inference of the speaking section of the Spanish module of the CCE. This assumption referred to participants' scores in the speaking section of the Spanish module of the CCE being related to their overall scores in the SRLT (see left side of Figure 28 above). Since a moderate positive relationship between the SRLT and the CCE medical Spanish speaking scores was found in this study, this assumption can be sustained.

5.5 RQ4: To what extent are participants' scores in the speaking section of the Spanish module of the CCE related to their production in the SRLT on linguistic measures of intelligibility, fluency, grammar, vocabulary, coherence, and effectiveness?

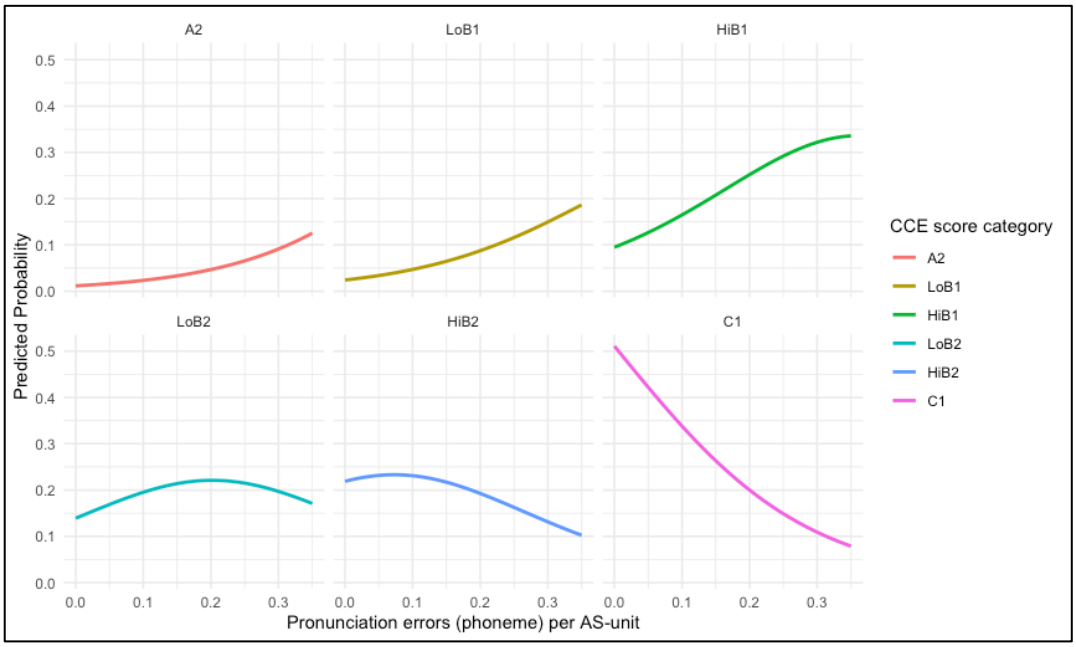
RQ4 further explored the relationship between participants' speaking scores in the Spanish module of the CCE and their performance in the SRLT by focusing on a range of linguistic measures from

their production in the SRLT. As was explained in the data analyses section within the methodology chapter, these linguistic measures were selected based on the linguistic scoring criteria included in the CCE and the SRLT scoring rubrics, the target construct of medical Spanish speaking ability for the U.S. medical workplace, a review of the literature, and the characteristics of the data collected for this study. Given the exploratory character of this RQ, a decision was made to investigate these linguistic measures and their relationship with participants' medical Spanish speaking scores in the CCE by linguistic scoring criteria.

Beginning with the intelligibility criterion, the results showed that two of the linguistic measures selected (i.e., phoneme pronunciation errors per AS-unit, and syllable pronunciation errors per AS-unit) were associated with low CCE medical Spanish speaking score attainment. To illustrate this, Figure 30 displays the ordinal regression model's predicted probabilities of CCE medical Spanish speaking score achievement based on production of phoneme pronunciation errors per AS-unit. In this figure, each CCE medical Spanish speaking score category is represented by a graph. These graphs are organized from left to right, top to bottom, with the lowest CCE medical Spanish speaking score category (i.e., A2) at the top left corner, and the highest CCE medical Spanish speaking score category (i.e., C1) at the bottom right corner. The x-axis in these graphs shows phoneme pronunciation errors per AS-unit, and the y-axis shows the predicted probability of achieving that specific CCE medical Spanish speaking score category. As can be seen in the top three graphs, higher production of phoneme pronunciation errors per AS-unit increases the predicted probabilities of achieving the three lowest CCE medical Spanish speaking score categories. In contrast, when looking at the bottom three graphs, higher production of phoneme pronunciation errors per AS-unit decreases the predicted probabilities of achieving the three highest CCE medical Spanish speaking score categories.

Figure 30

Predicted Probabilities of Achieving a Specific CCE Score Category based on Production of Phoneme Pronunciation Errors per AS-unit

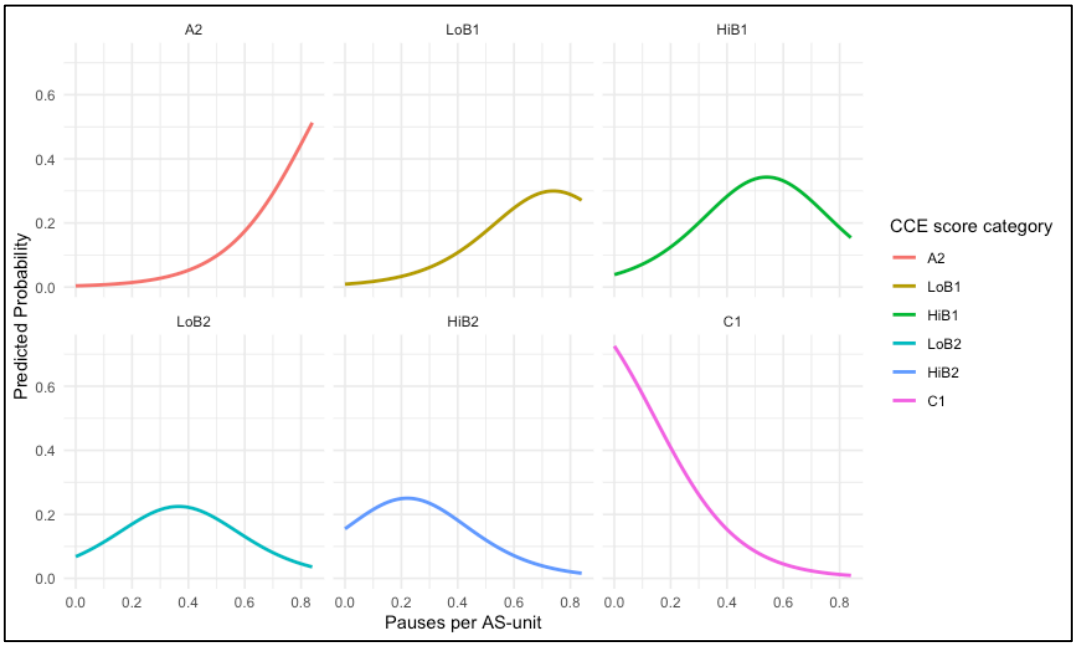


Some of the most common phoneme pronunciation errors that were found in participants' production in the SRLT included vocalic phoneme substitutions such as *colester/a/l* instead of *colester/o/l* (Participant 14; low-B2 CCE score) (English translation: cholesterol), *rec/i/ta* instead of *rec/e/ta* (Participant 14; low-B2 CCE score) (English translation: prescription), or *l/é/pidos* instead of *l/i/pidos* (Participant 15; high-B1 CCE score) (English translation: lipids). Regarding syllable pronunciation errors, some of the most common ones involved adding or eliminating vowels within syllable units. Some examples found were *medicam/i/entos* instead of *medicam/e/ntos* (Participant 18; low-B2 CCE score) (English translation: medication), equilibr/o/ instead of equilibr/i/o (Participant 18; low-B2 CCE score) (English translation: balance), or anteced/i/entes instead of anteced/e/ntes (Participant 18; low-B2 CCE score) (English translation: history).

Regarding the fluency criterion, pauses per AS-unit was found to be associated with lower CCE medical Spanish speaking scores. Figure 31 illustrates the ordinal regression model's predicted probabilities of CCE medical Spanish speaking score attainment based on production of pauses per AS-unit. This figure used the same organization pattern for the CCE medical Spanish speaking score categories as the one used in Figures 29 and 30 above. As shown in Figure 31, an increase in the production of pauses per AS-unit augments the predicted probabilities of attaining lower CCE medical Spanish speaking scores (see top three graphs), and decreases the predicted probabilities of obtaining higher CCE medical Spanish speaking scores (see bottom three graphs).

Figure 31

Predicted Probabilities of Achieving a Specific CCE Score Category based on Production of Pauses per AS-unit

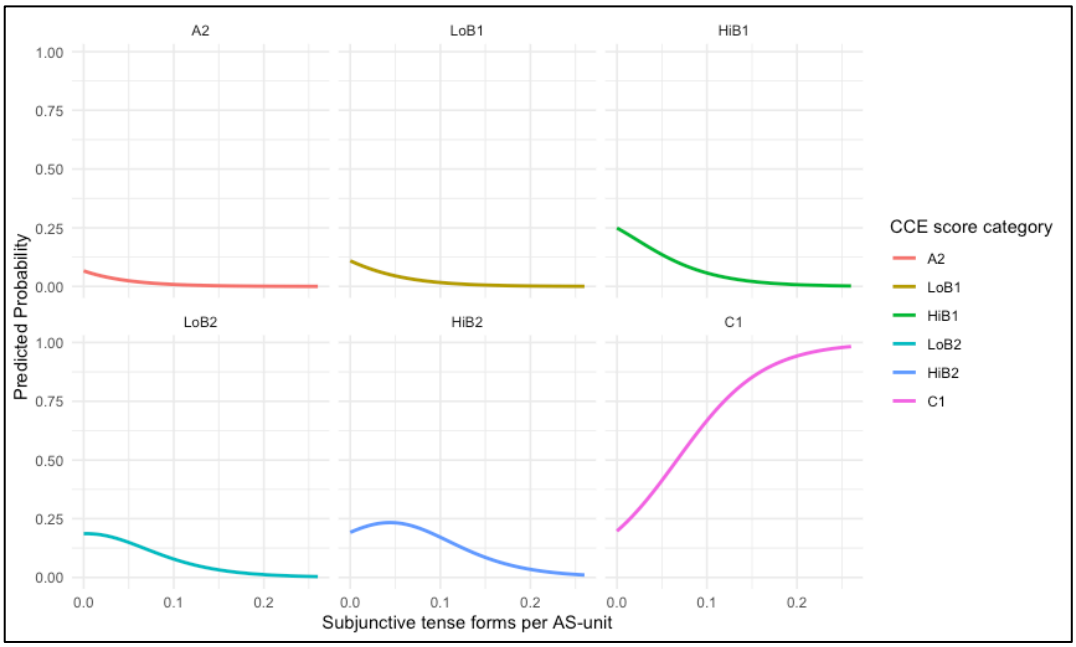


Among the three grammatical complexity measures selected for the grammar criterion in this study, only production of subjunctive tense forms per AS-unit was found to be associated with higher CCE medical Spanish speaking score achievement. Figure 32 illustrates the ordinal

regression model's predicted probabilities of CCE medical Spanish speaking score attainment based on production of subjunctive tense forms per AS-unit. As can be observed, the predicted probabilities of obtaining lower CCE medical Spanish speaking scores decrease as production of subjunctive tense forms per AS-unit increases (see top three graphs), whereas the predicted probabilities of obtaining higher CCE medical Spanish speaking scores increase as the production of subjunctive tense forms per AS-unit grows (see bottom three graphs).

Figure 32

Predicted Probabilities of Achieving a Specific CCE Score Category based on Production of Subjunctive Tense Forms per AS-unit

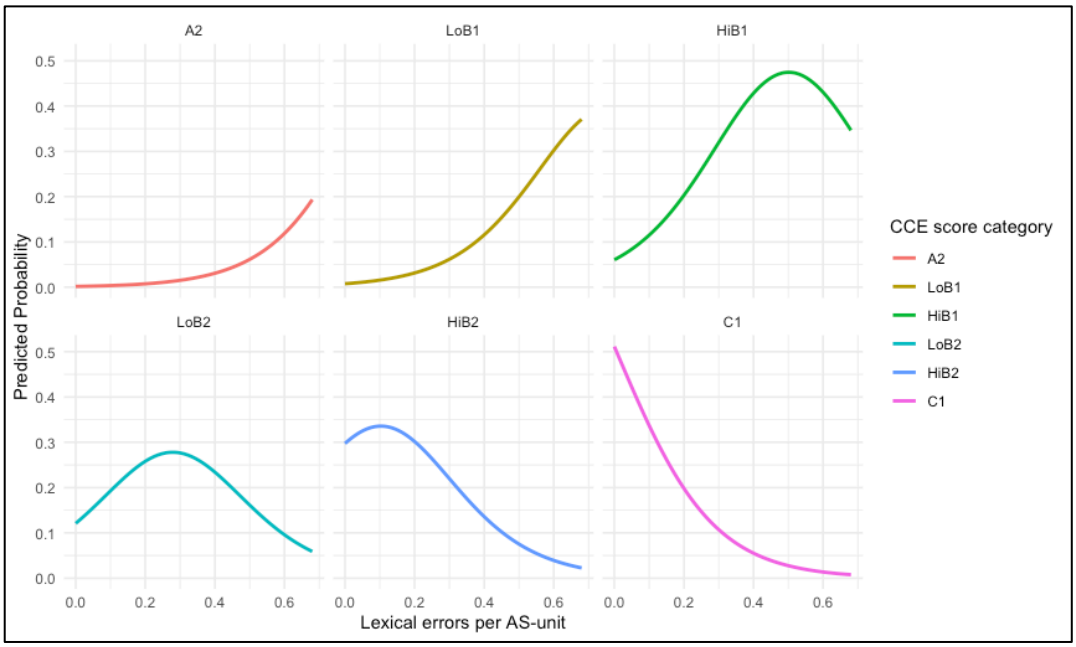


As for the grammatical accuracy measures selected for the grammar criterion, two of them were found to be associated with lower CCE medical Spanish speaking score attainment. These were lexical errors per AS-unit, and morphological errors per AS-unit. Figure 33 presents the ordinal regression model's predicted probabilities of CCE medical Spanish speaking score achievement based on production of lexical errors per AS-unit. When looking at the top three

graphs in Figure 33 representing the three lower CCE medical Spanish speaking score categories, it can be observed that the predicted probabilities of receiving these CCE medical Spanish speaking scores are higher when more lexical errors per AS-unit are produced. In contrast, when looking at the bottom three graphs, the predicted probabilities of achieving the three higher CCE medical Spanish speaking score categories decrease as production of lexical errors per AS-unit increase.

Figure 33

Predicted Probabilities of Achieving a Specific CCE Score Category based on Production of Lexical Errors per AS-unit



Some of the most common lexical errors that were found in participants' production in this study's SRLT were related to the use of terms influenced by English language or made-up words/neologisms. Some examples include *resultos* (Participant 9; high-B2 CCE score) to refer to medical laboratory test results, *eseserios* (Participant 66; low-B1 CCE score) to refer to physical

exercise, or *improbar tu colesterol* instead of *mejorar tu colesterol* (Participant 66; low-B1 CCE score) (English translation: to improve your cholesterol).

Regarding the linguistic measure selected for the vocabulary domain, there was no association observed between production of medical terms per AS-unit and CCE medical Spanish speaking score achievement. This could potentially be due to the fact that the SRLT task prompt that was given to the participants included the medical terms that they would need to use to complete the task, which could have limited the number of medical terms they ultimately used.

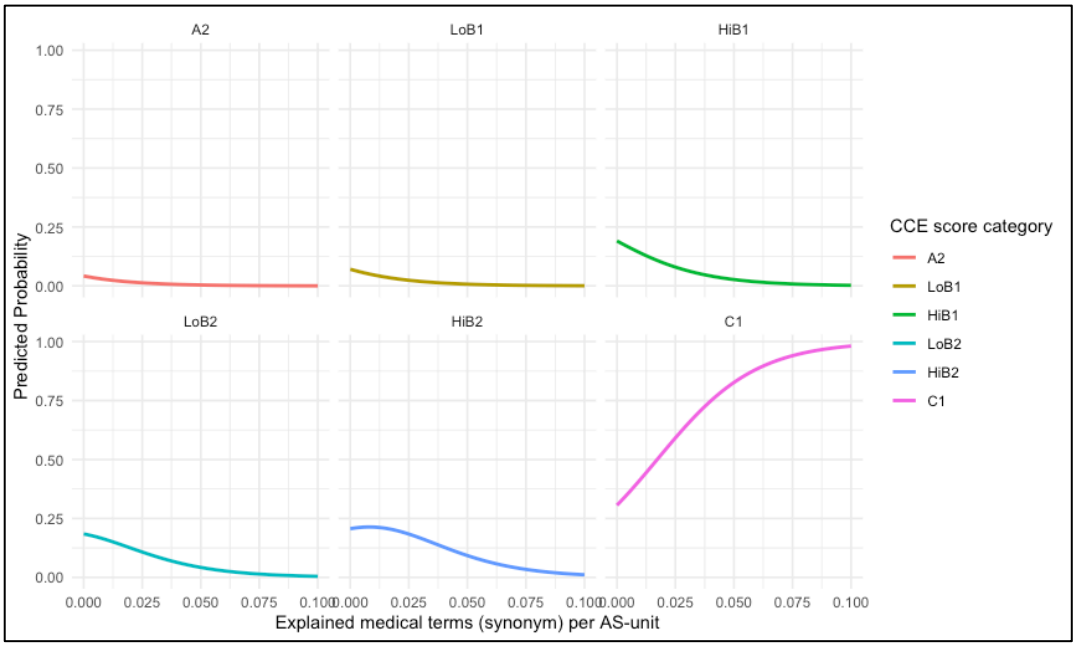
Concerning the linguist measures selected for the coherence criterion (i.e., coordinating conjunctions per AS-unit, and subordinating conjunctions per AS-unit), no association was found between any of them and CCE medical Spanish speaking score achievement. This might be associated with the results obtained for the clauses per AS-unit linguistic measure, which was one of the measures selected for the grammar complexity criterion. In this case, this measure was not associated with CCE medical Spanish speaking score achievement either. As Brooks and Swain (2014) explain, this could be due to the fragmentary characteristics of interactive dialogues where these linguistic features are not as common as in monologues (Brooks & Swain, 2014, p. 369). Future studies could use participants' production in the CCE—which used a monologue format in the speaking tasks—and their production in the SRLT—which employed a dialogue format—to examine the use of coordinating and subordinating conjunctions in these two formats.

One of the linguistic measures selected for the effectiveness criterion (i.e., explained medical terms by synonym per AS-unit) was found to be associated to higher CCE medical Spanish speaking score achievement. Figure 34 shows the ordinal regression model's predicted probabilities of CCE medical Spanish speaking score achievement based on production of explained medical terms by synonym per AS-unit. As can be seen particularly in the graphs for

the two highest CCE medical Spanish speaking score categories (i.e., high-B2 and C1), the predicted probabilities of achieving these scores increase as production of explained medical terms by synonyms per AS-unit increases.

Figure 34

Predicted Probabilities of Achieving a Specific CCE Score Category based on Production of Explained Medical Terms by Synonym per AS-unit



Some examples of explained medical terms by synonym that participants in this study produced in the SRLT are *"la glucosa, el azúcar en su sangre salió alta"* (Participant 11; C1 CCE score) (English translation: the glucose level or your blood sugar is high), *"eso pasa cuando uno tiene mucha grasa o muchos lípidos en su sangre"* (Participant 23; C1 CCE score) (English translation: that happens when someone has too much fat or lipids in their blood), "el colesterol que tiene la densidad más baja le referimos colesterol malo y eso no lo queremos tanto en nuestro cuerpo" (Participant 62; C1 CCE score) (English translation: low density cholesterol is known as bad cholesterol and we don't want too much of that one in our body).

The second assumption included in the argument-based structure model used to investigate the extrapolation inference of the speaking section of the Spanish module of the CCE referred to the relationship between participants' scores in this section and module of the CCE and the linguistic characteristics of their production in the SRLT. Given the proficiency-oriented focus of the CCE, it could broadly be expected that fewer errors and more complex language would be indicative of higher levels of performance. The results from the linguistic measures selected for this study and discussed in this section show that lower production of phoneme and syllable pronunciation errors per AS-unit, and lexical and morphological errors per AS-unit in the SRLT was associated to higher CCE medical Spanish speaking score achievement. In addition, higher production of subjunctive tense forms per AS-unit was associated to higher CCE medical Spanish speaking score attainment. Lower production of pauses per AS-unit, and higher production of explained medical terms by synonym per AS-unit were also associated with higher CCE medical Spanish speaking scores. Thus, these results provide evidence suggesting that SRLT productions characterized by lower production of pauses and the pronunciation and grammar errors mentioned, along with higher production of subjunctive tense forms and explained medical terms by synonym were associated to higher CCE medical Spanish speaking score achievement. Under the assumption that the linguistic characteristics of participants' productions in the SRLT and the CCE were similar, this evidence provides support to the second assumption included in the structure model which can therefore be sustained.

Nevertheless, the linguistic measures selected in this study could have been impacted by some of the constraints of the SRLT. As mentioned before, these included the insertion of the medical terms that participants needed to complete the task in the participants' task prompt, and the dialogue format. These could have influenced participants' use of medical terms, and

coordinating and subordinating conjunctions. Furthermore, the limited number of sub-tasks within the SRLT (e.g., asking follow-up questions, explaining results, delivering diagnosis, etc.), could have had an effect on participants' production of these linguistic measures as well. Additional tasks with different interlocutors, such as requiring the participant to update another medical professional on a patient's condition and progress for example, could provide an opportunity to elicit a more robust language sample. Future studies without the logistic limitations of this one could include additional sub-tasks at different levels of predicted difficulty to explore this potential effect of task constraints on participants' production of the selected linguistic measures. Also, the skewed nature of the participant pool towards higher proficiency levels in this study may have impacted the range of values obtained for some of the linguistic measures. A future study with more robust representation at the lower levels would help explore this possibility more fully.

5.6 RQ5: To what extent are participants' scores in the speaking section of the Spanish module of the CCE related to their self-assessment of medical Spanish speaking ability?

The results from the Spearman's rank-order correlation indicated a significant positive relationship between participants' self-assessment of their medical Spanish listening and speaking ability and their CCE medical Spanish speaking scores. The third assumption included in the argument-based structure model employed to investigate the extrapolation inference of the speaking section of the Spanish module of the CCE referred to the relationship between participants' scores in this module and section of the CCE and their self-assessment of medical Spanish listening and speaking ability. The positive correlation between the self-assessment scores and the CCE medical Spanish speaking scores provide evidence supporting this assumption. Therefore, this assumption can be sustained as well.

The results discussed in this section addressing RQ3-RQ5 provide evidence suggesting a positive relationship between participants' medical Spanish speaking scores in the CCE and their overall scores in the SRLT. Regarding the relationship between participants' CCE medical Spanish speaking scores and the linguistic characteristics of their production in the SRLT, the evidence obtained from the analyses suggests an association between a range of linguistic measures and participants' medical Spanish speaking scores on the CCE. A positive relationship was also found between participants' medical Spanish speaking scores in the CCE and their self-assessment of medical Spanish listening and speaking ability. As such, these findings provide evidence supporting the extrapolation inference of the speaking section of the Spanish module of the CCE, which was a video-mediated monologic task, to a dialogic task with a human interlocutor.

However, it is important to consider that even though relationships were found in the predicted directions, these relationships varied in their strength. In addition, this study's SRLT focused on a single specific medical scenario, and the population sample was small and skewed towards higher proficiency levels. Therefore, this study is one of the multiple pieces of evidence that should be collected and considered by stakeholders when interpreting and using the CCE medical Spanish speaking scores as well as the scores from similar LAPP.

Chapter 6: Conclusion

The present study found evidence suggesting that SRLT like the one developed for this project could be a suitable external evaluation criterion to examine the extrapolation inference of LAPP, specifically those used in the medical domain. The procedures that were used in the development of this SRLT and described in the methodology chapter were found to be useful for the design of this SRLT. Furthermore, the results from the MFRM analysis indicated that the measurement obtained from this study's SRLT was appropriate to be used in the examination of the extrapolation inference of the speaking section of the Spanish module of the CCE. Accordingly, it could be inferred that SRLT like the one employed in this study could be used in future studies investigating this inference in other LAPP.

In reference to the extrapolation inference of the speaking section of the Spanish module of the CCE, the results from this study provide support for this inference. More studies should be conducted to further examine this inference, considering the recommendations and limitations below.

The first limitation that should be taken into account when interpreting the results from this study is the participants' sample size (N = 82). This sample was relatively small for the type of analyses that were employed, and the sample was skewed towards more proficient speakers, though this is representative of the population taking the CCE. Nevertheless, as Linacre (1994) sustains, the main goal of the MFRM is to elicit information that can assist researchers in the evaluation and use of their measurement instruments. This idea was followed in this study and for this reason the results obtained from the MFRM analysis were considered sufficient to conclude that the measurement obtained through this study's SRLT was appropriate for the purpose of this

project. However, it is acknowledged that additional useful information could be obtained with a larger sample of participants, and by having all participants' productions rated by all three raters.

In addition, using multiple ordinal regression analyses to examine the relationship between participants' medical Spanish speaking scores in the CCE and the linguistic measures selected for each linguistic scoring criterion could have also influenced the number of associations found. Examining linguistic variables by scoring criteria (i.e., intelligibility, fluency, grammar, vocabulary, coherence, and effectiveness) was chosen for ease of interpretation. It was for this reason that a decision was made to set the alpha level at 0.01, report 99% confidence intervals, and interpret the R^2 coefficients with caution. However, examining variables piecemeal may obscure interactions between them that either increase or decrease their relationship with the CCE scores. Future studies including a larger number of participants could allow additional or alternative analyses such as structural equation models. This could shed more light about the relationship between the linguistic measures selected, participants' medical Spanish speaking scores in the CCE and the SRLT, and the target construct of medical Spanish speaking ability for the U.S. medical workplace.

Additionally, future research could also examine more in depth the relationship between the linguistic measures selected for this study and medical Spanish speaking proficiency from a developmental perspective. This could help elucidate which linguistic measures would reflect a pattern of medical Spanish speaking proficiency development. Furthermore, it would also be interesting to examine the relationship between medical Spanish speaking proficiency development and general Spanish speaking proficiency development using these linguistic measures and additional ones. These additional linguistic measures could be selected using linguistically naïve listeners. This would allow the investigation of which linguistic features are

more related and relevant to medical Spanish speaking proficiency development, and if they are similar/different from those related to general Spanish speaking proficiency.

Moreover, a better designed self-assessment tool could also be used as an additional external evaluation criterion. The self-assessment used in this study targeted medical Spanish listening and speaking ability within each descriptor. Future studies could employ a self-assessment with descriptors that address these abilities separately instead of in combination to allow for more insight into how medical professionals perceive their linguistic abilities in their additional language.

Another important aspect to consider when conducting studies like this one is the inclusion of participants and simulated patients for the SRLT for whom Spanish is an additional language. In this study, the simulated patient in the SRLT was the researcher who is a Spanish NS. In addition, there were participants in this study who had been speaking Spanish in different contexts since age five. In multilingual context such as the U.S. medical workplace, is common to find spoken interactions among individuals where all of them are using one of their additional languages, and none of them are using their native language(s). Therefore, future studies should consider this aspect by including more participants and simulated patients with this profile.

The results from this study could also have some theoretical implications for the fields of language test validity, language learning, and language assessment. In regard to language test validity, this study introduces SRLT as an alternative/additional external evaluation criterion for test validity research. Furthermore, it provides support for the argument-based framework and its usefulness for this type of research. The SRLT and the description of its design could also be helpful for language learning and for the research areas of TBLT and TBLA. As the results from this study show, the SRLT was an effective instrument to measure medical Spanish speaking

ability. This SRLT could also be used as a target task in language learning programs following the TBLT methodology. The steps that were followed to develop this SRLT could also serve as model for future test validity studies using this instrument, and for TBLT and TBLA practitioners to develop similar tasks for language learning and assessment purposes.

Appendices

Appendix A

The *CanopyCredential* exam rubrics

Table A1. The CCE scoring rubric for linguistic competence (CanopyCredential, n.d.)

LINGUISTIC COMPETENCE		
C1	**Intelligibility**	Always easy to understand with minimal or no strain for the listener. The minimal influence of the L1 does not cause any misunderstanding.
	Fluency	Mostly lengthy utterances that are almost effortless, with a few occasional hesitations and unnatural pauses, when the topic is difficult.
	Grammar	Consistently accurate grammar, but with a few occasional difficult-to-spot errors, especially while talking about complex and unfamiliar topics.
	Vocabulary	A broad range of vocabulary including the occasional use of relatively low frequency words and colloquialisms. Expert-level use of medical terms and expressions.
	Coherence	Well organized and smoothly flowing narratives with sufficient organizational patterns.
B2	**Intelligibility**	Most of the time easy to understand but with occasional strain for the listener. The influence of L1 may cause minor pronunciation or prosodic errors.
	Fluency	Some relatively lengthy utterances with some effort and hesitation, but long pauses are not common.
	Grammar	Mostly accurate grammar. Errors are not very common and oftentimes corrected on the spot.
	Vocabulary	A good range of vocabulary for a wide range of general topics, but no use of low frequency words and colloquialisms. Competent use of medical terms and expressions.
	Coherence	Mostly organized, but at times disorganized and jumpy, especially when elaborating on complex ideas with longer utterances.

(continued)

Table A1. The CCE scoring rubric for linguistic competence (CanopyCredential, n.d.) (continued)

B1	**Intelligibility**	At times difficult to understand with some strain for the listener. The noticeable influence of L1 sometimes causes pronunciation or prosodic errors.
	Fluency	No lengthy utterances, but oftentimes effortless speech, especially when talking about basic ideas. Rare long pauses are noticeable.
	Grammar	Noticeable inaccuracy but the errors rarely cause misunderstanding.
	Vocabulary	Working range of vocabulary for both everyday situations and medical terms and expressions. While gaps in vocabulary are noticeable and hesitations are frequent, the speaker is able to communicate a message, mostly without resorting to L1.
	Coherence	Disorganized and jumpy, especially when discussing unfamiliar topics and elaborating on complex ideas with longer utterances.
A2	**Intelligibility**	Sometimes difficult to understand with severe strain for the listener. Very strong influence of L1 often causes pronunciation or prosodic errors.
	Fluency	Utterances beyond short and isolated phrases and sentences, but speech is constantly interrupted through many pauses, false starts, and reformulations.
	Grammar	Noticeable inaccuracies with obvious and basic errors that sometimes cause misunderstanding.
	Vocabulary	Very limited range of vocabulary for both everyday situations and medical terms and expressions. Gaps in vocabulary knowledge causes lots of frequent hesitations and repetitions. May rely on L1 when unable to retrieve a word.
	Coherence	No clear organization. Mainly relies on the use of listing cohesive devices. Clear connections between ideas are often missing.

Table A2. The CCE scoring rubric for socio-cultural communicative competence (CanopyCredential, n.d.)

SOCIO-CULTURAL COMMUNICATIVE COMPETENCE (Ability to deliver accurate content effectively and impartially)		
4 Adept	**Effectiveness and Impartiality**	**Full** communicative-task goal achievement. **Full** delivery of **completely** accurate information. **Always** provides information in a language accessible to patients. **Smooth** transitions between formal/technical and informal/non-technical language. The speed of speech is **always easy** for the patient to follow. **Fully clear and obvious** demonstration of empathy, rapport, and respect. **Completely** avoids stereotyping and using a condescending, sarcastic, or rude tone.
In comparison to the score of 4:		
3 Competent	**Effectiveness and Impartiality**	**Slightly less than** full communicative-task goal achievement. **Slightly less than full delivery** of **almost completely** accurate information. **Almost always** provides information in a language accessible to patients. **Mostly smooth** transitions between formal/technical and informal/non-technical language. The speed of speech is **almost always easy** for the patient to follow. **Slightly less clear and obvious** demonstration of empathy, rapport, and respect. There might be some **rare signs** of stereotyping or use of a condescending, sarcastic, or rude tone.
In comparison to the scores of 3 and 4:		
2 Effective	**Effectiveness and Impartiality**	**Partial** communicative-task goal achievement. **Partial but still acceptable** delivery of **mostly** accurate information. **With effort** provides information in a language accessible to patients. **Occasional issues** transitioning between formal/technical and informal/non-technical language. The speed of speech is **usually easy** for the patient to follow. **Occasional and less clear** demonstration of empathy, rapport, and respect. There might be some **occasional** signs of stereotyping or use of a condescending, sarcastic, or rude tone.
In comparison to the scores of 2, 3 and 4:		
1 Ineffective	**Effectiveness and Impartiality**	**Limited** delivery of **somewhat** accurate information. **Incomplete** communicative-task goal achievement. **Cannot** provide information in a language accessible to patients. **Obvious issues** transitioning between formal/technical and informal/non-technical language. The speed of speech is **not easy** for the patient to follow. **Lack of** demonstration of empathy, rapport, and respect. There might be **obvious signs** of stereotyping or use of a condescending, sarcastic, or rude tone.

Appendix B

Materials that were used for this research project

Figure B1. Flyer to recruit participants for this research project.

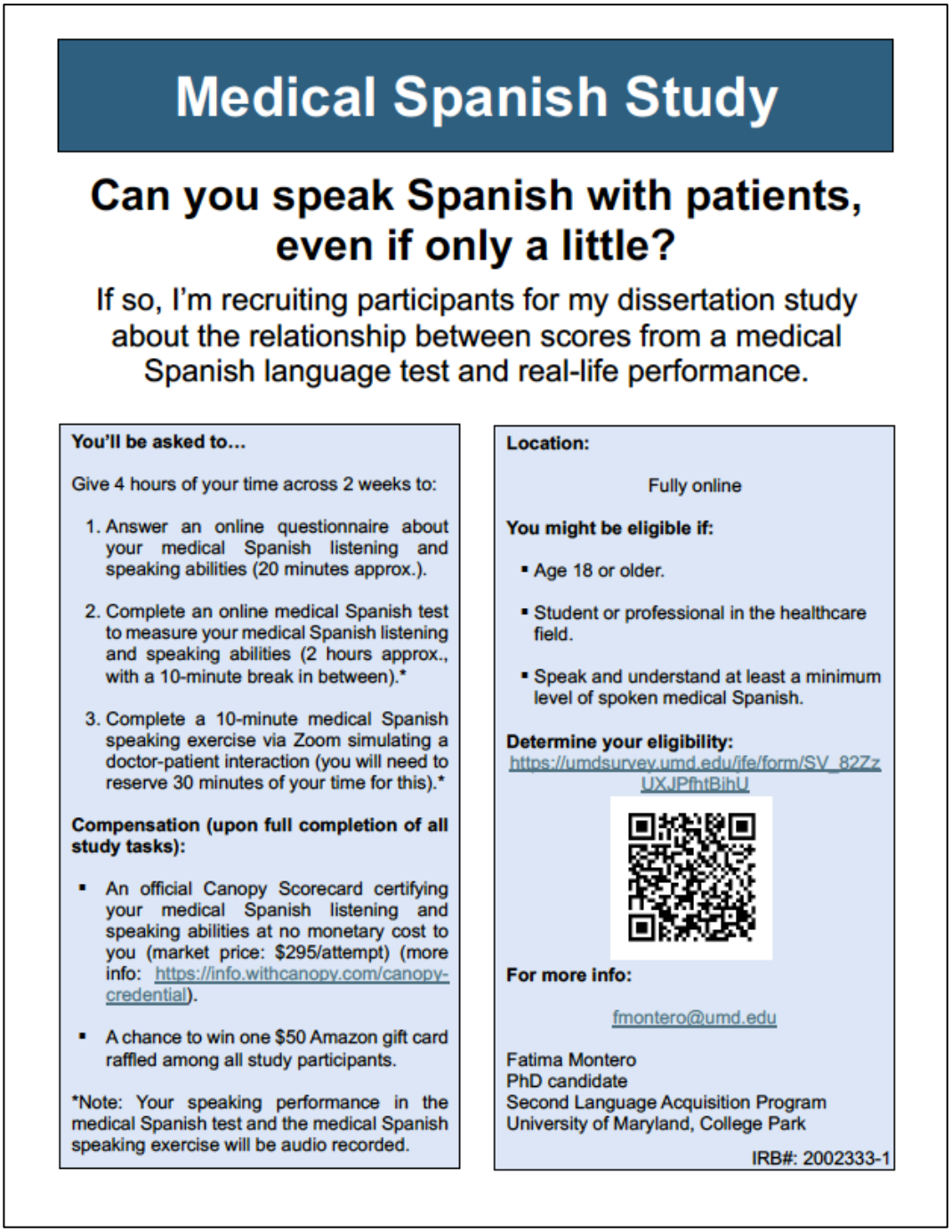


Figure B2. Participants' bio-data survey.

1. First name.
2. Last name.
3. E-mail.
4. Age, in years.
5. Current professional status:
    a. Student in the healthcare field.
    b. Professional in the healthcare field.
    c. All of the above.
    d. None of the above.
6. What option better describes your medical Spanish speaking and listening abilities?
    a. I can understand and communicate simple medical information and sustain very simple medical interactions.
    b. I can communicate medical information and sustain medical interactions with some fluency. I can understand familiar language normally encountered in medical interactions. Sometimes I might ask for clarification.
    c. I can communicate medical information and sustain medical interactions fluently, with few hesitations. I can understand most language encountered in medical interactions.
    d. I can communicate medical information and sustain medical interactions fluently and effectively. I can completely understand language encountered in medical interactions.
7. At what age (in years) did you start using Spanish in the following environments? (Type N/A when needed).
    a. At home:
    b. With friends:
    c. At school:
    d. At work:
8. What was/is the primary language of instruction for your medical education (including med school, nursing school, etc.)?
9. Have you ever taken medical Spanish courses?
    a. Yes.
    b. No.
10. If you answered "Yes" to the previous question, how many medical Spanish courses/levels have you completed?
11. How often do you use your medical Spanish listening and speaking abilities in the following contexts?

	Never	Rarely	Sometimes	Regularly	Always
At school					
At work					

Figure B3. Confidentiality agreement for the SRLT raters.

I (rater's fist and last name) agree to participate in the research project *Simulated real-life tasks as a tool to investigate the extrapolation inference of language assessments for professional purposes* (IRBNet Package: 2002333-1) as a rater. I agree to protect participants' identity and confidentiality by not sharing, recording, or saving any data or information from their audio recorded simulated medical phone call with anyone, anywhere.

Rater's Signature: _____
Date: _____

Table B1. Judging plan used to assign participants' performances in the SRLT to the raters.

		Raters		
Proficiency level (from participants' scores in Canopy Exam)	Participant	A	B	C
A2	3*	X	X	X
	12*	X	X	X
	51	X	X	
B1 LB1	6	X	X	
LB1	36	X		X
LB1	43	X		X
LB1	64		X	X
LB1	66		X	X
HB1	4	X	X	
HB1	8	X	X	
HB1	13	X	X	
HB1	15	X	X	
HB1	24	X	X	
HB1	26	X		X
HB1	37	X		X
HB1	45	X		X
HB1	55	X		X
HB1	74		X	X
HB1	77		X	X
HB1	78		X	X
HB1	82		X	X

(continued)

Table B1. Judging plan used to assign participants' performances in the SRLT to the raters (continued).

B2 LB2	14	X		X
LB2	18	X	X	
LB2	29	X	X	
LB2	30	X	X	
LB2	32	X	X	
LB2	47	X		X
LB2	50	X		X
LB2	57	X		X
LB2	59	X		X
LB2	63		X	X
LB2	68		X	X
LB2	75		X	X
HB2	5*	X	X	X
HB2	9	X	X	
HB2	16*	X	X	X
HB2	20	X	X	
HB2	46	X	X	
HB2	49	X	X	
HB2	52	X		X
HB2	53	X		X
HB2	54	X		X
HB2	65	X		X
HB2	71		X	X
HB2	72		X	X
HB2	73		X	X
HB2	76		X	X
HB2	81		X	X

(continued)

Table B1. Judging plan used to assign participants' performances in the SRLT to the raters (continued).

C1	1	X	X	
	2	X	X	
	7	X	X	
	10	X	X	
	11	X	X	
	17	X	X	
	19	X	X	
	21	X	X	
	22	X	X	
	23	X	X	
	25	X	X	
	27	X		X
	28	X		X
	31	X		X
	33	X		X
	34	X		X
	35	X		X
	38	X		X
	39	X		X
	40	X		X
	41	X		X
	42	X		X
	44		X	X
	48		X	X
	56		X	X
	58		X	X
	60		X	X
	61		X	X
	62		X	X
	67		X	X
	69		X	X
	70		X	X
	79		X	X
	80*	X	X	X
* These participants' productions were used in the raters' training and rated by all the raters.				

Table B2. Spearman's correlation matrix for all the variables included in this study.

Variable	Canopy exam scores	SRLT scores	Self-assessment scores	Pronunciation errors (phoneme)	Pronunciation errors (syllable)	Pronunciation errors (word stress)	Pauses	Clauses	Clitic pronouns	Subjunctive tense forms	Lexical errors	Morphological errors	Syntactic errors	Medical terms	Coordinating conjunctions	Subordinating conjuctions	Explained medical terms (definition)	Explained medical terms (synonym)	Explained medical terms (context)
Canopy exam scores		0.74***	0.51***	-0.54***	-0.38***	-0.27*	-0.64***	0.42***	0.53***	0.49***	-0.75***	-0.77***	-0.14	0.03	0.26*	0.36**	-0.03	0.37***	0.28*
SRLT scores			0.66***	-0.65***	-0.3**	-0.31**	-0.59***	0.49***	0.59***	0.53***	-0.79***	-0.69***	-0.23*	0.06	0.33**	0.35**	0.12	0.38***	0.28*
Self-assessment scores				-0.41***	-0.26*	-0.4***	-0.48***	0.43***	0.45***	0.31**	-0.59***	-0.48***	-0.18	0.17	0.33**	0.26*	0.14	0.29**	0.31**
Pronunciation errors (phoneme)					0.32**	0.24*	0.45***	-0.36***	-0.38***	-0.45***	0.59***	0.62***	0.27*	0.08	-0.23*	-0.13	0.12	-0.21	-0.15
Pronunciation errors (syllable)						0.12	0.25*	-0.09	-0.13	0.1	0.3**	0.38***	0.11	0.02	-0.1	-0.03	0.21	-0.2	-0.23*
Pronunciation errors (word stress)							0.27**	-0.01	-0.09	-0.17	0.32**	0.25*	0.36**	-0.1	-0.14	0.02	-0.09	-0.02	-0.12
Pauses								0.01	-0.35**	-0.46***	0.59***	0.46***	0.17	0.17	0.07	-0.04	-0.06	-0.23*	-0.07
Clauses									0.46***	0.26*	-0.33**	-0.38***	0.03	0.21	0.63***	0.74***	0.04	0.22*	0.33**
Clitic pronouns										0.43***	-0.5***	-0.49***	0.07	0.1	0.31**	0.37***	0.16	0.3**	0.22*
Subjunctive tense forms											-0.62***	-0.55***	-0.25*	0.07	0.08	0.16	0.03	0.21	0.12
Lexical errors												0.75***	0.3**	-0.16	-0.29**	-0.2	0.02	-0.3**	-0.28*
Morphological errors													0.24*	-0.06	-0.2	-0.2	0.14	-0.31**	-0.27*
Syntactic errors														-0.09	-0.04	0.06	-0.06	0.08	-0.1
Medical terms															0.27*	0.33**	0.22*	0.05	0.09
Coordinating conjunctions																0.41***	-0.06	0.08	0.18
Subordinating conjuctions																	0.14	0.22*	0.3**
Explained medical terms (definition)																		0.03	-0.28*
Explained medical terms (synonym)																			0.04
Explained medical terms (context)																			

References

ACTFL. (2019). *Making languages our business: Addressing foreign language demand among U.S. employers.* Retrieved February 4, 2025, from https://www.actfl.org/uploads/files/general/MakingLanguagesOurBusiness_FullReport.pdf

Aryadoust, V. (2016). Gender and academic major bias in peer assessment of oral presentations. *Language Assessment Quarterly*, *13*(1), 1-24. https://doi.org/10.1080/15434303.2015.1133626

Aryadoust, V., & Liu, S. (2015). Predicting EFL writing ability from levels of mental representation measured by Coh-Metrix: A structural equation modeling study. *Assessing Writing*, *24*, 35-58. https://doi.org/10.1016/j.asw.2015.03.001

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press.

Bond, T. G., Yan, Z., & Heene, M. (2021). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th edition). Routledge. https://doi.org/10.4324/9780429030499

Bonilla, C. L. (2015). From number agreement to the subjunctive: Evidence for Processability Theory in L2 Spanish. *Second Language Research, 31*(1), 53-74. https://doi.org/10.1177/0267658314537291

Brooks, L., & Swain, M. (2014). Contextualizing performances: Comparing performances during TOEFL iBT™ and real-life academic speaking activities. *Language Assessment Quarterly*, *11*(4), 353-373. https://doi.org/10.1080/15434303.2014.947532

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English for academic purposes speaking tasks* (TOEFL Monograph

No. MS-29). Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2005.tb01982.x

Brown Nielson, K., & DeKeyser, R. (2019). Working memory and planning time as predictors of fluency and accuracy. *Journal of Second Language Studies, 2*(2), 281-316. https://doi.org/10.1075/jsls.19004.bro

Bryfonski, L., & McKay, T. H. (2019). TBLT implementation and evaluation: A meta-analysis. *Language Teaching Research*, *23*(5), 603-632. https://doi.org/10.1177/1362168817744389

Bui, G., & Huang, Z. (2018). L2 fluency as influenced by content familiarity and planning: Performance, measurement, and pedagogy. *Language Teaching Research, 22*(1), 94-114.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language™*. Routledge.

Chapelle, C. A., & Voss, E. (2013). Evaluation of language tests through validation research. In A. J. Kunnan (Ed.), *The companion to language assessment* (Vol. 3) (pp. 1079-1097). John Wiley & Sons. https://doi.org/10.1002/9781118411360.wbcla110

Chapelle, C. A., & Voss, E. (Eds.). (2021). *Validity argument in language testing: Case studies of validation research*. Cambridge University Press. https://doi.org/10.1017/9781108669849

Christiansen, P. (2022, January 26). *R tutorial: Ordinal regression* [Video]. Youtube. https://www.youtube.com/watch?v=rrRrI9gElYA

Christensen, R. (2023). *ordinal-Regression Models for Ordinal Data* (R package version 2023.12-4.1). https://CRAN.R-project.org/package=ordinal

Cuddy, M. M., Liu, C., Ouyang, W., Barone, M. A., Young, A., & Johnson, D. A. (2022). An examination of the associations among USMLE Step 3 scores and the likelihood of

disciplinary action in practice. *Academic Medicine*, *97*(10), 1504-1510. DOI: 10.1097/ACM.0000000000004775

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., & James, M. (2006). *Analysis of discourse features and verification of scoring levels for independent and integrated prototype writing tasks for new TOEFL* (TOEFL Monograph No. MS-30). Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2005.tb01990.x

Cumming, A., Kantor, R., & Powers, D. E. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype tasks: An investigation on raters' decision making and development of a preliminary analytic framework* (TOEFL Monograph Series, Rep. No. 22). Educational Testing Service. http://www.ets.org/Media/Research/pdf/RM-01-04.pdf

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal, 86*(1), 67-96. https://doi.org/10.1111/1540-4781.00137

Dong, T., Zahn, C., Saguil, A., Swygert, K. A., Yoon, M., Servey, J., & Durning, S. (2017). The associations between clerkship objective structured clinical examination (OSCE) grades and subsequent performance. *Teaching and Learning in Medicine*, *29*(3), 280-285. DOI: 10.1080/10401334.2017.1279057

Dursun, A., & Li, Z. (2021). A systematic review of argument-based validation studies in the field of language testing (2000-2018). In C. A. Chapelle & E. Voss (Eds.), *Validity argument in language testing: Case studies of validation research* (pp. 54-70). Cambridge University Press. DOI: 10.1017/9781108669849.005

Educational Testing Service. (n. d.). *TOEFL program history. Volume 6.* Retrieved February 15, 2025, from https://www.ets.org/pdfs/toefl/toefl-ibt-insight-s1v6.pdf

Elder, C., & McNamara, T. (2016). The hunt for "indigenous criteria" in assessing communication in the physiotherapy workplace. *Language Testing*, *33*(2), 153-174. https://doi.org/10.1177/0265532215607398

Enright, M. K., & Schedl, M. (1999). *Reading to learn and reading to integrate* (internal report). Educational Testing Service.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied linguistics, 21*(3), 354-375. https://doi.org/10.1093/applin/21.3.354

Gaillard, S. (2014). *The elicited imitation task as a method for French proficiency assessment in institutional and research settings*. [Unpublished doctoral dissertation]. University of Illinois at Urbana-Champaign.

Garnier, S., Ross, N., Rudis, R., Camargo, A. P., Sciaini, M., & Scherer, C. (2024). *viridis(Lite) - Colorblind-Friendly Color Maps for R* (viridis package version 0.6.5).

Gavin, B. H. Y. (2014). Task readiness: Theoretical framework and empirical evidence from topic familiarity, strategic planning, and proficiency levels. In *Processing perspectives on task performance* (pp. 63-94). John Benjamins.

Gohel, D., Moog, S., & Heckmann, M. (2025). *officer: Manipulations of Microsoft Word and PowerPoint documents* (Version 0.6.8). https://CRAN.R-project.org/package=officer

Gohel, D., & Skintzos, P. (2024). *flextable: Functions for tabular reporting* (Version 0.9.7). https://CRAN.R-project.org/package=flextable

Hale, G. A., Taylor, C., Bridgeman, B., Carson, J., Kroll, B., & Kantor, R. (1996). *A study of writing tasks assigned in academic degree programs* (TOEFL Research Rep. No. 54). Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1995.tb01678.x

Harden, R. M., Stevenson, M., Downie, W. W., & Wilson, G. M. (1975). Assessment of clinical competence using objective structured examination. *British medical journal*, *1*(5955), 447-451. https://doi.org/10.1136/bmj.1.5955.447

Harrell, Jr. F. (2025). *Hmisc: Harrell Miscellaneous* (Version 5.2-3). https://CRAN.R-project.org/package=Hmisc

Harsch, C., Ushioda, E., & Ladroue, C. (2017). Investigating the predictive validity of TOEFL iBT® test scores and their use in informing policy in a United Kingdom university setting. *ETS Research Report Series*, *2017*(1), 1-80. https://doi.org/10.1002/ets2.12167

Hohenstein, C., & Lévy-Tödter, M. (Eds.). (2020). *Multilingual healthcare: A global view on communicative challenges*. Springer Nature.

Isaacs, T., Hu, R., Trenkic, D., & Varga, J. (2023). Examining the predictive validity of the Duolingo English Test: Evidence from a major UK university. *Language Testing*, *40*(3), 748-770. https://doi.org/10.1177/02655322231158550

Isbell, D. R., Crowther, D., Kim, J., & Kim, Y. (2024, October 25). *TOEFL Essentials and speaking performance in academic settings: Investigating the relationships between test tasks and non-test academic tasks* [Conference session]. East Coast Organization of Language Testers (ECOLT) 2024 Conference, Georgetown University, Washington, DC, United States.

Isbell, D. R., Crowther, D., & Nishizawa, H. (2023). Speaking performances, stakeholder perceptions, and test scores: Extrapolating from the Duolingo English test to the university. *Language Testing*, *41*(2), 233-262. https://doi.org/10.1177/02655322231165984

Jung Youn, S. (2023). Test design and validity evidence of interactive speaking assessment in the era of emerging technologies. *Language Testing, 40*(1), 54-60. https://doi.org/10.1177/02655322221126606

Kane, M. T. (1992). An argument-based approach to validity. *Psychological bulletin*, *112*(3), 527-535. https://doi.org/10.1037/0033-2909.112.3.527

Kane, M. T. (2001). Current concerns in validity theory. *Journal of educational measurement, 38*(4), 319-342. http://www.jstor.org/stable/1435453

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of educational measurement*, *50*(1), 1-73. https://doi.org/10.1111/jedm.12000

Kelly-Riley, D., & Elliot, N. (2014). The WPA outcomes statement, validation, and the pursuit of localism. *Assessing writing*, *21*, 89-103. https://doi.org/10.1016/j.asw.2014.03.004

Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing, 35*(4), 477-499. https://doi.org/10.1177/0265532217710049

Komasawa, N., Terasaki, F., Nakano, T., & Kawata, R. (2020). Relationships between objective structured clinical examination, computer-based testing, and clinical clerkship performance in Japanese medical students. *PLoS One 15*(3), 1-8. DOI: 10.1371/journal.pone.0230792

Kumazawa, T., Shizuka, T., Mochizuki, M., & Mizumoto, A. (2016). Validity argument for the VELC Test® score interpretations and uses. *Language Testing in Asia*, *6*, 1-18. https://doi.org/10.1186/s40468-015-0023-3

Krusen, N. E., & Martino, M. N. (2020). Occupational therapy students' perceptions of OSCE: A qualitative descriptive analysis. *Journal of Occupational Therapy Education, 4*(1), 7. https://doi.org/10.26681/jote.2020.040107

Lado, R. (1961). *Language testing*. Longman.

LaFlair, G. T., & Staples, S. (2017). Using corpus linguistics to examine the extrapolation inference in the validity argument for a high-stakes speaking assessment. *Language Testing*, *34*(4), 451-475. https://doi.org/10.1177/0265532217713951

Li, Z. (2015). Using an English self-assessment tool to validate an English Placement Test. *Papers in language testing and assessment*, *4*(1), 59-96. DOI: 10.58379/RPHI9026

Linacre, J. M. (1994). *Many-facet Rasch measurement* (2nd edition). MESA Press.

Linacre, J. M. (2024a). *Facets computer program for many-facet Rasch measurement* (version 4.2.4). https://www.winsteps.com

Linacre, J. M. (2024b). *A user's guide to FACETS Rasch-model computer programs* (program manual 4.1.8). https://www.winsteps.com/index.htm

Linacre, J. M. (2025, January 20). *Facets Tutorial PDFs.* Winsteps. https://winsteps.com/facetman/tutorialpdfs.htm

Llosa, L., & Malone, M. E. (2019). Comparability of students' writing performance on TOEFL iBT and in required university writing courses. *Language Testing*, *36*(2), 235-263. https://doi.org/10.1177/0265532218763456

Long, M. H. (1985). A role for instruction in second language acquisition: Task-based language teaching. *Modelling and assessing second language acquisition*, *18*(1), 77-99.

Long, M. H. (2015). *Second language acquisition and task-based language teaching*. John Wiley & Sons.

Macqueen, S., Pill, J., & Knoch, U. (2016). Language test as boundary object: Perspectives from test users in the healthcare domain. *Language Testing*, *33*(2), 271-288. https://doi.org/10.1177/0265532215607401

Macqueen, S., Pill, J., & Knoch, U. (2021). Trust the test: Score-user perspectives on the roles of language tests in professional registration and skilled migration. *Papers in Language Testing and Assessment*, *10*(1), 49-69. DOI: 10.58379/MAKH4553

Malone, M. E., & Montee, M. (2014). Stakeholders' beliefs about the TOEFL iBT® test as a measure of academic language ability. *ETS Research Report Series*, *2014*(2), 1-51. https://doi.org/10.1002/ets2.12039

Mangiafico, S. S. (2024). *rcompanion: Functions to Support Extension Education Program Evaluation* (Version 2.4.36). Rutgers Cooperative Extension. New Brunswick, NJ. https://CRAN.R-project.org/package=rcompanion

McNamara, T. F. (1996). *Measuring second language performance*. Longman.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Macmillan.

Meyer, D., Zeileis, A., Hornik, K., Friendly, M. (2024). *vcd: Visualizing Categorical Data* (R package version 1.4-13). https://CRAN.R-project.org/package=vcd

Needs analysis. The Canopy test of medical Spanish (n.d.). [Unpublished manuscript].

Nguyen, T. H. H., Nguyen, B. T. T., Hoang, G. T. L., Pham, N. T. H., & Dang, T. T. C. (2024). Computer-delivered vs. face-to-face score comparability and test takers' perceptions: The case of the two English speaking proficiency tests for Vietnamese EFL learners. *Language Testing in Asia*, *14*(1), 6. https://doi.org/10.1186/s40468-024-00277-1

Norris, J. M. (2009). Task-based teaching and testing. In M. Long & C. J. Doughty (Eds.), *Handbook of language teaching* (pp. 578–594). Blackwell.

Norris, J. M. (2016). Current uses for task-based language assessment. *Annual Review of Applied Linguistics*, *36*, 230-244. https://doi.org/10.1017/S0267190516000027

Norris, J. M. (2018). Task-based language assessment: Aligning designs with intended uses and consequences. *JLTA Journal, 21*, 3-20. DOI: 10.20622/jltajournal.21.0_3

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied linguistics, 30*(4), 555-578. https://doi.org/10.1093/applin/amp044

Oh, S. R. (2018). *Investigating test-takers' use of linguistic tools in second language academic writing assessment*. [Unpublished doctoral dissertation]. Teachers College, Columbia University.

Ortega, P., Cisneros, R., & Park, Y. S. (2023). Spanish medical jargon: A new metric for improving patient-centered communication with Spanish-speaking patients. *Patient Education and Counseling, 109*, 107644. https://doi.org/10.1016/j.pec.2023.107644

Qian, D. D. (2009). Comparing direct and semi-direct modes for speaking assessment: Affective effects on test takers. *Language Assessment Quarterly*, *6*(2), 113-125. https://doi.org/10.1080/15434300902800059

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* MESA Press.

R Core Team. (2024). *R: A Language and Environment for Statistical Computing* (Version 4.4.2). R Foundation for Statistical Computing. https://www.R-project.org/

Real Academia Nacional de Medicina de España. (2012). *Diccionario de términos médicos. Editorial médica panamericana.* https://dtme.ranm.es/index.aspx

Revelle, W. (2024). *psych: Procedures for Psychological, Psychometric, and Personality Research* (R package version 2.4.12). Northwestern University. https://CRAN.R-project.org/package=psych

Riazi, A. M. (2016). Comparing writing performance in TOEFL-iBT and academic assignments: An exploration of textual features. *Assessing Writing*, *28*, 15-27. https://doi.org/10.1016/j.asw.2016.02.001

Rosenfeld, M., Leung, P., & Oltman, P. K. (2001). *The reading, writing, speaking, and listening tasks important for academic success at the undergraduate and graduate levels* (TOEFL Monograph No. 21). Educational Testing Service. http://www.ets.org/Media/Research/pdf/RM-01-03.pdf

Ross, S. J. (2017). *Interviewing for language proficiency*. Palgrave Macmillan.

Santos, V. D. (2017). *A computer-adaptive test of productive and contextualized academic vocabulary breadth in English (CAT-PAV): Development and validation.* [Unpublished doctoral dissertation]. Iowa State University.

Schlegel, B., & Steenbergen, M. (2020). *brant: Test for Parallel Regression Assumption* (R package version 0.3-0). https://CRAN.R-project.org/package=brant

Sokolowski, J. A., & Banks, C. M. (Eds.). (2009). *Principles of modeling and simulation: A multidisciplinary approach*. John Wiley & Sons.

Statistics Solutions. (2025, February 4). *Ordinal regression.* https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/ordinal-regression/

Tennant, A., & Pallant, J. (2006). Unidimensionality matters! (A tale of two Smiths?). *Rasch Measurement Transactions, 20,* 1048-1051. Retrieved January 26, 2025, from https://www.rasch.org/rmt/rmt201c.htm

The theoretical framework. The Canopy test of medical Spanish (n.d.). [Unpublished manuscript].

Toulmin, S. E. (2003). *The uses of argument*. Cambridge University Press.

United States Medical Licensing Examination. (2025, February 16). *Computer-based case simulations.* Retrieved February 16, 2025, from https://www.usmle.org/step-3-test-question-formats/computer-based-case-simulations

Validation Study of Canopy Credential, a Test of Spanish Proficiency in the Medical Domain (n.d.). [Unpublished manuscript].

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V.,… Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4* (43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A Grammar of Data Manipulation* (R package version 1.1.4). https://CRAN.R-project.org/package=dplyr

Will, F. L. (1960). [Review of the book *The uses of argument,* by S. E. Toulmin]. *The Philosophical Review, 69*(3), 399–403. https://doi.org/10.2307/2183556

Wind, S. A., & Engelhard Jr, G. (2013). How invariant and accurate are domain ratings in writing assessment?. *Assessing Writing, 18*(4), 278-299. http://dx.doi.org/10.1016/j.asw.2013.09.002

Woodward-Kron, R., & Elder, C. (2016). A comparative discourse study of simulated clinical roleplays in two assessment contexts: Validating a specific-purpose language test. *Language Testing*, *33*(2), 251-270. https://doi.org/10.1177/0265532215607399

Youn, S. J. (2013). *Validating task-based assessment of L2 pragmatics in interaction using mixed methods.* [Unpublished doctoral dissertation]. University of Hawai'i at Manoa.

Zhang, Y. (2006, October 13). *Statistical review of the TOEFL iBT April-June 2006 administrations.* [Paper presentation]. TOEFL Committee of Examiners Meeting, Philadelphia, PA, United States.