



EtoE Interpreter Testing Study Report

May 15, 2021

Published by:
Certification Commission for Healthcare Interpreters
1725 I Street NW, Suite 300
Washington, DC 20006
www.cchicertification.org

Copyright ©2021, Certification Commission for Healthcare Interpreters
Published in the United States of America

Acknowledgements

This *Report* and administration of the CHI™ and ETOE™ examinations were conducted on behalf of the Certification Commission for Healthcare Interpreters (CCHI) by Prometric LLC, under the supervision of the psychometrician **Oksana Naumenko**, the author of *Part II* of this report. This document is copyrighted and is intended for the sole use of CCHI.

This study could not have been completed without the effort and diligence of the principal investigator for the study **Natalya Mytareva**, M.A., CoreCHI™, CCHI Executive Director, the author of *Part I* and *III* of the report, and the advice and counsel of **James P. Henderson**, Ph.D., Credentialing Examination Consulting, LLC.

This study is made possible in part by a grant from the **Robert Wood Johnson Foundation**, by contributions from **Certified Languages International, Cross Cultural Communication Systems, Inc™**, other corporate and individual donors. CCHI thanks all the contributors and supporters.

CCHI thanks all the experts¹ who, throughout the years of this project, generously contributed their time, energy, and expertise toward the completion of this study.

CCHI Commissioners

Margarita S. Bekker, CoreCHI™, Lead Interpreter (Russian), Education and Training, Stanford University Medical Center (CA) (Commissioner term ended April 2021)

Jaime Fatás-Cabeza, USCCI, English/Spanish, CHI™-Spanish, Director of the Undergraduate Program in Translation and Interpretation, Department of Spanish and Portuguese, University of Arizona (AZ)

Vonessa Costa, CoreCHI™, Director of Multicultural Affairs and Patient Services, Cambridge Health Alliance (MA)

H. Valerie Huang, M.A., CHI™-Mandarin, Language Services Manager, Nationwide Children's Hospital (OH); CCHI Secretary

Dylanna Jackson, M.A., Director, International Institute of Erie, PA Office of USCRI (PA); CCHI Public member

Mina Kini, M.S.W., M.S., Language Access Consultant (TX), CCHI Treasurer

Eliana Lobo, M.A., CoreCHI™, Language access policy consultant and medical interpreter trainer, Lobo Language Access (WA)

Francisco J. Martinez, M.A., CHI™-Spanish, Language Coach & Medical Interpreter, Children's Mercy Hospitals and Clinics (CMHC) (MO)

Idolly F. Oliva, M.B.A., Manager, Interpreter Services, Fairview Health (MN), CCHI Chair

Edna Y. Quartey, CHI™-Spanish, Spanish interpreter, (MI) (Commissioner term ended October 2020)

Erin Rosales, B.A., CPLP®, CHI™-Spanish, VP and Director of Interpreter Development, Connecting Cultures, Inc (WI)

Mateo C. Rutherford, M.A., MATI, CHI™-Spanish, Technology and Systems Manager, Interpreting Services Dept., UCSF Health (CA); CCHI Vice Chair

E. Zoe Schutzman, MA, NYS CCI, CHI™-Spanish, Program Manager / Educator & Staff Development Specialist II, University of New Mexico Hospitals - Health Sciences Center (NM)

Jorge U. Ungo, Strategic Healthcare Account Executive, LanguageLine Solutions (TX)

¹ See the list of the members of the EtoE National Task Force in *Appendix A*.

CCHI expresses its gratitude to the volunteer subject matter experts (SMEs) who helped develop the ETOE™ exam content and rating scales. The dedication of these individuals – who volunteered countless hours to CCHI – was critical in creating the ETOE™ examination:

Sura Al Khalidi, CHI™- Arabic (CA)
Mutaz Al Mudaris, CHI™- Arabic (PA)
Yasmeen Albarghouthi, CHI™- Arabic (OR)
Chi-Wei Chang, R.N., CHI™- Mandarin (MI)
Carla Chavez Moreno, Ph.D. (MD)
Suzanne Couture, CHI™-Spanish (WI)
Ping Cross, CHI™- Mandarin (TN)
Jose Cruz, CHI™- Spanish, CMI (TN)
Jaime Diaz, CHI™- Spanish (IL)
Maria Donley, CHI™- Spanish (IL)
Lois Feuerle, J.D., Ph.D. (OR)
Jeffrey Gabbitas, Ph.D. (AZ)
Emil Gilmanov, CMI, Russian (MN)
Elisa Gustafson, CHI™- Spanish (MN)
Erika Hernandez, CHI™- Spanish (NC)
William (Will) Hester, CHI™- Spanish, CMI (IL)
Jacqueline Hinds, CHI™- Spanish (OR)
Dan Kristie, B.A. (PA)
Katherine Langan, Ph.D., CHI™- Spanish (IA)
Sandy Maloney, CHI™- Spanish (MN)
Myrlande Merisme, Haitian Creole interpreter (NY)
Laura Neri, CHI™- Spanish (CA)
Laura Onofre, CHI™- Spanish (CA)
Linda Pollack-Johnson, CoreCHI™, Italian interpreter (PA)
Myriam Prias, CHI™- Spanish (FL)
Ronda Rankin, CoreCHI™, NIC, ASL interpreter (NE)
Rosemary Rodriguez, CHI™- Spanish (VA)
Karin Ruschke, M.A., German interpreter (IL)
Aida Schneider, CHI™- Arabic (OH)
Tatyana Sorensen, CoreCHI™, Russian interpreter (UT)
Sarah Stockler-Rex, CHI™- Spanish (OH)
Indira Sultanic, Ph.D., CHI™- Spanish (VA)
Qing Xiu (Linda) Sun, CHI™- Mandarin (NV)
Manyee Tang, CHI™- Mandarin, CMI (MA)
Icarus Tsang, Psy.D. (2021) (CA)
Rio Zamarron, MHPE (MO)

Contents

Acknowledgements	2
Introduction	6
Part I. EtoE Examination Development	7
Selection Process of Subject Matter Experts.....	7
Test Content Development	7
Test Blueprint and Scoring Rubrics Development	10
Rater Training	11
Study Participants Recruitment and Demographics.....	12
Who Participated in the Study?	12
Part II. EtoE Study Follow-up Validation Report by Prometric LLC	19
INTRODUCTION	20
ETOE MEASURE DEVELOPMENT BACKGROUND	20
Examination Form	21
Rubric and Rater Handbook	21
EtoE Administration	23
EtoE Scoring	23
RESEARCH QUESTIONS	23
RESULTS	24
DATA INTEGRITY	25
ITEM ANALYSIS.....	25
ETOE MEASURE DESCRIPTIVE STATISTICS	26
RATER AGREEMENT	29
RUBRIC SCALE SUMMARIES.....	30
RESEARCH QUESTION RESULTS	33
Exploratory Pearson Correlations between CHI™ Scores and EtoE Scores and between Each Language-specific Set of CHI™ Scores and EtoE Scores	33
Logistic Regression of CHI™ Exam Passing Status on the Weighted EtoE Score	35
Simple Linear Regression of the CHI Scores on the Weighted Item Type Score	36
Confirmatory Factor Analyses for the EtoE Measure of Interpreting Ability	37
CONCLUSIONS	38
Part III. EtoE Study Additional Observations	40
Professional Affiliation and Experience.....	40
English Language Acquisition	42
Healthcare Interpreting Education.....	43
Additional Confirmatory Analyses.....	46
Conclusion	48

Appendix A. Designing an English-to-English Interpreting Performance Test: Recommendations of the National Task Force (January 2019)	49
Acknowledgements.....	49
Contents.....	50
Healthcare Interpreter Competencies and Job Tasks.....	50
Test Item Types.....	51
Item Content (Scripts and Texts).....	74
Scoring Recommendations	74
Test Taker Questionnaire and Preparation Guide.....	75
Conclusion.....	76
Appendix B. Performance Item Template for the EtoE Examination	77
Appendix C. Initial EtoE Item Type General Review Form	78
Appendix D. Final EtoE Item Type General Review Form	80
Appendix E. EtoE Study Participation Questionnaire	82
Appendix F. ETOE™ Examination Guide	88
Appendix G. Sample Rubric: Reading Comprehension	98
Appendix H. Item Analysis Indices Description	101
Appendix I. Standardized Factor Loadings from CFA Models 1, 2, 3	102

Introduction

The Certification Commission for Healthcare Interpreters (CCHI) has administered the national certification programs for healthcare interpreters since 2010. The two currently available certifications – Core Certification Healthcare Interpreter™ (CoreCHI™) and Certified Healthcare Interpreter™ (CHI™ - Spanish, Arabic, Mandarin) – are aimed at the entry-level healthcare interpreter.

As part of a continuous evaluation process of its certification programs, CCHI reviewed the current knowledge regarding valid and efficient assessment of interpreting skills of interpreters of *any* language. The goal is to explore if cognitive interpreting skills can be measured via a standardized oral performance monolingual test in English (English-to-English, EtoE) so that this test can be used for interpreters of less common languages (i.e., languages of lower incidence) for whom creating a separate dual-language oral performance exam is unfeasible.

The EtoE project consisted of four phases:

- I. Feasibility Review: Discussions with stakeholders and focus groups² – Fall of 2017
- II. Study Design: The EtoE National Task Force and consultations with psychometricians – 2018
- III. EtoE Exam Creation: by CCHI’s volunteer SMEs and Prometric, LLC – 2019
- IV. EtoE Study: Test delivery to study participants and study results report – 2020-21.

The Covid-19 pandemic and ensuing test site closures in the U.S. in the spring of 2020 interrupted the study. As a result, the Commissioners extended testing until November 2020 and the analysis until spring 2021. A preliminary report was published in May 2020, to provide some preliminary results.³

This *Report* provides results of the comparison of CCHI candidates’ performance on the monolingual EtoE exam to the dual-language CHI™ certification exam. These findings offer data-based evidence to the possibility of measuring cognitive interpreting skills responsible for a successful conversion of meaning from one language into another in a monolingual format.

A total of 249 interpreters of Arabic, Mandarin and Spanish took the EtoE and CHI™ examinations between January 24 and November 3, 2020, with 247 of them also completing the *EtoE Study Participant Questionnaire* (*Appendix E*). All 249 participants have completed both exams, and 177 of those had the full sets⁴ of the exam items completed.

The report consists of *Part I* describing the EtoE exam development and study participants, *Part II* containing the psychometric analyses prepared by Prometric LLC for CCHI, *Part III* providing some additional observations relevant for interpreter educators and the profession at large, and *Appendices*. *Appendix A* is the National Task Force *Recommendations on Designing the English-To-English Interpreting Performance Test*. *Appendix B* is a sample *Performance Item Template for the EtoE Examination* that has been used by SMEs to write items for the exam. The item review forms are provided in *Appendices C* and *D*. *Appendix E* contains the *EtoE Study Participation Questionnaire*. *Appendix F* is the *EtoE Examination Guide* that participants were encouraged to review before taking the exams in order to familiarize themselves with the test tasks; it also contains sample exam items. *Appendix G* provides an example of the scoring scales used by raters to score the EtoE exam. *Appendix H* clarifies some statistical terms used in Prometric’s report of *Part II*, while *Appendix I* contains loading factors of the three confirmatory factor analysis models.

² See the resulting whitepaper *Assessing Healthcare Interpreting Performance Skills in and English-to-English Format* at https://cchicertification.org/uploads/CCHI_EtoE_Interpreter_Performance_Assessment.pdf

³ The full text is at https://cchicertification.org/uploads/CCHI-ETOE_Study_Preliminary_Report.pdf, accessible from the EtoE Project webpage at <https://cchicertification.org/etoe/>.

⁴ Incomplete exams occurred due to technical issues or user errors during test administration.

Part I. EtoE Examination Development

Selection Process of Subject Matter Experts

In developing the EtoE examination for the study, CCHI followed the same procedures for test construction it employs for developing and maintaining its certification exams. In all steps of the process CCHI is guided by the best practices and standards of the National Commission for Certifying Agencies (NCCA).

In January 2019, CCHI recruited volunteer subject matter experts (SMEs) via a public announcement on its website,⁵ through social media, and direct emails to CCHI's certificants and e-news subscribers. The Test Development Steering Committee selected 36 SMEs⁶ for three phases of the process: item writing (16 SMEs), item review (22 SMEs), and test form/scoring development (12 SMEs).

Ten of the SMEs participated in more than one of the phases to ensure the continuity of the process. Also, ten of the SMEs have been previously involved in the development of CCHI's certification examinations. 28 SMEs are CCHI-certified interpreters; six SMEs are not practicing interpreters (they are interpreter educators or interpreter managers), one is a certified court interpreter, and one is a novice interpreter preparing for certification.

By language diversity, the SMEs represent interpreters and speakers of Arabic, ASL, Bosnian, Cantonese, French, German, Italian, Haitian Creole, Mandarin, Russian, and Spanish. By language acquisition, eleven SMEs were native speakers of English, five speak more than two languages, eight have been raised bilingual. Geographically, the SMEs represent these countries of origin: Argentina, Bosnia, Colombia, Costa Rica, Germany, Haiti, Hong Kong, Iraq, Jordan, Mexico, Palestine, People's Republic of China, Russia, Puerto Rico, Taiwan, and the USA, and the following states of residence in the U.S.: AZ, CA, FL, IA, IL, MA, MD, MI, MN, MO, NC, NE, NV, NY, OH, OR, PA, TN, UT, VA, and WI.

The SMEs represent all modalities of interpreting – in-person, over-the phone and video remote, and both free-lance and staff healthcare interpreters. They interpret for a variety of settings – hospitals, clinics, public health, insurance companies. Most have additional experience of either interpreting in a non-healthcare setting (education, court) or translating or language teaching. Among the SMEs are representatives of other healthcare professions (nurse, physician, psychologist, and social worker). Their general education level varies from Associate degree to Ph.D./J.D., and the years of healthcare interpreting range from 2 to over 20.

All the SMEs signed the conflict of interest and security agreements and adhered to CCHI's test development security procedures.

Test Content Development

One of the differences of the EtoE exam of this study from a usual certification exam that has influenced the test development process is its dual purpose. This EtoE exam is meant to allow CCHI to:

- explore a potential correlation between pass/fail results on this monolingual exam and on the existing dual-language CHI™ exams, and
- identify which item *types* have a potential correlation and/or predictive ability, so that in case of a positive correlation a future certification exam could be developed incorporating these types of items.

⁵ See at <https://cchicertification.org/volunteer-to-create-etoe-test/>.

⁶ See their names on pp. 2-3 in the *Acknowledgements* section of this report.

Based on the National Task Force *Recommendations on Designing the English-To-English Interpreting Performance Test (Appendix A)*, the SMEs were tasked with developing items of the following ten types listed in Table 1.

Table 1. Item Types Recommended for the EtoE Exam by the National Task Force

Type	Mode of Input
Reading Comprehension	Text-to-Audio
Shadowing	Audio-to-Audio
Finish the Sentence	Audio-to-Audio
Restate the Message	Audio-to-Audio
Listening Comprehension	Audio-to-Audio
Memory	Audio-to-Audio
Equivalence	Mode of Input
Medical Concepts	Multiple-choice
Fill-in-the-Blank	Mode of Input
Bilingual Reformulation*	

*Due to the technical limitations of test delivery, this item type is not included in the current EtoE test and study. However, the items of this type were created, and CCHI may utilize them in its future test development work.

A performance item for the EtoE exam (similarly to the CHI™ exam items) consists of five components:

- the directions to the test taker,
- the script, i.e., the content that candidates will be manipulating as instructed,
- the audio recording of the script (for most item types),
- the timing of the item, i.e., an estimate of how many seconds a minimally competent interpreter should spend to successfully complete the task,
- the scoring guide – a document for human raters which contains possible candidate responses, agreed-upon conventions of assigning scores to specific versions of the performance, and references on the item’s subject matter.

Training of SMEs

All SMEs participated in the live virtual training facilitated by the principal investigator Natalya Mytareva, who is CCHI’s chief Test Development and Content Management Officer. For item writers and reviewers, the training consisted of three synchronous two-hour calls held on February 2, 6, and 9, 2019. The training covered the following topics and utilized the procedures used by CCHI’s SMEs who develop the other certification exams:

- Introductions and meeting objectives
- Overview of the EtoE Project, purpose of the EtoE and CHI™ examinations
- CCHI’s certification target audience
- Item and Test Development Process overview and goals
- SMEs’ responsibilities
- How-to Primer for writing and reviewing performance items
- EtoE National Task Force Recommendations
- Performance Item Writing Guide
- CCHI’s Performance Item Conventions

Item reviewers also participated in the fourth synchronous two-hour training (February 13, 2019) which, in addition to defining their group’s specific objectives and responsibilities, included these topics:

- How to Review Performance Items
- Overview of Scoring Models

SMEs of the test form/scoring development group were required to review the recordings of the previous SME trainings before participating in the synchronous meetings.

All SMEs have read the National Task Force *Recommendations on Designing the English-To-English Interpreting Performance Test (Appendix A)*.

Item Writing

The item writing was conducted remotely, both synchronously and asynchronously. After the virtual synchronous SME training meetings, SMEs received individual writing assignments. The principal investigator assigned to each SME the *types* of items and the number of items per each type that they were supposed to create. In order to ascertain the diversity of the items content, the assignments also specified which healthcare specialties could be addressed in items. This distribution was based on the SMEs' indicated preference.

The SMEs utilized an item writing template CCHI has used for developing its other performance exams, slightly modified for this project (see *Appendix B*). In addition to creating a script for the item, SMEs also created possible model responses for three different levels:

- experienced/skilled interpreter (highest score),
- minimally competent interpreter (passing score),
- not-yet competent interpreter or non-interpreter (failing score).

This was done in order to both develop a more robust item and to lay the foundation for the scoring method and rubrics for that item type.

The SMEs created their items independently of one another in the course of 10-12 days, communicating as needed with the principal investigator. The items were submitted via a secure cloud-based platform, and CCHI's staff performed a preliminary editorial review and confirmed the accuracy of the content (reference checks). All items were then distributed to the SMEs for asynchronous review and comments.

The SMEs proceeded to improve the items during four synchronous two-hour meetings on February 16, 20, 23, and 27, 2019. The item writers concluded their participation in the project by completing online the *Initial EtoE Item Type General Review Form* (see *Appendix C*).

Item Review

The second group of SMEs – item reviewers – were, first, divided into three groups: two groups were assigned to review three item types (out of ten initial item types), and one group was assigned four item types. The grouping was based on the commonality of skills that various item types were designed to measure, e.g., Reading and Listening Comprehension items were grouped together, as were the Restate the Meaning and Medical Equivalence items. Separating the review into three groups allowed for a more in-depth, focused analysis of the items themselves, the possible model responses, and laid the foundation for the final selection of the items for the test form.

The SMEs started their work by independently reviewing the assigned items and completing the following steps:

- a. timing themselves performing the task required by each item,
- b. offering edits to the item scripts,
- c. completing the online *Initial EtoE Item Review Form (Appendix C)*.

The principal investigator compiled the SME's responses and comments and eliminated from further review the items that were marked by all reviewers as weak. Then, each of the three groups met virtually for a two-hour meeting facilitated by the principal investigator (on March 2, 16, and 23, 2019) to finalize the review of their items.

For the next step of this phase, the reviewers met as one group to complete the item development process. These virtual two-hour meetings took place on March 28, 30, and April 6, 2019. During the meetings, the SMEs evaluated each item by applying this *Item Validation Checklist*:

1. Is the item content representative of the job of an entry-level healthcare interpreter? Is the subject matter (topic) significant and relevant for the profession?
2. Is the level of the speech/language complexity representative of the speech/language encountered in real-life healthcare interactions between patients and providers?
3. Is the item free of inconsistencies (regional, cultural, educational level, etc.)?
4. Does the item provide reasonable opportunity for success regardless of the candidate's regional or cultural background and regardless of which language is their native one?
5. Does the item provide sufficient linguistic material to measure the key skill/subskill it is intended to measure?
6. On the scale of 1-10 (ten being the most difficult), assess the item's difficulty.

The SMEs then approved the final version of the item scripts for recording. CCHI engaged volunteers (SMEs who participated in the CHI™ exams development) to record the items. The following criteria were applied to the recordings: presence of female/male voice talents, neutrality of the native-English-speaker accents, clear diction of voice talents, and conversational pace of speech at a rate of approximately 110–150 words per minute.

The SMEs' asynchronous work consisted of:

- a. timing themselves performing the task required by each item,
- b. completing online the *Final EtoE Item Type General Review Form* (see *Appendix D*).

The SMEs made final testing decisions reflected in the directions to test takers such as how many times candidates are allowed to play the audio for a specific item type, and if notetaking is allowed. During the last meeting, 33 items were selected for the test form construction. The SMEs made recommendations regarding the test form construction and development of scoring rubrics.

Test Blueprint and Scoring Rubrics Development

CCHI contracted Prometric LLC to develop the test blueprint, complete the development of scoring rubrics, build the EtoE exam, and administer it to candidates.

The test blueprint and scoring rubric development was facilitated by Prometric's staff - Krystal Fitzgerald, M.M., M.Ed., Measurement and Testing Technical Advisor, and Oksana Naumenko, Psychometrician. The two-hour SME meetings took place on November 8, 14, 15, and 20, 2019 via teleconference.

The description of this phase is provided in *Part II*, Section "EtoE Measure Development Background" of this report.

The final step of the SMEs' asynchronous work (via online forms and email communication) consisted in approving the item recordings, finalizing the timing of all items, and approving the overall time for the EtoE exam and the directions to test takers.

Rater Training

CCHI recruited human raters for the EtoE exam following the same process and requirements as for the CHI™ certification exams⁷. A total of 17 raters were selected initially, and 14 of them completed the rater training and rated the candidate responses.

The raters' panel has the following characteristics:

- Ten raters are currently practicing healthcare interpreters; four raters have been practicing healthcare interpreters in the recent past and are now either interpreter educators or industry advisors.
- Three raters are native English speakers, two are “true” bilinguals (raised speaking English and the other language), and nine have a near-native command of English.
- They represent the following languages of interpreting: Arabic, German, Italian, Mandarin, Russian, and Spanish (of different regional variants).
- Some of the raters are current raters of the CHI™ exams, and seven raters participated in the development of the EtoE exam and scoring rubrics.

CCHI's principal investigator Natalya Mytareva created and conducted the rater training. Throughout the training, the raters communicated with the principal investigator via email and phone calls as needed.

The rater training consisted of two phases:

1. self-paced, asynchronous completion of three online training modules within 3 weeks, and
2. five virtual synchronous meetings held as two-hour teleconferences on February 5, 6, 11, 12, and 19, 2020.

The meetings were recorded, and the recordings were available to the raters throughout the rating process.

The online Rater Training Module 1 consisted of one video lecture, a video recording of the webinar for the EtoE test takers, a self-assessment questionnaire, and a quiz. The objectives for Module 1 are:

- understand CCHI's rater's responsibilities,
- know the principles of analytic rubric rating method,
- know the purpose and structure of the EtoE exam.

The online Rater Training Module 2 consisted of one video lecture, self-study of the *EtoE Rater Handbook*, and a quiz, with the objective that raters will be able to apply the scoring scales of the EtoE exam.

The online Rater Training Module 3 consisted of raters applying the scoring scales to the model responses (audio recordings) with various automated feedback if their scores did not match the scores pre-determined by SMEs for these responses. Raters were required to repeat the assignments of this module and review previous modules as needed until their score matched the pre-determined score for each item. Raters communicated with the principal investigator via email as needed.

The second phase of the training consisted of rating the actual candidate responses (anchors) pre-selected by the principal investigator from the pool of the exams that had already been completed by that time. Several responses at different level of performance were selected for each item. First, raters would listen to the response and record their score on a specific scale, then their scores are revealed, and raters are asked to explain their score. The process is repeated until the consensus is reach for each item. Through this iterative process the raters learn to apply the scoring scales correctly and consistently. Some responses resulted in the necessity to create specific “rating conventions” for that item. For example, how to rate an item if a candidate did not follow the directions correctly, or if a candidate's response was in their non-English language.

⁷ See CCHI's webpage for details: <https://cchicertification.org/volunteer-for-cchi/>.

Study Participants Recruitment and Demographics

CCHI recruited volunteer participants for the EtoE Study among Arabic, Mandarin and Spanish interpreters via direct email to over 1,000 candidates who had not taken the dual-language CHI™ certifications exam, and to its 15,000 newsletter subscribers, via website⁸ and social media announcements, and via direct contact with interpreters at interpreter conferences throughout 2019.

All study participants had to have been deemed eligible for CCHI’s certification (see CCHI’s eligibility requirements at <https://cchicertification.org/certifications/eligibility/>) prior to participating in the study. The participants were provided an incentive in the form of a \$100 discount off the CHI™ exam fee. The candidates were asked to complete the *EtoE Study Participant Questionnaire* (see *Appendix E*) before they scheduled the testing appointment. All responses were self-reported and have not been verified by CCHI.

The information provided in this part of the *Preliminary Report* is based on the responses of 247 participants (of 249 total) who completed the *EtoE Study Participant Questionnaire*, with the exception of the category “Language of interpreting” for which the information for all 249 participants is present. However, the comparisons involving the ETOE™ and CHI™ exams are based on 176 responses (one participant of the 177 complete sets of both exams has not submitted the *Questionnaire*).

Who Participated in the Study?

The demographic data collected about the study participants is consistent with the demographics of the respondents to our national *Job Task Analysis Survey*⁹ in 2016, confirming that the study sample is representative of currently practicing healthcare interpreters.

The study was designed for interpreters of Arabic, Mandarin and Spanish, and participants reflected the distribution of these languages among candidates of the CHI™ certification program¹⁰ (Table 2).

Table 2. Language of interpreting

Language	Study Count	Study Percent	2020 Count	2020 Percent	2018* Count	2018 Percent
Arabic	31	13%	113	12%	185	18%
Mandarin	23	9%	58	6%	77	8%
Spanish	195	78%	785	82%	744	74%
Total	249	100%	956	100%	1,006	100%

*2019 data is omitted because exams were administered only for two testing windows instead of four.

The first question asked the participants if interpreting or translation was their **main profession** (means of earning a living). 177 participants responded “yes,” and for 70 participants interpreting or translation was secondary occupations.

Regarding **age**, the largest number of participants falls into the range of 31 to 40 years. CCHI requires certificants to be at least 18 years old (Table 3).

⁸ See CCHI’s webpage at <https://cchicertification.org/etoe/register-for-etoe-exam/>.

⁹ P. 9 of the Report on CCHI’s 2016 Job Task Analysis Study at https://cchicertification.org/uploads/CCHI_JTA2016_Report.pdf.

¹⁰ See CCHI’s Annual reports at <https://cchicertification.org/about-us/annual-reports/>.

Table 3. What is your age?

Age Group	Count	Percent
18 to 20 years	2	1%
21 to 30 years	53	21%
31 to 40 years	68	28%
41 to 50 years	66	27%
51 to 60 years	43	17%
61 years and over	15	6%
Total	247	100%

The next demographic question asked about **gender identity**. 74% (182) of participants were female, and 26% (65) - male.

The largest number of participants (95) stated that they currently **work** as a freelancer (contractor); however, almost as many (82) reported being a staff interpreter (Table 4). It is important to keep in mind that participants could work as supervisors and trainers of healthcare interpreters, meaning, participants indicating that they do not interpret in healthcare settings are still considered qualified candidates.

Table 4. What is your current employment status in relation to healthcare interpreting?

Employment Status	Count	Percent
I am a staff interpreter	82	33%
I am a freelancer (contractor)	95	38%
I am a volunteer	17	7%
I'm a dual-role interpreter, with interpreting as a secondary responsibility	26	11%
I don't interpret in healthcare settings	27	11%
Total	247	100%

Table 5 describes the **types of settings** respondents interpret in (multiple responses were accepted). 63% indicated that they interpret regularly in healthcare settings, however, only 22% interpret in health care exclusively.

Table 5. In what settings do you interpret regularly? Select ALL that apply

Type of settings	Count	Percent
Healthcare	54	22
Healthcare/other	101	41
Business	2	1
Business/other community	1	0
Conference/other community	1	0
Legal/schools	5	2
Telephone/video (all settings)	58	23
Other	7	3
I'm mostly a translator	9	4
None of the above	11	4
Total	247	100%

Two questions asked the participants about their **experience in healthcare interpreting**: the objective question referred to the specific number of years they have been interpreting (Table 6), and the subjective one referred to their personal perception of their experience across the professional continuum (Table 7).

Table 6. How many years of healthcare interpreting experience do you have?

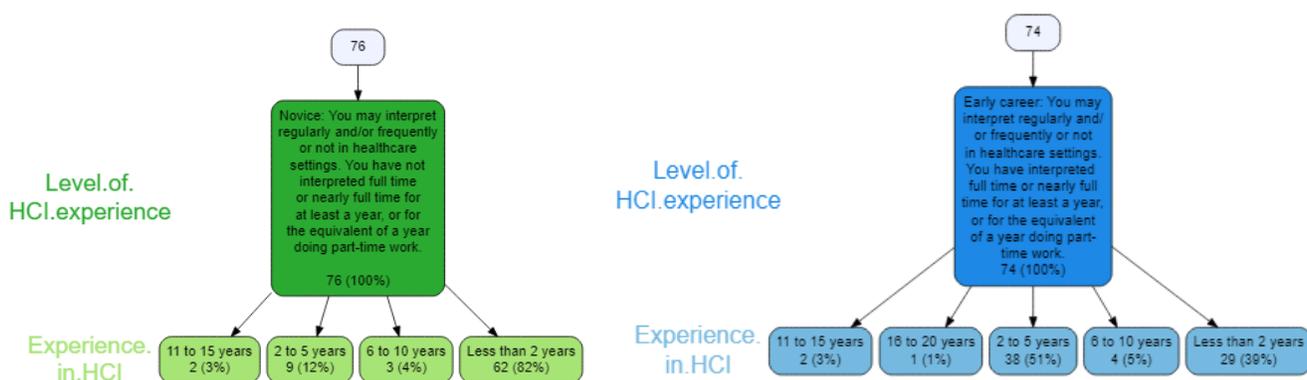
Years of HCI Experience	Count	Percent
Less than 2 years	93	38%
2 to 5 years	75	30%
6 to 10 years	45	18%
11 to 15 years	23	9%
16 to 20 years	7	3%
21 years or more	4	2%
Total	247	100%

61% of participants identified themselves as “novice” or “early career,” thus, representing the target audience of the certification, *i.e.*, entry-level interpreters (table 6).

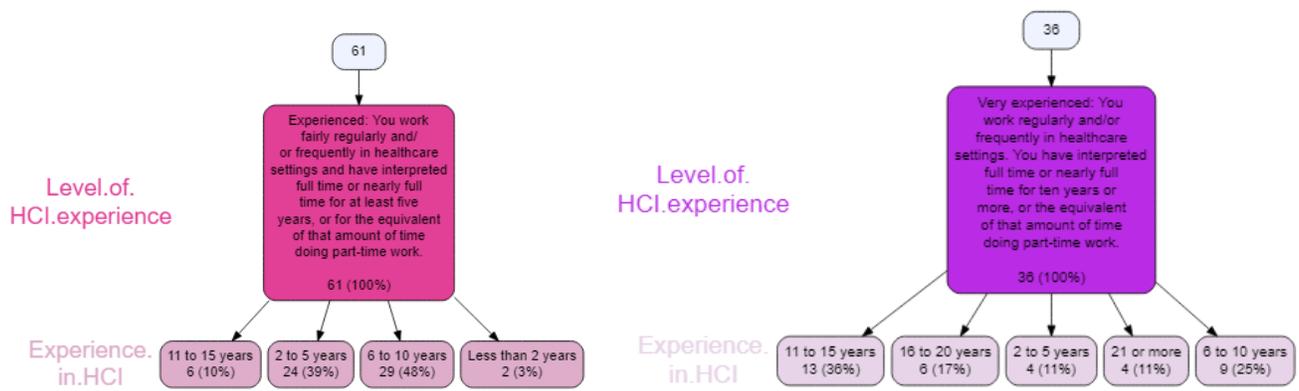
Table 7. How much healthcare interpreting experience do you have?

Level of HCI Experience (Subjective) ¹¹	Count	Percent
Novice	76	31%
Early career	74	30%
Experienced	61	25%
Very experienced	36	15%
Total	247	100%

Of interest is distribution of years of interpreting within the subjectively selected levels of respondents’ experience. Refer to the following charts.



¹¹ See the definitions of the levels in Appendix E.



Frequency of interpreting in healthcare settings is described in Table 8. 63% of respondents interpret fewer than 21 hours per week. This factor is important to reflect the potential audience of the EtoE exam correctly. Most representatives of the potential audience will have a similar frequency of interpreting since it is comprised of interpreters of languages of lower incidence who rarely have full-time employment as interpreters.

Table 8. How many hours do you interpret per week in healthcare settings specifically?

Hours of interpreting in HC settings, per week	Count	Percent
Less than 2 hours	68	28%
3-20 hours	86	35%
21 - 40 hours	81	33%
41 hours and over	12	4%
Total	247	100%

CCHI’s **general education** requirement is a minimum of high school-level diploma. 80% of participants exceeded this requirement, with 63% obtaining a four-year degree or higher. Table 9 summarizes responses to this question.

Table 9. Which of the following most closely describes the highest level of formal education (from any country) that you have completed?

Education Level	Count	Percent
High school diploma/GED or equivalent	49	20%
Associate degree (any major/specialization)	43	17%
Bachelor’s degree (any major/specialization)	93	38%
Master’s degree (any major/specialization)	52	21%
Doctoral degree (any major/specialization)	9	4%
Post-doctoral degree (any major/specialization)	1	0.04%
Total	247	100%

The monolingual modality of the EtoE exam requires candidates to perform language-related tasks that are atypical for a healthcare interpreter. Yet, these tasks require skills and abilities that are components of interpreting and are often part of **university-level courses in interpreting, translation, or linguistics** (Table 10). 48% of participants stated that they had completed a university-level training in interpreting. 14% had completed university-level courses in two of these three language specialties.

Table 10. Did you have any university-level training in interpreting (regardless of the setting type) (e.g., an interpreting course at a community college, college, or university)/ translation/ linguistics?

Type of University-Level Language Specialty	Count	Percent (of 247)
Interpreting	119	48%
Translation	57	23%
Linguistics	71	29%
Two of the above specialties	37	14%
All three	30	12%

Table 11 describes the distribution of respondents according to the nature of the **acquisition of the Language Other Than English (LOTE)**. 62% of participants were native speakers of LOTE, which is consistent with the profession’s distribution. 37 participants (23%) were native speakers of English who acquired the LOTE via either formal or informal learning. 24 participants (15%) identified themselves as heritage speakers and consisted only of the Spanish interpreters. Heritage speaker is defined as a person who speaks the non-English language most exclusively at home with family and friends, while growing up and living in an English-speaking country.

Table 11. How did you acquire your non-English interpreting language?

LOTE Acquisition	Count	Percent
Native speaker (of LOTE)	150	61%
Second language learner: formal learning (college, etc.)	52	21%
Second language learner: informal learning (self-taught)*	8	3%
Heritage speaker*	37	15%
Total	247	100%

* These groups contained only Spanish interpreters.

CCHI’s eligibility criterion of **healthcare interpreter training** requires a minimum of 40 hours of instruction in healthcare interpreting (up to 5 hours may be in another language field or in a healthcare specialty). 68% of participants reported having completed more than that amount (Table 12). It is important to remember that none of the participants in the study had yet been certified by CCHI.

Table 12. How much formal (academic or non-academic) training do you have in healthcare interpreting (including continuing education and conferences)?

HCI Training	Count	Percent
Less than 40 instructional hours	11	4%
40 instructional hours	69	28%
41-65 instructional hours	73	30%
66-100 instructional hours	42	17%
over 100 instructional hours	50	20%
Associate degree in healthcare interpreting	1	0.4%
Bachelor’s degree in healthcare interpreting	0	0
Master’s degree in healthcare interpreting	1	0.4%
Total	247	100%

Since our profession is seeing a rapid growth of training opportunities specific to healthcare interpreting, it is important to observe which **formats of professional education** interpreters seeking certification choose. These formats are described in Table 13. The most popular choice – 38% – was a non-academic program of 40+ hours duration, with in-person instruction. However, academic programs seem to be gaining popularity, with 25% of participants reporting this method. The share of online modality of interpreter education is represented at 18%, exceeding the on-the-job training (15%).

Table 13. How did you receive the majority of training in healthcare interpreting?

Format of Obtaining HCI Education	Count	Percent
an academic program in medical interpreting of 45 hours (3 credits in U.S.) in duration (any country)	22	9%
an academic program in medical interpreting of <u>more than 45 hours</u> in duration (any country)	40	16%
a non-college program of 40-100 hours with in-person instruction (e.g., <i>Bridging the Gap</i> , <i>The Community Interpreter</i> , <i>The Art of Interpretation</i> , etc.)	93	38%
a non-college program of 40-100 hours with online instruction	32	13%
a combination of in-person workshops and conferences	11	4%
a combination of online courses and webinars	13	5%
on-the-job training	36	15%
Total	247	100%

Table 14 describes the **amount of specific skill-based training** respondents received prior to this study. While 63% of respondents have studied consecutive interpreting for more than 25 hours, only 34% dedicated more than 25 hours to simultaneous interpreting training, and even fewer – 29% allocated that amount of time to sight translation training.

Table 14. Estimate how many hours of training (with an instructor) you have had in consecutive/simultaneous/sight translation interpreting (regardless of the setting)

	Consecutive		Simultaneous		Sight Translation	
	Count	Percent	Count	Percent	Count	Percent
0-2 hours	18	7%	64	26%	76	31%
3-6 hours	20	8%	41	17%	43	17%
7-12 hours	35	14%	36	14%	37	15%
13-24 hours	20	8%	22	9%	20	8%
25-45 hours	59	24%	39	16%	34	14%
more than 45 hours	95	39%	45	18%	37	15%
Total	247	100%	247	100%	247	100%

Table 15 describes levels of **deliberate exposure of respondents to English and LOTE**. Overall, respondents have less deliberate exposure to LOTE in both reading and watching/listening modalities than to English. This is not surprising since respondents reside and work in the U.S.

Table 15. How much time do you spend reading/watching or listening in English and in LOTE (any type of content)?

Language	Reading (count)		Watching or Listening (count)	
	ENG	LOTE	ENG	LOTE
0 time	1	2	4	14
less than 30 minutes a week	3	19	3	18
1 hour a week	15	38	9	37
2-7 hours a week	94	122	100	126
8-14 hours a week	65	40	63	31
more than 15 hours a week	69	26	68	21

The observations gleaned from the *EtoE Study Questionnaire* display a significant diversity among the study participants. This diversity makes it possible to infer that this group is reasonably representative of the healthcare interpreting profession in general, and the results of the study could be applicable to interpreters of other languages.

Part II. EtoE Study Follow-up Validation Report by Prometric LLC



EXPERTISE. INTEGRITY. CHOICE.
Measured by your success.



Prepared By:

Oksana Naumenko, Psychometrician
Charis Walikonis, Assessment Design Specialist
Assessment Services
Prometric LLC

This report is prepared on behalf of CCHI by Prometric LLC. This document is copyrighted and is intended for the sole use of CCHI. Copyright 2021 by Prometric LLC

INTRODUCTION

The Certification Commission for Healthcare Interpreters (CCHI) endeavored to commission test development and psychometric services to conduct a study of English-to-English (EtoE) cognitive interpreting skills. The CCHI currently provides language-specific oral performance exams to prospective healthcare interpreters in Arabic, Mandarin, and Spanish. In order to certify candidates in lower incidence languages, the CCHI would like to determine whether an EtoE cognitive interpreting skills exam would be a substitute for the current CHI™ language-specific oral performance exam. If the EtoE exam were found to be highly predictive, then it would open up the possibility for the CCHI to offer EtoE interpreter exams to candidates with proficiency in lower incidence languages for which creating a separate language-specific oral performance exam is cost-prohibitive.

Prometric assisted CCHI in the development and administration of the EtoE oral performance exam. This report provides a technical overview of the analyses and the results to bolster the validity of the EtoE score use. More specifically, the overarching goal of the preliminary study was to gauge the promise of the EtoE as a substitute for a bilingual assessment and to also gauge whether the EtoE is comparable to the current CHI™ language-specific oral exam. These results will help the CCHI to determine a plan of action for the future regarding how to test candidates for proficiency in their ability to interpret from English into a lower incidence language. Prometric does not suggest that the EtoE exam as a language-independent construct is comprehensive. While the EtoE exam cannot replace the CHI™ dual-language performance exam, it may have value if used as a component of an assessment portfolio for interpreters of lower incidence languages.

ETOE MEASURE DEVELOPMENT BACKGROUND

In 2017-2018, CCHI has conducted several stakeholder engagement activities to ascertain the necessity and feasibility of the English-to-English (EtoE) oral performance exam for healthcare interpreters. The concept received wide support, and the whitepaper about the EtoE project was published in May 2018 (see http://cchicertification.org/uploads/CCHI_EtoE_Interpreter_Performance_Assessment.pdf).

In summer of 2018, CCHI formed the EtoE National Task Force Panel of 22 experts to further plan the project. The panelists and CCHI staff met remotely and consulted with Dr. James P. Henderson of Credentialing Examination Consulting, LLC (formerly Scantron) about the possible design of the EtoE test and of the corresponding study. Based on these meetings, CCHI developed the necessary items and rubrics for the study.

Prometric Test Development staff completed the following two prerequisite test development steps before creating a test form for the EtoE study:

1. Review recordings of the item development and review sessions facilitated by CCHI; and
2. Participate in 2–3 teleconference sessions on scoring and test blueprint creation.

This first prerequisite step enabled Prometric content developers to gain a deep understanding of the types of items that appear on the EtoE, the thought process behind creating these items, and the skills and abilities these items are intended to measure. The item bank includes several types of test items that are designed to measure candidates' cognitive interpreting skills, including (1) audio input-audio output, (2) text input-audio output, and (3) traditional multiple choice.

During the second prerequisite step, Prometric staff was more actively involved in the test development process for the EtoE. With the exception of the traditional multiple-choice items, most of the items on the exam were scored by human raters. Prometric worked with CCHI subject matter experts (SMEs) via teleconference to

develop new scoring rubrics/scoring guides for the different types of items that require human raters. Prometric also provided CCHI raters training preparation on the new item types and scoring rubrics/scoring guides.

In addition, Prometric staff participated in a teleconference session to develop the test blueprint. A well-defined and detailed test blueprint allows test developers to use the content and psychometric requirements to guide the selection and placement of items on the test form. Once the test blueprint is established, Prometric test developers are able to select from the newly developed test items to create new EtoE test forms, considering that enough evidence exists to support the use of the EtoE exam to judge candidates' ability to interpret in a practical setting.

Examination Form

The EtoE exam consists of 33 tasks subdivided into nine task types: 1) Reading Comprehension; 2) Shadowing; 3) Finish the Sentence; 4) Restate the Message; 5) Listening Comprehension and Speech Production; 6) Memory Capacity; 7) Equivalence of Medical Terminology; 8) Synonyms (Multiple Choice); and 9) Fill-in-the-blank.

Table 1 shows the item type order, the input-output mode, the relative weight to be applied to each item type (not executed for the simple preliminary study), the number of items per type, and the number of input and output audio files the type requires. The three modes of input/output are text-to-audio, audio-to-audio, and multiple choice. Due to the exploratory nature of this study, item type weights were not applied to candidate scores prior to analysis execution.

TABLE 1. ETOE SCALE BLUEPRINT

Order	Type	Mode*	Weight	Items	Audio Input Files	Audio Output Files
1	Reading Comprehension	T-to-A	10%	1	0	1
2	Shadowing	A-to-A	12.5%	1	1	1
3	Finish the Sentence	A-to-A	8%	5	5	5
4	Restate the Message	A-to-A	12.5%	8	8	8
5	Listening Comprehension	A-to-A	15%	1	1	1
6	Memory	A-to-A	15%	8	8	8
7	Equivalence	T-to-A	10%	3	0	3
8	Medical Concepts	MCSR	9%	3	0	0
9	Fill-in-the-Blank	T-to-A	8%	3	3	3
TOTAL			100%	33	26	30

* Mode/Item Type per input: T-to-A = Text-to-Audio; A-to-A = Audio-to-Audio; MCSR = Multiple Choice Single Response

Rubric and Rater Handbook

A detailed scoring rubric, along with the rater handbook, were developed to guide raters in their judgments of participants' interpreting quality (see *Appendix G* for an example of the EtoE rubric and Key Elements for item type Reading Comprehension and Speech Production). The rater handbook described each of the nine item types, discussed the knowledge, skills and abilities to be assessed by each type, and provided guidance on making the scoring decision. In addition, for each EtoE item, notes were provided with examples of correct, partially correct, and incorrect responses.

Each of the item types on the examination required a specific rubric, each of which is comprised of three to five of the following scales, or criteria: 1) Quality of Speech; 2) Task Completion; 3) Accuracy and Cohesion/Coherence; 4) Lexical Content; and 5) Grammar. For example, item type “Shadowing” requires the simultaneous repetition of an audio recording. To evaluate a candidate’s ability to complete this task, scales Task Completion, Quality of Speech, and Accuracy and Cohesion/Coherence are considered. On the other hand, “Reading Comprehension and Speech Production” is evaluated using all five scales, as the task’s purpose is to assess the quality of English speech (Grammar) as well maintaining meaning (Lexical Content). Table 2 defines the nature of each of the scales.

TABLE 2. ETOE SCALE DESCRIPTORS

Scale	Scale Description
1. Quality of Speech	Focuses on the physical characteristics of the speech produced. Physical characteristics include false starts, self-corrections, repetitions, pronunciation, articulation, volume control, pace, and intonation.
2. Task Completion	Refers to completion of the task in a relevant manner, from the point of view of following the instructions.
3. Accuracy and Cohesion/Coherence	Focuses on relevance (logical response) and correctness (medical concepts) of the information. Cohesion focuses on the degree to which sentences (or different parts of one sentence) are connected so that the flow of ideas is easy to follow. Coherence is the quality of being understandable. Here, errors include omissions of information, additions of information, and disorganized flow of ideas.
4. Lexical Content	Refers to accurate rendition of “units of information” and maintaining register (when possible/applicable). Units of information can be individual words, groups of words, or phrases that communicate a single concept. Register is a variety of language used for a particular purpose or in a particular social setting. Here, errors include inaccurate re-statements of a unit of information, incorrect usage of words/phrases, and unjustified changes of register.
5. Grammar	Includes the set of rules that govern how sentences, phrases, and words are put together in a given language (keeping in mind generally accepted speech patterns). Examples of errors in grammar include verb tense, number, gender, word order, and incomplete thoughts.

Each of the five scales identifies levels of candidate performance as 0 – Unqualified; 1 – Limited; 2 – Competent; and 3 – Accomplished, and provides detailed descriptions as to the meaning of the label for each specific scale in the form of behavioral anchors. Key elements were also provided to the raters to assist in focusing on specific features within each part of the scale.

The EtoE Scoring rubrics provided specific guidelines as to the behavioral features necessary to be present in a candidate’s response for providing each of the score levels above. For example, what features of a candidate’s audio output should be considered when evaluating Grammar in a Reading Comprehension and Speech Production task? What were the characteristics of the candidate’s product that resulted in a score of “2” on the Accuracy and Cohesion scale for the Reading Comprehension task? The key elements comprise the essential features of a scale, and raters were encouraged to consider them during the rating process.

EtoE Administration

The CCHI recruited over 300 volunteer candidates to participate in this study. However, due to the global Coronavirus Disease Pandemic, also known as COVID-19, many of the recruited participants were unable to have been administered the EtoE measure. A total of 249 candidates participated in the EtoE study, some of whom were excluded due to incomplete exams originating from technical issues during testing. As originally intended, the participants in the study also had taken one of the three CCHI oral language-specific examinations – CHI™-Arabic, CHI™-Mandarin, and CHI™-Spanish. Both exams were administered on the same day at Prometric’s secure test centers located throughout North America.

EtoE Scoring

Except for the multiple-choice items, items on the EtoE exam were scored by human raters. Raters were trained and calibrated on the scoring rubrics and guides using actual EtoE measure. Prometric provided the raters with online access to audio files of the English prompts and to the candidate’s audio responses. Raters used a specialized rating system calibrated specifically for the EtoE exam to complete their scoring activities. Each participant’s response was scored by at least two raters, with a chief rater providing the final rating if two raters were more than one scale point apart on any one of the item averages. Raters were allowed to give half-points. If two raters’ scores on a particular item were separated by more than one point, then a chief rater scored that item and was weighted twice in the scoring process.

The scoring procedure for each participant was as follows. 1) At constructed response (CR) item level, scale scores (where scales vary by item, for example, all five scales are used for Reading Comprehension, but only Quality of Speech, Task Completion, and Accuracy and Cohesion/Coherence, criteria were used for Shadowing) were averaged for each rater; 2) the simple absolute difference between raters was computed and evaluated against 1; 3) if the average difference was larger than one, a third rater was assigned to judge the item. 4) To compute the final item score for each participant, all rater means were averaged, where the weights for raters were: 1 for Rater 1, 1 for Rater 2, and 2 for Rater 3 (i.e., chief rater was treated as two raters). Candidates total scores were obtained by aggregating across the 33 items with the maximum score being 99 points.

In this report, the original EtoE scores were weighted as shown in the EtoE Scale Blueprint (Table 1). Specifically, for each candidate and item, ratings were first averaged across the rater pair/triplet. Then, an item-type specific weight was applied to obtain the final weighted item score for each candidate. Finally, the item scores were summed across the measure to obtain the final weighted EtoE scores.

RESEARCH QUESTIONS

Three main research questions were identified by CCHI, evidence towards which will help to establish the validity of the EtoE measure score interpretations and uses:

- 1) What is the relationship between the Arabic, Mandarin, and Spanish CHI™ exam scores and the EtoE passing scores?
 - a. What is the relationship between the CHI™ and EtoE scores across all CHI™ exams?
 - b. What is the relationship between the CHI™ and EtoE scores for groups that took the Arabic, Mandarin, and Spanish CHI™ exam?
 - c. Do candidates who pass the CHI™ exam also pass the EtoE exam?

- 2) Does the EtoE exam measure cognitive interpreting skills? Is the structure of the cognitive interpreting construct the same across groups that take the Arabic, Mandarin, and Spanish CHI™ exam?
- 3) Does EtoE item type predict candidate performance passing status on the CHI™ exam?
 - a. Does EtoE item type predict candidate passing status on the CHI™ exam?
 - b. Does EtoE item type predict candidate score on the CHI™ exam?

The following table outlines the methodology proposed to be used to address the above research questions:

TABLE 3. ANALYSES TO ANSWER RESEARCH QUESTIONS

Research Question	Methodology	Interpretation of Supporting Evidence
1a	Exploratory Pearson Correlations between CHI™ Scores and EtoE Scores Overall	Higher positive correlation index indicates a stronger relationship between the CHI™ and EtoE.
1b	Exploratory Pearson Correlations between Each Language-specific Set of CHI™ Scores and EtoE Scores	Higher positive correlation index indicates a stronger relationship between the CHI™ and EtoE for the particular language group.
1c	Logistic Regression of CHI™ Exam Passing Status on the EtoE Passing Status/CHI™-square Test	Higher log-odds indicate how much more/less likely passing CHI™ candidates are to pass the EtoE exam.
2*	Overall/Multigroup Structural Equation Model of the Interpreting Cognitive Skills Construct	Best fitting factor model functioning similarly across groups indicates that the construct is invariant to group membership and thus does not dispute that the EtoE exam measures a language-independent construct.
3a	Logistic Regression of the CHI™ Passing Status on the EtoE Item Type Score	Larger logistic regression coefficients indicate higher predictive value of passing status.
3b	Simple Linear Regression of the CHI™ Scores on the Item Type Score	Larger regression coefficients indicate higher predictive value of passing status.

Note. *Due to the small samples of the CHI™-Arabic and Mandarin tests, the Multigroup SEM analysis was not conducted.

RESULTS

Prior to answering the above research questions, the quality of the newly developed EtoE items was evaluated to establish that any further analyses are based on reliable data. For this, a classical item analysis was conducted to shed light on the average item quality information, and item-specific item quality information. Item quality was primarily based on each item's correlation with the total scores on the EtoE exam (i.e., item discrimination). Psychometric exam quality was also judged using the Classical Test Theory Cronbach's alpha/KR-20 statistics using the accepted quality ranges. It should be noted that causal interpretations should not be drawn from any of the evidence provided by the described analytical methods. For further detail on the reported indices please see *Appendix H*.

DATA INTEGRITY

Incomplete data were removed from the analyses after a thorough review of rater comments by CCHI. Specifically, CCHI provided lists of IDs of participants for each item on the exam that were flagged for removal due to incorrect interpretation of the rubric, audio quality, or candidate behavior that was not consistent with the guidelines of the rubrics. Subsequently, out of 249 available cases, item-level data of 72 candidates (for one or more item) were removed. In subsequent analyses, candidates were removed listwise if the total score was important to the calculation of any statistical indices. For research questions (RQs) 1 through 3, it was important to maintain the sample of candidates that had complete ratings on all items and all rubric scales. Analyses within these RQ groups depended on the accurate calculation of the total score (raw or weighted), so candidates were removed if they had any item ratings missing. Due to this, the sample size for RQ1 and RQ3 was 177. For RQ2, it was important to maximize information at the item level, thus the sample size for analyses under that category used the original 249 candidates.

In this report, the original EtoE scores were weighted as shown in the EtoE Scale Blueprint (Table 1). Specifically, for each candidate and item, ratings were first averaged across the rater pair/triplet. Then, an item-type specific weight was applied to obtain the final weighted item score for each candidate. Finally, the item scores were summed across the measure to obtain the final weighted EtoE scores.

ITEM ANALYSIS

Rater averages on the EtoE exam were obtained for each rated item across 177 participants in the EtoE study. These values, along with multiple choice item scores, were then evaluated in a classical test theory item analysis for relative easiness/difficulty and discrimination of items (Table 4). The item-total correlation (ITC) values are indices of discrimination and were evaluated against $ITC < 0.20$ criterion. None of the items fell below the flagging criteria for discrimination, signaling that most EtoE items are useful in predicting a person's total score. An item was flagged as "hard" if the scale mean fell below 0.30 points for multiple choice items, and below 1 point for rated items, and flagged as easy if the scale mean was above 0.90 for multiple choice items and above 2.50 for rated items. Nine items were flagged as relatively easy, which included all the "Finish the Sentence" items, all "Fill-in-the-Blank" items, the Shadowing item, and the shortest of the Memory items. The overall reliability for the 33-item scale was measured by Cronbach's Alpha, which was at 0.92, indicating a very high index of internal consistency. The average score on the EtoE scale was 66.78 out of 99 points (67.45% of total possible), with a standard deviation of 9.83 ($N = 177$). When weighted by the item type, the average score is 7.50 out of 10.815 possible weighted points (69.34% of total possible), with a standard deviation of 1.20. Weighting the final EtoE score has the effect of adjusting the raw percent correct upwards due to higher representation of certain item types that proved to be relatively easy, on average (e.g., 8 out of 33 items were Memory, which have a relatively high weight of 15% on the measure).

TABLE 4. AVERAGE DIFFICULTY AND DISCRIMINATION OF ETOE ITEMS

Item	Item Mean	Item-Total Correlation	Alpha if deleted	Hard	Easy	Low ITC	Type
1*	0.65	0.26	0.92				Medical Concepts
2*	0.61	0.26	0.92				Medical Concepts
3*	0.81	0.40	0.92				Medical Concepts
133	2.50	0.48	0.92		X		Shadowing
134	1.71	0.56	0.91				Restate the Meaning
135	1.79	0.57	0.91				Restate the Meaning

Item	Item Mean	Item-Total Correlation	Alpha if deleted	Hard	Easy	Low ITC	Type
136	1.82	0.58	0.91				Restate the Meaning
137	1.75	0.47	0.92				Restate the Meaning
138	1.73	0.55	0.91				Restate the Meaning
139	1.90	0.46	0.92				Restate the Meaning
140	1.73	0.58	0.91				Restate the Meaning
141	1.74	0.56	0.91				Restate the Meaning
142	2.12	0.38	0.92		X		Fill-in-the-Blank
143	2.70	0.32	0.92		X		Fill-in-the-Blank
144	2.60	0.20	0.92		X		Fill-in-the-Blank
145	2.77	0.41	0.92		X		Finish the Sentence
146	2.73	0.33	0.92		X		Finish the Sentence
147	2.63	0.40	0.92		X		Finish the Sentence
148	2.88	0.37	0.92		X		Finish the Sentence
149	2.61	0.38	0.92		X		Finish the Sentence
150	2.20	0.64	0.91				Equivalence
151	2.06	0.65	0.91				Equivalence
152	2.18	0.62	0.91				Equivalence
153	2.35	0.41	0.92				Reading Comprehension
154	2.27	0.46	0.92				Listening Comprehension
158	2.51	0.52	0.91		X		Memory
159	2.22	0.55	0.91				Memory
160	1.94	0.64	0.91				Memory
161	1.84	0.57	0.91				Memory
162	2.40	0.56	0.91				Memory
163	1.40	0.60	0.91				Memory
164	1.62	0.62	0.91				Memory
165	1.64	0.55	0.91				Memory

Note. *Multiple choice items, thus item means are equivalent to an item p-value, or proportion of participants that correctly answered the item.

ETOE MEASURE DESCRIPTIVE STATISTICS

The overall calculated score was slightly negatively skewed, meaning that the large bulk of study participants did well on the measure. The weighted median score was 7.51 (69% of the maximum scale score), once all cases with any missing data were removed (N = 177). Table 5 contains descriptive statistics associated with the total EtoE score, whereas Table 6 contains correlations between items on the entire scale, with item type indicated.

TABLE 5. OVERALL ETOE SCORE DISTRIBUTION

Raw		Weighted	
Index	Value	Index	Value
N	177	N	177
Mean	66.40	Mean	7.50
SD	9.83	SD	1.20
Median	66.78	Median	7.51
Min	30.03	Min	3.43
Max	84.77	Max	9.82
Range	54.74	Range	6.39
Skew	-0.49	Skew	-0.33
Kurtosis	0.55	Kurtosis	0.09
SE	0.74	SE	0.09

In general, the expected patterns of the correlation matrix should be such that items within the same item-type grouping are related more to each other than items outside of the groupings. For example, the expectation is that items within the Memory Capacity item type group should be more highly correlated to each other (blue) than to items within the Reading Comprehension item-type group if they are truly distinct components of measuring interpreting competency. However, all items on the scale, regardless of item-type group/competency, should be expected to be positively correlated if they are measuring the same overall trait of Interpreting Ability.

Because only some item type groups are similar to competencies in that they incorporate a set of skills, abilities and knowledge (versus only skills, or even innate aptitude, such as the measure of Memory Capacity), some relationship patterns may be interpreted differently. For example, with only three Medical Concepts items, it is difficult to establish patterns, because the three multiple choice items could belong to vastly different medical concept topics.

It is important to note that these relationships can be more easily observed with a sufficient number of items within an item-type group. Specifically, with only one item in the Shadowing, Reading Comprehension and Listening Comprehension groups, it would be difficult to establish a reliable understanding of whether the item itself is representative of the competency and whether it is the quality of the item, or the actual concept that it represents is affecting its relationships with items within other item-type groups. It should be noted, though, that the length of each of these single items is significantly longer than items of the other types, which precluded the inclusion of multiples. Additionally, the Shadowing item is intended to represent a unique skill of simultaneity of the interpreter's cognitive resource management, which corresponds to a bilingual simultaneous interpreting item on the CHI™ dual-language performance exam.

The correlation matrix should also be interpreted hand in hand with item quality information. Special attention should be paid to items with lower discrimination values. For example, item 144 in the Fill-in the-Blank group had the lowest item discrimination value and this item also shows negligible, low, or even negative correlations with other items on the measure. This item may be confusing or may have a possible second key. Similarly, Medical Concepts items 1 and 2, which had some of the lower discrimination values, appear to be the least strongly correlated to most items on the scale (red).

Conceptually, many of the patterns appear logical. For example, Memory Capacity, an item group that appears to be appropriately homogenous, is related moderately to Listening Comprehension, Equivalence, and Restate the Meaning item groups, all of which conceptually require higher working memory capacity for better performance. Overall, the relationships based on the new sample are similar to those in the preliminary study (see https://cchicertification.org/uploads/CCHI-ETOE_Study_Preliminary_Report.pdf).

TABLE 6. INTER-ITEM PEARSON PRODUCT MOMENT CORRELATIONS BY ITEM TYPE

Item	1	2	3	133	134	135	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150	151	152	153	154	158	159	160	161	162	163	164			
2	0.12	MC																																	
3	0.15	0.17		SW																															
133	0.00	0.23	0.16																																
134	0.19	0.12	0.26	0.27																															
135	0.22	0.16	0.13	0.37	0.43																														
136	0.24	0.19	0.11	0.35	0.28	0.41																													
137	0.16	0.36	0.10	0.24	0.34	0.31	0.40																												
138	0.15	0.17	0.20	0.31	0.40	0.25	0.41	0.41																											
139	0.21	0.10	0.21	0.24	0.33	0.14	0.28	0.12	0.32																										
140	0.25	0.23	0.29	0.24	0.54	0.41	0.32	0.33	0.33	0.29																									
141	0.22	0.21	0.18	0.35	0.46	0.54	0.41	0.33	0.28	0.22	0.50																								
142	0.07	0.10	0.35	0.35	0.28	0.26	0.23	0.21	0.25	0.21	0.18	0.22																							
143	0.00	0.01	0.17	0.06	0.21	0.17	0.17	-0.06	0.18	0.23	0.25	0.20	0.09																						
144	0.10	0.09	0.11	0.04	-0.05	0.06	0.11	0.20	0.18	0.14	0.09	0.06	0.09	0.08																					
145	-0.07	0.04	0.16	0.28	0.21	0.31	0.25	0.19	0.16	0.12	0.30	0.26	0.28	0.15	0.14																				
146	-0.04	0.11	0.23	0.19	0.17	0.19	0.18	0.20	0.12	0.14	0.22	0.16	0.17	0.12	0.11	0.39																			
147	0.07	0.11	0.21	0.28	0.24	0.17	0.08	0.10	0.15	0.22	0.29	0.31	0.15	0.20	0.06	0.38	0.37																		
148	-0.04	0.08	0.27	0.26	0.26	0.16	0.11	0.22	0.17	0.25	0.23	0.20	0.17	0.12	0.00	0.34	0.38	0.41																	
149	0.03	-0.06	0.22	0.16	0.22	0.21	0.23	0.12	0.23	0.22	0.33	0.28	0.18	0.16	0.02	0.34	0.08	0.29	0.27																
150	0.23	0.10	0.30	0.31	0.32	0.37	0.34	0.31	0.36	0.40	0.33	0.35	0.27	0.20	0.31	0.36	0.26	0.30	0.27	0.23															
151	0.26	0.16	0.27	0.28	0.44	0.38	0.37	0.32	0.40	0.35	0.38	0.29	0.22	0.20	0.20	0.26	0.12	0.25	0.24	0.34	0.61														
152	0.27	0.19	0.26	0.25	0.46	0.38	0.38	0.32	0.45	0.36	0.37	0.32	0.26	0.24	0.16	0.27	0.15	0.26	0.22	0.29	0.58	0.66													
153	0.06	0.13	0.13	0.31	0.31	0.26	0.26	0.20	0.22	0.27	0.41	0.23	0.15	0.17	0.02	0.29	0.14	0.25	0.23	0.31	0.23	0.33	0.27												
154	0.13	0.17	0.13	0.31	0.27	0.40	0.41	0.22	0.23	0.15	0.16	0.29	0.30	0.17	0.10	0.16	0.15	0.12	0.09	0.17	0.29	0.30	0.22	0.15											
158	0.25	0.11	0.23	0.20	0.19	0.32	0.38	0.25	0.20	0.21	0.30	0.22	0.22	0.31	0.10	0.29	0.26	0.25	0.21	0.25	0.42	0.32	0.34	0.17	0.37										
159	0.12	0.10	0.28	0.27	0.29	0.33	0.37	0.23	0.41	0.27	0.28	0.37	0.24	0.17	0.16	0.17	0.12	0.25	0.17	0.25	0.37	0.38	0.39	0.11	0.20	0.30									
160	0.19	0.19	0.32	0.35	0.39	0.38	0.34	0.36	0.43	0.28	0.36	0.26	0.19	0.26	0.17	0.26	0.18	0.19	0.24	0.19	0.46	0.45	0.45	0.20	0.26	0.39	0.46								
161	0.13	0.14	0.24	0.22	0.31	0.25	0.40	0.22	0.34	0.40	0.40	0.29	0.13	0.27	0.11	0.10	0.17	0.24	0.18	0.25	0.38	0.44	0.32	0.30	0.21	0.31	0.46	0.51							
162	0.16	0.17	0.13	0.28	0.41	0.36	0.33	0.28	0.34	0.29	0.35	0.44	0.15	0.21	0.07	0.27	0.25	0.35	0.26	0.31	0.34	0.35	0.30	0.31	0.34	0.33	0.28	0.31	0.37						
163	0.23	0.16	0.31	0.33	0.35	0.47	0.45	0.28	0.34	0.21	0.40	0.34	0.22	0.23	0.04	0.11	0.13	0.11	0.19	0.19	0.42	0.43	0.36	0.21	0.38	0.34	0.44	0.47	0.46	0.38					
164	0.20	0.19	0.20	0.30	0.39	0.40	0.48	0.32	0.42	0.25	0.34	0.35	0.19	0.23	0.12	0.16	0.20	0.15	0.19	0.18	0.37	0.43	0.29	0.27	0.46	0.38	0.35	0.44	0.39	0.47	0.54				
165	0.09	0.18	0.24	0.30	0.25	0.25	0.37	0.33	0.33	0.23	0.24	0.30	0.16	0.25	0.22	0.21	0.20	0.22	0.15	0.15	0.30	0.33	0.34	0.09	0.31	0.26	0.44	0.54	0.39	0.32	0.43	0.50			

Note. SW = Shadowing. RC = Reading Comprehension. LC = Listening Comprehension.

Table 7 provides descriptive statistics for the EtoE measure by Language Acquisition group (correlations can be found in the *Results* section).

TABLE 7. DESCRIPTIVE STATISTICS BY LANGUAGE ACQUISITION GROUP

Group	N	EtoE (Weighted) Score				CHI (Scaled) Score			
		Mean	SD	Min	Max	Mean	SD	Min	Max
Heritage speaker	23	7.62	1.22	3.43	9.18	458.39	66.86	300	559
Native speaker of LOTE*	111	7.18	1.09	3.44	9.82	472.98	55.14	330	590
Non-native speaker of LOTE*	43	8.27	1.12	5.51	9.77	474.79	55.90	338	559

*LOTE = Language Other Than English

RATER AGREEMENT

Since each item score for each rater consisted of the average of multiple scales, causing a lessened likelihood of any item score to be exactly the same between two raters, rater agreement was calculated for multiple thresholds, as percent of item scores that were similar between raters. Chief ratings were excluded from this analysis. After removing multiple choice items, and items with partial ratings, and ratings that were flagged as problematic by CCHI, rater agreement was calculated as the proportion of the sum of instances the difference between rater scores was less than X out of all item score comparisons (X being a point difference of 0.25, 0.50, 0.75, and 1). Table 8 expresses percent agreement between two raters across all possible ratings overall and for each non-multiple-choice item. Note that more complex (e.g., generalizability, multi-faceted Rasch modeling) can be applied with larger samples to understand inter-rater reliability and precision of measurement. Overall, rater agreement was very good, with 87% of the ratings between two raters being within one point of each other.

TABLE 8. RATER AGREEMENT WITHIN VARIOUS THRESHOLDS

Threshold	.25 points	.50 points	.75 points	1 point	Absolute	Pearson	Item Type	Average
Overall	43%	61%	79%	87%	24%	0.69	Mixed	
133	60%	73%	89%	94%	35%	0.79	Shadowing	94%
134	34%	57%	80%	92%	10%	0.52	Restate the Meaning	86%
135	41%	62%	82%	89%	8%	0.46	Restate the Meaning	
136	33%	53%	72%	85%	6%	0.48	Restate the Meaning	
137	33%	53%	71%	79%	8%	0.35	Restate the Meaning	
138	28%	44%	72%	84%	4%	0.47	Restate the Meaning	
139	32%	49%	77%	86%	8%	0.36	Restate the Meaning	
140	40%	56%	82%	91%	9%	0.53	Restate the Meaning	
141	28%	45%	74%	82%	6%	0.31	Restate the Meaning	
142	40%	62%	80%	86%	31%	0.67	Fill-in-the-	91%

Threshold	.25 points	.50 points	.75 points	1 point	Absolute	Pearson	Item Type	Average
							Blank	
143	68%	86%	92%	95%	56%	0.83	Fill-in-the-Blank	
144	63%	76%	84%	90%	58%	0.71	Fill-in-the-Blank	
145	64%	80%	92%	95%	53%	0.74	Finish the Sentence	94%
146	75%	88%	93%	97%	60%	0.75	Finish the Sentence	
147	49%	67%	79%	89%	37%	0.50	Finish the Sentence	
148	82%	93%	97%	98%	73%	0.80	Finish the Sentence	
149	52%	75%	85%	92%	41%	0.73	Finish the Sentence	
150	34%	56%	72%	84%	12%	0.54	Equivalence	79%
151	22%	40%	60%	72%	6%	0.37	Equivalence	
152	34%	56%	72%	82%	12%	0.49	Equivalence	
153	31%	53%	71%	85%	6%	0.22	Reading Comprehension	85%
154	32%	55%	77%	88%	8%	0.48	Listening Comprehension	88%
158	59%	70%	82%	88%	40%	0.71	Memory	83%
159	51%	66%	83%	89%	33%	0.79	Memory	
160	38%	57%	77%	82%	14%	0.62	Memory	
161	34%	58%	80%	85%	13%	0.62	Memory	
162	51%	64%	79%	85%	33%	0.66	Memory	
163	29%	44%	64%	72%	13%	0.56	Memory	

Item-specific rater agreement indices indicate that agreement between raters depends on the specific item, and do not appear to show strong trends for item type for item groups. Notably, the lowest correlation between rater scores was found for the Reading Comprehension item (Pearson = 0.22). Overall, within-item type group agreement was exhibited for the Shadowing item and the Finish the Sentence items, with the percentage of rating pairs within 1 point of each other being 94% for both. The lowest agreement for item type was found in the Equivalence item group (79%).

RUBRIC SCALE SUMMARIES

Rubric scale scores were analyzed across 177 candidates with valid ratings on all items and scales. Specifically, rubric scale scores were averaged across two or three raters for each item. Table 9 represents the average rubric score by item type. Note that rubric scales differ depending on the item type; the count listed in Table 9 represents the total number of rubric scales that exists for each item. For example, Listening Comprehension is measured by four scales (Accuracy & Cohesion/Coherence, Grammar, Lexical Content, and Quality of Speech) and only one Listening Comprehension item is included on the EtoE measure. Therefore, the count of scales used across all Listening Comprehension items is 4. Across the four scales, the average candidate item score on the Listening

Comprehension item was 2.18 with a standard deviation of 0.28. The average rating across all raters, all candidates, and all items was 2.04 (possible range 0 to 3).

TABLE 9. AVERAGE SCALE SCORES BY ITEM TYPE

Item Type	Average	SD	Count
Equivalence	2.05	0.24	15
Fill in the Blank	2.34	0.44	12
Finish the Sentence	2.62	0.17	20
Listening Comprehension	2.18	0.28	4
Memory Capacity	1.84	0.52	24
Reading Comprehension	2.37	0.12	4
Restate the Meaning	1.70	0.64	40
Shadowing	2.35	0.09	3
Grand Total	2.04	0.58	122

Table 10 represents a more detailed breakdown of scores per item type by rubric scale. Specifically, the average rubric scale score, standard deviation, and count of items is listed for each item type. For example, for item type Equivalence of Meaning (3 items of this type exist on the EtoE), the average candidate score across all rubric scales was 2.05. Across all candidates, the lowest scoring rubric scale on this item type was Lexical Content (Average = 1.68), whereas the highest scoring rubric scale was Task Completion (Average = 2.28). Across all item types, Restate the Meaning was the most challenging type of task (Average = 1.70), whereas Finish the Sentence was the easiest type of task (Average = 2.62).

TABLE 10. AVERAGE SCALE SCORES BY ITEM TYPE AND RUBRIC SCALE

Scale	Average	SD	Count
Equivalence	2.05	0.24	15
Accuracy & Cohesion/Coherence	1.90	0.05	3
Grammar	2.13	0.05	3
Lexical Content	1.68	0.10	3
Quality of Speech	2.26	0.08	3
Task Completion	2.28	0.07	3
Fill in the Blank	2.34	0.44	12
Accuracy & Cohesion/Coherence	2.03	0.53	3
Grammar	2.61	0.21	3
Lexical Content	2.10	0.49	3
Quality of Speech	2.63	0.15	3
Finish the Sentence	2.62	0.17	20
Accuracy & Cohesion/Coherence	2.47	0.18	5
Grammar	2.75	0.08	5
Lexical Content	2.54	0.15	5
Quality of Speech	2.73	0.07	5
Listening Comprehension	2.18	0.28	4
Accuracy & Cohesion/Coherence	1.84		1
Grammar	2.49		1
Lexical Content	2.07		1

Scale	Average	SD	Count
Quality of Speech	2.30		1
Memory Capacity	1.84	0.52	24
Accuracy & Cohesion/Coherence	1.49	0.48	8
Quality of Speech	2.30	0.14	8
Task Completion	1.72	0.49	8
Reading Comprehension	2.37	0.12	4
Grammar	2.51		1
Lexical Content	2.28		1
Quality of Speech	2.25		1
Task Completion	2.43		1
Restate the Meaning	1.70	0.64	40
Accuracy & Cohesion/Coherence	1.57	0.15	8
Grammar	2.32	0.08	8
Lexical Content	1.55	0.08	8
Quality of Speech	2.38	0.08	8
Task Completion	0.68	0.09	8
Shadowing	2.35	0.09	3
Accuracy & Cohesion/Coherence	2.32		1
Quality of Speech	2.28		1
Task Completion	2.44		1
Grand Total	2.04	0.58	122

Table 11 represents a more detailed breakdown of rubric scale scores by item type. Specifically, the average item score, standard deviation, and count of items is listed for each rubric scale. It is important to emphasize that differences in rubric scales exist across item types, and slightly different criteria may be used to judge candidate performance depending on the required task. For example, an Accuracy & Cohesion/Coherence rubric scale was used for seven out of nine item types on the EtoE measure, and the average rubric scale score was 1.82. Across all candidates and item types, the lowest scoring rubric scale was Task Completion (Average = 1.47), whereas the highest scoring rubric scale was Grammar (Average = 2.45). Under Task Completion, the lowest scoring item type was Restate the Meaning (Average = 0.68).

TABLE 11. AVERAGE SCALE SCORES BY RUBRIC SCALE AND ITEM TYPE

Scale	Average	SD	Count
Accuracy & Cohesion/Coherence	1.82	0.48	29
Equivalence	1.90	0.05	3
Fill in the Blank	2.03	0.53	3
Finish the Sentence	2.47	0.18	5
Listening Comprehension	1.84		1
Memory Capacity	1.49	0.48	8
Restate the Meaning	1.57	0.15	8
Shadowing	2.32		1
Grammar	2.45	0.24	21
Equivalence	2.13	0.05	3
Fill in the Blank	2.61	0.21	3

Scale	Average	SD	Count
Finish the Sentence	2.75	0.08	5
Listening Comprehension	2.49		1
Reading Comprehension	2.51		1
Restate the Meaning	2.32	0.08	8
Lexical Content	1.94	0.45	21
Equivalence	1.68	0.10	3
Fill in the Blank	2.10	0.49	3
Finish the Sentence	2.54	0.15	5
Listening Comprehension	2.07		1
Reading Comprehension	2.28		1
Restate the Meaning	1.55	0.08	8
Quality of Speech	2.42	0.20	30
Equivalence	2.26	0.08	3
Fill in the Blank	2.63	0.15	3
Finish the Sentence	2.73	0.07	5
Listening Comprehension	2.30		1
Memory Capacity	2.30	0.14	8
Reading Comprehension	2.25		1
Restate the Meaning	2.38	0.08	8
Shadowing	2.28		1
Task Completion	1.47	0.74	21
Equivalence	2.28	0.07	3
Memory Capacity	1.72	0.49	8
Reading Comprehension	2.43		1
Restate the Meaning	0.68	0.09	8
Shadowing	2.44		1
Grand Total	2.04	0.58	122

RESEARCH QUESTION RESULTS

Exploratory Pearson Correlations between CHI™ Scores and EtoE Scores and between Each Language-specific Set of CHI™ Scores and EtoE Scores

In order to foreshadow the relationship between the new EtoE exam and the CHI™ exam, the total EtoE scores were correlated with the total CHI™ scores after the simultaneous administration of both exams. The Pearson correlation coefficient has the theoretical range from -1 to 1, with more negative indices indicating an inverse relationship between two sets of scores (i.e., as scores increase on Test A, scores decrease on Test B), and more positive indices indicating a direct relationship between two sets of scores (i.e., as scores increase on Test A, scores also increase on Test B). Moderate to strong relationships suggest reciprocal predictive power between two

variables. The following guidelines will be used to label the observed correlation coefficients (Hinkle, Wiersma, Jurs, 2003¹²):

Size of Correlation	Interpretation
.90 to 1.00 (–.90 to –1.00)	Very high positive (negative) correlation
.70 to .90 (–.70 to –.90)	High positive (negative) correlation
.50 to .70 (–.50 to –.70)	Moderate positive (negative) correlation
.30 to .50 (–.30 to –.50)	Low positive (negative) correlation
.00 to .30 (.00 to –.30)	Negligible correlation

In total, out of 249 original EtoE candidates, 177 had scores on all items and also had a score on one of the three CHI™ examinations: CHIARA, CHIMAN, and CHISPA. To calculate the preliminary correlations between the CHI™ exams and the EtoE, the data sample was again reduced listwise. Table 12 shows the correlations between the EtoE score and the three CHI™ exams. As can be observed from the table, all CHI™ exam correlations range from low to moderate positive. Correlations were also provided by language acquisition category: Native Speaker of Language Other than English (LOTE), Non-Native Speaker of LOTE, and Heritage Speaker. Note that correlations were not provided when the group size was smaller than 15. As with previously reported indices, these preliminary indicators should be viewed with caution, as some are based on a very small sample (i.e., CHIARA and CHIMAN).

TABLE 12. CORRELATION BETWEEN THE CHI™ EXAMS AND THE ETOE MEASURE

	Scale	Raw EtoE	Weighted EtoE	N
Overall	CHISPA	0.45	0.44	136
	CHIARA	0.53	0.51	24
	CHIMAN	0.34	0.35	17
	All CHI	0.48	0.47	177
Native Speaker of LOTE	CHISPA	0.64	0.63	73
	CHIARA	0.51	0.48	21
	CHIMAN	0.34	0.35	17
	All CHI	0.62	0.61	111
Heritage Speaker	CHISPA	0.46	0.47	23
	CHIARA	NA	NA	0
	CHIMAN	NA	NA	0
	All CHI	0.46	0.47	38
Non-native Speaker of LOTE*	CHISPA	0.28	0.28	40
	CHIARA	NA	NA	3
	CHIMAN	NA	NA	0
	All CHI	0.29	0.29	43

*Non-native Speaker of LOTE = Native Speaker of the English language.

¹² Hinkel, D. E., Wiersma, W., & Jurs, S. G. (2003). Applied statistics for the behavioral sciences. Boston: Houghton Mifflin Company.

Logistic Regression of CHI™ Exam Passing Status on the Weighted EtoE Score

Logistic Regression is a modeling method that applies to situations where the predictor variables (e.g., EtoE passing status) are either on a continuous or categorical scale, and the criterion (e.g., CHI™ passing status) is a binary variable (i.e., only two possible outcomes). Other covariates may be introduced in the model to evaluate the relative importance of the covariate in question, the EtoE passing status. The comparative fit of models containing and excluding the EtoE passing status covariate, as well as the relative statistical importance of the EtoE passing status covariate regression coefficient were used to judge whether the EtoE is able to predict passing status on the CHI™ exams.

Because the EtoE does not have a cut point, instead of using the passing status of the EtoE as a predictor, the EtoE total score was used. Table 13 displays the crosstabulation of passing status across the three CHI™ exams and overall. Due to the small samples in the smaller exams, the overall pass/fail status was used in the logistic regression results.

TABLE 13. PASS/ FAIL STATUS BY CHI EXAM AND OVERALL

Status	CHISPA	CHIMAN	CHIARA	Total
Fail	43 (32%)	4 (24%)	14 (58%)	93 (34%)
Pass	93 (68%)	13 (76%)	10 (42%)	156 (66%)
Total	136	17	24	177

The logistic regression coefficients give the change in the log-odds of the outcome for a one unit increase in the predictor variable. Two models were fit to the data. First, an intercept model, which simply confirms that overall passing probability of 66% (see Table 13) was fit. Model 2 added the single EtoE score predictor of the passing status on the CHI™ exam (regardless of language). The intercept in Model 2 indicates the log odds of a candidate with an EtoE score of zero passing the CHI™ exam. Table 14 shows the coefficients associated with both models and the associated standard error, statistical significance levels, probability of finding an effect when there is actually none in the population, and the confidence intervals associated with the coefficient.

TABLE 14. LOGISTIC REGRESSION RESULTS

Model	Coefficients	Estimate	Standard Error	z-value	p	CI Low 2.5%	CI High 97.5%
Model 1 (Intercept Only)	Intercept	0.64	0.16	4.06	<0.001	0.34	0.96
Model 2	Intercept	-5.08	1.20	-4.23	<0.001	-7.54	-2.82
	EtoE	0.78	0.16	4.74	<0.001	0.47	1.11

The intercept model (Model 1) shows the log-odds of passing the CHI™ exam (0.64); when converted to probability, this index is 0.66, or the probability of passing the CHI™ exam, overall. In Model 2, the intercept coefficient is interpreted as the log-odds of a candidate who scored a zero on the EtoE of passing the CHI™ exam. When converted to probability, this amounts to 0.01 probability of passing. The EtoE coefficient shows that for every unit increase in the EtoE score, the log-odds of passing (versus not passing) the CHI™ exam increases by 0.78, which is a statistically significant increase ($p < .001$). We are 95% confident that the actual population effect falls between log-odds of 0.47 and 1.11. In terms of odds ratios, this can be interpreted as an increase of odds of passing a CHI™ exam by a factor of 2.18 for every point increase in the EtoE score (recall in these data this is a weighted score ranging from 3.43 to 9.42). Overall, the results suggest that the weighted EtoE is a viable predictor of the CHI™ exam passing status.

Simple Linear Regression of the CHI Scores on the Weighted Item Type Score

Linear regression is a modeling method that allows to predict continuous outcomes (e.g., CHI™ total scores) via any type of predictor variable. The comparative fit of models containing and excluding the Item Type Score covariate, as well as the relative statistical importance of the Item Type Score covariate regression coefficient was used to judge whether the Item Type is able to predict CHI™ exam scores.

Weighted item scores were averaged together to create “Item Type Scales.” Exceptions were item types that were represented by only one item each (i.e., Shadowing, Reading Comprehension, and Listening Comprehension). Table 15 shows the relationships between item types and the CHI™ exam scores. Most item groups displayed low to moderate relationships with other item groups, with several pairs exhibiting negligible results (e.g., Medical Concepts and Shadowing, $r = 0.25$). The two item types observed to be strongest predictors of CHI™ scores were Restate the Meaning ($r = 0.40$) and Memory Capacity ($r = 0.42$). The latter two scales were also well related to each other ($r = 0.72$), and moderately related to most other item types. Other, more modest relationships are bolded in Table 15.

TABLE 15. CORRELATIONS BETWEEN WEIGHTED ITEM TYPES AND CHI™ SCALE SCORE

Scale	CHI™	1	2	3	4	5	6	7	8
1. Medical Concepts	0.33								
2. Restate the Meaning	0.40	0.45							
3. Fill in the Blank	0.26	0.27	0.41						
4. Finish the Sentence	0.30	0.18	0.46	0.32					
5. Equivalence of Meaning	0.39	0.39	0.64	0.42	0.44				
6. Memory Capacity	0.42	0.41	0.72	0.41	0.44	0.63			
7. Shadowing	0.25	0.19	0.45	0.26	0.34	0.32	0.41		
8. Reading Comprehension	0.23	0.16	0.41	0.18	0.37	0.32	0.30	0.31	
9. Listening Comprehension	0.11	0.21	0.41	0.31	0.21	0.31	0.46	0.31	0.15

The item types with the highest correlation to the CHI™ measures were chosen as the predictors. Specifically, Restate the Meaning, Equivalence, and Memory Capacity were used as predictors. These three indicators were able to account for 26% variance in CHI™ scores. Adding other item types to the predictive model did not improve R^2 (variance explained). Table 16 shows the results of the multiple regression analysis. Only Memory Capacity showed statistically high enough predictive power for the CHI™ scores. However, both Restate the Meaning and Equivalence were statistically significant when individually paired with Memory as predictors. Current results give evidence that some of the EtoE item types are related to and are predictive of the existing CHI™ measure. Note that the correlations are slightly lower than suggested in the preliminary study due to a modified weighting structure for the item scores.

TABLE 16. REGRESSION ANALYSIS RESULTS PREDICTING CHI™ SCORES

Coefficient	Estimate	Standard Error	t-value	p-value	Variance Explained
Intercept	325.89	19.82	16.44	<0.001	
Memory Capacity	246.41	77.36	3.185	<0.01	21%
Equivalence	138.27	99.85	1.385	0.17	17%
Restate the Meaning	196.21	137.49	1.427	0.16	23%

Confirmatory Factor Analyses for the EtoE Measure of Interpreting Ability

A confirmatory factor analysis (CFA) measurement model is essentially a linear regression model in which the main predictor, the factor, is latent or unobserved. It is theorized that the factor or factors are influencing the way in which candidates respond to items on an exam or a measure. Multiple models can be specified to describe a factor model, and these models can then be compared in terms of fit to the data. Maximum likelihood (ML) estimation was utilized to examine the measurement and structural model fit. All analyses were conducted using the R Package [lavaan](#). Prior to the specification of the confirmatory models, CCHI had outlined sets of items that could be grouped based on cognitive aspects of the tasks.

For the CFA study, multiple competing measurement models were identified based on Item Type. Note that Reading Comprehension and Medical Concepts items were not included due to the items' low relationships with other items and their practical fit/lack of fit in practical healthcare interpreting situations. The sparse data matrix with N = 249 candidates was used to take advantage of partial item-level information.

Model 1: One-Factor model (Overall Healthcare Interpreting Skill)

Model 2: Two Correlated Factors Model:

- 1) Factor 1: Restate the Meaning, Equivalence, Memory Capacity, Shadowing, Listening Comprehension
- 2) Factor 2: Fill-in-the-Blank, Finish the Sentence

Model 3: Three Correlated Factors Model:

- 1) Factor 1: Restate the Meaning, Equivalence
- 2) Factor 2: Fill-in-the-Blank, Finish the Sentence
- 3) Factor 3: Memory Capacity, Shadowing, Listening Comprehension

The ML χ^2 -square-goodness-of-fit statistic and a combination of absolute and incremental global fit indices were used to evaluate model fit. The χ^2 -square test is sensitive to sample size and provides a measure of exact fit. Unlike the χ^2 -square statistic, absolute and incremental fit indices are advantageous because they assess the degree of fit (Hu & Bentler, 1998).

Thus, three other recommended fit statistics (Martens, 2005) were used to assess model fit. The Root Mean Square Error of Approximation (RMSEA; Steiger & Lind, 1980) is an index of fit that adjusts for model complexity (i.e., fit is calculated holding the number of degrees of freedom constant). An RMSEA value of zero suggests that the model fits exactly; a value of .06 or less indicates relatively good fit, and a value above .08 indicates poor fit (Browne & Cudeck, 1992; Hu & Bentler, 1999). The Standardized Root Mean Square Residual (SRMR; Bentler, 1995) is a standardized index of the average size of the discrepancy between the elements in the observed and predicted covariance matrices (Kline, 2005). A value less than or equal to .08 indicates good fit (Hu & Bentler, 1999). The Comparative Fit Index (CFI; Bentler, 1990) is interpreted as the proportion of improvement in overall model fit when compared to a null model. That is, the CFI compares the overall model fit to that of a model in which all relationships are set to zero. The CFI ranges between 0 and 1, with .95 or higher indicating a good fit (Hu & Bentler, 1999). Marsh, Hau and Wen (2004) point out that acceptable fit should be judged by the researchers, not golden rules, and many indicators of model fit should be examined (e.g., residuals).

TABLE 17. FIT INDICES ACROSS THREE MODELS

Model	SRMR	RMSEA	CFI	χ^2	df	χ^2 -change	df-change	p
One Factor	0.069	0.067	0.81	684.96	377			
Two Factor	0.069	0.061	0.84	633.31	376	51.65	1	<.0001
Three Factor	0.069	0.059	0.85	615.90	374	17.41	2	<.001

As displayed in Table 17, the overall fit for Model 1 was not adequate in that only SRMR index was in the bounds of acceptable fit (SRMR > .08; RMSEA > .06; CFI < .95; $\chi^2(377) = 684.96$, $p < .0001$). Further, the one-factor structure may not be conclusively confirmed given the statistically significant χ^2 statistic. The two-factor model improved on Model 1, however both the RMSEA and the CFI were still below the thresholds of acceptable fit, and the χ^2 statistic was significant. Finally, Model 3 significantly improves on Model 1 in terms of data fit, and although with much unexplained variance left in the data ($\chi^2(374) = 615.90$, $p < .0001$) had adequate values for the SRMR and RMSEA indices.

As a follow up analysis, a fourth model was created based on the observations from the EFA and the CFA. The EFA returned very low factor loadings on both a two-factor and a three-factor solution specifying uncorrelated factors (i.e., items were grouped into a maximum of two factors that were theorized to be unrelated to each other, e.g., “working memory” versus “recall”) from “Fill in the Blank” item group. CFA Model 4 capitalized on these items’ lack of relationships with other items on the EtoE measure by excluding them from the factor structure. However, no significant improvement to model-data fit was observed.

To further explore local misfit of the data to the models, it is useful to review standardized factor loadings. Standardized loadings can be compared and help provide insight into potentially problematic items. Factor loadings from each of the three models have been included in *Appendix I*.

CONCLUSIONS

This report consists of a summary seeking to provide evidence for the EtoE measurement tool, an exploratory assessment with the purpose of measuring interpreting ability without considering bilingual ability in healthcare settings. Generally, the EtoE shows promise in revealing candidates’ ability to interpret within a healthcare environment. Within the study, multiple indicators of the measure’s validity and reliability were addressed. Comparisons of relationships of the EtoE to the Certified Health Interpreting (CHI™) across groups of interest showed that scores of native speakers of the Language Other Than English (LOTE) had stronger relationships to the CHI™ scores than those of heritage speakers and nonnative speakers of the LOTE (i.e., native speakers of English). Classical test analyses showed that most items had moderate to strong discrimination, indicating a moderately high level of distinguishing between candidate performance on this measure. Rater reliability was high, indicating that rater training and rubric quality enabled raters to use the rubric consistently.

The EtoE measure displayed moderate relationships with the current bilingual performance-based measures, the CHI™ exams (administered in Spanish, Mandarin, and Arabic). The strength of the relationships depended on the language for the assessment; however, low sample sizes encumbered the interpretation of the Arabic and Mandarin language test relationships due to a low volume of candidates that were able to take both the CHI™ and the EtoE.

Several confirmatory factor models were evaluated to assess whether groups of items based on the type of tested cognitive ability would fit the data adequately across all CHI™ test takers. Multiple absolute and relative fit indices were evaluated. Based on the fit values and the comparisons, it appeared that interpreting skill in the context of healthcare services displays a better relative fit for a multi-factor model than a one-factor model. However, relevant factors may cross over in the cognitive abilities necessary to complete the items designed to measure them and were therefore moderately correlated.

As in the preliminary study results, particular sections of the assessment did not appear to relate strongly to the overall measure of interpreting ability. Specifically, Medical Concepts, Fill-in-the-Blank, and Reading Comprehension did not relate well to other item types and the overall interpreting ability scores. Listening Comprehension had moderate relationships to the rest of the assessment sections; however, these indices should also be viewed with caution given that only one item was assigned to this category.

Multiple choice questions relating to Medical Concepts do not appear to reveal healthcare interpreting skill levels. Moreover, more items targeting Listening Comprehension and Reading Comprehension need to be developed in

order to make any conclusions about the relationships between these item types and the attribute of interpreting skill.

Item type, the variable used to determine “factors” for the confirmatory factor analyses included in this study, may not be the best grouping indicator for further factor analyses. Certain items may share the cognitive functions necessary to respond to multiple types of items in a way that distinguishes between persons with different healthcare interpreting ability. Those items were grouped together to understand further the construct of interpreting ability; however, items within Fill-in-the Blank and Finish the Sentence groupings should be reviewed for their ability to distinguish between lower levels of healthcare interpreting ability. Improvements to those items through item revisions may need to focus on making the items more challenging.

Overall, the EtoE is a promising measure that requires additional revision and piloting prior to use for high-stakes testing. Its strengths include robust rubrics and rater training, high-quality items, and the prospect of measuring interpreting skill in languages of lower incidence. It is significant that score relationships between the EtoE and the CHI™ exam were strongest for native speakers of the LOTE: the intended target audience for the EtoE exam is the lower incidence language interpreter population, which is largely comprised of native speakers of the LOTE.

To improve the measure, we recommend revisions to items that revealed suboptimal item statistics and eliminating item types that as a group revealed low relationships to the rest of the assessment scores. After refining the measure, recommended next steps are to expand the study with targeted analyses based on a larger sample size.

Part III. EtoE Study Additional Observations

CCHI considers the EtoE Study of utmost importance not only for interpreters and certification development decisions but also for the interpreter educators. The observations presented in this Part are based on the responses of 176 participants who completed both the EtoE and CHI™ exams and filled out the *EtoE Study Participation Questionnaire (Appendix E)*. The information presented here is of descriptive nature, and further analysis of the data will be pursued.

To allow for a rough comparison of participants' performance on the CHI™ and EtoE exams, data about the EtoE scores of 60 and 70 is provided as a guide, on the assumption that the EtoE exam's passing point would not be lower than 60% (the EtoE maximum raw unweighted score is 99 points).

Caution should be exercised when interpreting the data provided below both because some subgroups have statistically low numbers of responses and because only the unweighted scores are reported.

Professional Affiliation and Experience

It is common knowledge that many interpreters in the U.S. work across several industry sectors (healthcare, legal, educational, conference, business) or have a second occupation, which can be either language-related (translation, language teaching, interpreter education) or healthcare-related (nursing, allied health, managing language services) or totally unrelated to either interpreting or health care. Additionally, the cohort of interpreters seeking CCHI's certification is diverse in relation to their experience due to the voluntary nature of healthcare interpreter certification in the U.S.

As anticipated, participants for whom interpreting or translation were the main occupation and main means of livelihood, performed better on both the EtoE and CHI™ examinations (Table 1).

Table 1. Examination scores and interpreting/translation being the main occupation

	Yes	No
Participants Count, N	128	48
Passed CHI™, Count	90	25
Passed CHI™, %	70%	52%
EtoE score 60 & higher, Count	107	35
EtoE score 60 & higher, %	84%	73%
EtoE score 70 & higher, Count	71	19
EtoE score 70 & higher, %	55%	40%

It appears that freelance and staff interpreters perform in a similar way on the exams (Table 2). It is hard to assess of employment status has any impact on the other sub-groups of participants since they are so low in numbers.

Table 2. Examination scores and employment status

	Freelancer	Staff	Dual-role	Volunteer	Don't interpret in HC
Participants Count, N	71	59	17	11	18
Passed CHI™, Count	51	43	6	4	11
Passed CHI™, %	72%	73%	35%	36%	61%
EtoE score 60 & higher, Count	58	50	12	9	13
EtoE score 60 & higher, %	82%	85%	71%	82%	72%
EtoE score 70 & higher, Count	35	37	6	6	6
EtoE score 70 & higher, %	49%	63%	35%	55%	33%

While the level of participants' experience in healthcare interpreting seems to correlate with their performance on the EtoE and CHI™ exams, the data demonstrates that both exams are appropriately designed for the entry-level interpreter (Tables 3 and 4). The results also confirm that, in the context of the U.S. healthcare interpreting industry, **experience alone cannot be a measure of the interpreter's qualifications**: 35% of participants who perceive themselves as "experienced" or "very experienced" failed the CHI™ certification examination and one-fifth of them (19%) performed below the EtoE exam score of 60 (Table 3). The objectively reported years of experience (Table 4) demonstrate the same: of 59 participants whose experience is reported as 6 years and more, 39% failed the CHI™ certification examination and one-fifth of them (20%) performed below the EtoE exam score of 60.

Table 3. Examination scores and healthcare interpreting experience (subjective)

	Novice	Early career	Experienced	Very experienced
Participants Count, N	57	47	46	26
Passed CHI™, Count	33	35	30	17
Passed CHI™, %	58%	74%	65%	65%
EtoE score 60 & higher, Count	43	41	38	20
EtoE score 60 & higher, %	75%	87%	83%	77%
EtoE score 70 & higher, Count	27	24	28	11
EtoE score 70 & higher, %	47%	51%	61%	42%

Table 4. Examination scores and healthcare interpreting experience (years)

	Less than 2 years	2 to 5 years	6 to 10 years	11 to 15 years	16 to 20 years	21 or more
Participants Count, N	68	49	36	14	5	4
Passed CHI™, Count	46	33	24	8	1	3
Passed CHI™, %	68%	67%	67%	57%	20%	75%
EtoE score 60 & higher, Count	57	38	30	11	3	3
EtoE score 60 & higher, %	84%	78%	83%	79%	60%	75%
EtoE score 70 & higher, Count	35	25	19	6	2	3
EtoE score 70 & higher, %	51%	51%	53%	43%	40%	75%

While levels or years of experience do not seem to correlate to success on either examination, frequency of interpreting per week (in hours), regardless of the setting, seems to display such a correlation (Table 5). Participants who interpret more than 20 hours per week appear to have better chances of passing either examination.

Table 5. Examination scores and all hours of interpreting (in any setting)

	Less than 2 hours	3-20 hours	21 - 40 hours	41 hours and over
Participants Count, N	31	67	67	11
Passed CHI™, Count	18	40	47	10
Passed CHI™, %	58.1%	59.7%	70.1%	90.9%
EtoE score 60 & higher, Count	23	53	58	8
EtoE score 60 & higher, %	74.2%	79.1%	86.6%	72.7%
EtoE score 70 & higher, Count	13	30	42	5
EtoE score 70 & higher, %	41.9%	44.8%	62.7%	45.5%

English Language Acquisition

One of the concerns of the Commissioners, the EtoE Study National Task Force members, the SMEs who developed and rated the EtoE exam, and the interpreters, is whether or not the monolingual EtoE exam would give advantage to interpreters for whom English is a native language or to heritage speakers who acquired English as part of their living and schooling in the U.S. from an earlier age than the interpreters for whom English is second language. The data presented in Table 6 indicates that native English and heritage speaker participants (last column) appear to have an advantage of earning a score of 60 and higher on the EtoE exam: 87.9% compared to 76.6% of non-native English-speaking participants. And while this native English/heritage speaker group is relatively small (n=66), the displayed advantage is worthy of consideration. The Commission is planning to explore the possibility of compensating for this via an adjusted eligibility requirement of language proficiency in the LOTE for this subgroup of applicants. At the same time, it is important to keep in mind that most of interpreters of languages of low incidence (the target audience for the future ETOE™ examination) are not native speakers of English but are native speakers of a LOTE. Spanish interpreters who participated in this study represent the largest language group where the proportion of native English and heritage Spanish speakers is somewhat significant. This fact is not surprising given the state of the world languages education in the U.S. where most university programs that teach to the level of language mastery needed for interpreting are for Spanish.

Table 6. Examination scores and English/LOTE acquisition*

	Non-native English speakers	Native English speakers	Heritage speakers of LOTE**	Native English & Heritage speakers combined
Participants Count, N	111	43	23	66
Passed CHI™, Count	72	30	14	44
Passed CHI™, %	64.9%	69.8%	60.9%	66.7%

	Non-native English speakers	Native English speakers	Heritage speakers of LOTE**	Native English & Heritage speakers combined
EtoE score 60 & higher, Count	85	39	19	58
EtoE score 60 & higher, %	76.6%	90.7%	82.6%	87.9%
EtoE score 70 & higher, Count	46	33	13	46
EtoE score 70 & higher, %	41.4%	76.7%	56.5%	69.7%

* This category has full data for 177 respondents.

**Only Spanish-speaking participants are represented in this group.

Healthcare Interpreting Education

In the U.S., healthcare interpreting remains a profession where the majority of practitioners receive non-academic education in interpreting and in healthcare-related subjects relevant to interpreters. Over the last four-five years, more academic programs, usually at the Associate degree or vocational certificate level, as well as non-academic programs of over 40-hours duration have been emerging. It is important to observe these trends and measure if these emerging programs adequately prepare the interpreter for the certification exam.

Table 7 presents the results according to the duration of healthcare interpreter training. Keep in mind that seven participants, who responded having less than 40 hours of training, meet this CCHI eligibility requirement. All of them have had 30 or 35 hours of training in healthcare interpreting specifically, and 5 or 10 hours in related fields, e.g., court or conference interpreting, translation, or healthcare specialty. (Some CCHI certification candidates are physicians, nurses, or other allied health professionals, especially, if they have received their medical education overseas). Increase in duration of training beyond 40 hours indicates somewhat higher passing rate for the CHI™ exams and better performance on the EtoE exam. However, the subgroups have relatively low numbers of participants, especially in the categories of “Less than 40 hours” and “66-100 hours.”

Table 7. Examination scores and duration of training in healthcare interpreting

	Less than 40 instructional hours	40 instructional hours	41-65 instructional hours	66-100 instructional hours	Over 100 instructional hours*
Participants Count, N	7	48	52	27	42
Passed CHI™, Count	1	31	35	15	33
Passed CHI™, %	14.3%	64.6%	67.3%	55.6%	78.6%
EtoE score 60 & higher, Count	5	38	45	20	34
EtoE score 60 & higher, %	71.4%	79.2%	86.5%	74.1%	81.0%
EtoE score 70 & higher, Count	3	21	32	12	22
EtoE score 70 & higher, %	42.9%	43.8%	61.5%	44.4%	52.4%

* Includes 1 respondent with an Associate Degree in HCI and 1 respondent with a master’s degree in HCI. Both passed the CHI™ exam; respondent with the Associate degree had the EtoE score over 70 points, and the one with the master’s degree – over 60 points.

While dividing the data into subgroups by the method of acquiring healthcare interpreting education (Table 8) produces subsets with even lower number of participants, the information is of interest to educators. Further

research is needed about the subgroup of participants who indicated completion of 45-hour academic courses (3 U.S. credits). This is the least represented group, and the low performance markers may be due to various factors, including non-alignment of curriculum with the practice of healthcare interpreting, courses focusing on the interpreter knowledge rather than interpreting skills, monolingual instruction, etc. Overall, more data is needed to allow for meaningful insights.

Table 8. Examination scores and method of healthcare interpreting training

	Academic, 45 hours	Academic, over 45 hours	Non- academic, in-person, 40-100 hours	Non- academic, online, 40- 100 hours	Combination of workshops/ courses less than 40 hours each (any modality)	On-the job training
Participants Count, N	15	26	72	19	17	27
Passed CHI™, Count	5	20	50	13	9	18
Passed CHI™, %	33.3%	76.9%	69.4%	68.4%	52.9%	66.7%
EtoE score 60 & higher, Count	9	23	58	18	13	21
EtoE score 60 & higher, %	60.0%	88.5%	80.6%	94.7%	76.5%	77.8%
EtoE score 70 & higher, Count	5	14	42	8	7	14
EtoE score 70 & higher, %	33.3%	53.8%	58.3%	42.1%	41.2%	51.9%

As anticipated, participants who completed university-level courses in interpreting, translation, or linguistics performed better on both examinations (Table 9). Interestingly, courses in translation appear to be most effective which might be explained by the fact that, on the one hand, they focus on transfer of meaning compared to linguistics courses, and, on the other, they tend to be more “established” in terms of curriculum and methods of teaching than interpreting ones.

Table 9. Examination scores and university-level courses in interpreting, translation, or linguistics

	University courses in interpreting		University courses in translation		University courses in linguistics	
	Yes	No	Yes	No	Yes	No
Participants Count, N	80	96	39	137	49	127
Passed CHI™, Count	54	61	33	82	38	77
Passed CHI™, %	68%	64%	85%	60%	78%	61%
EtoE score 60 & higher, Count	67	75	36	106	43	99
EtoE score 60 & higher, %	84%	78%	92%	77%	88%	78%
EtoE score 70 & higher, Count	43	47	27	63	28	62
EtoE score 70 & higher, %	54%	49%	69%	46%	57%	49%

Review of the data regarding the duration of training in a specific interpreting mode and performance on the examinations (Tables 10, 11, and 12) did not produce any definitive findings. It appears that training below seven (7) hours of duration is not effective enough to impact performance on the CHI™ examination. Overall, duration as the only variable of training does not seem to provide meaningful information. Further analysis of several factors is needed.

Overall, it could be inferred that in the context of the U.S. healthcare interpreting industry, *interpreter training alone cannot be a measure of the interpreter’s qualifications.*

Table 10. Examination scores and hours of training in consecutive mode

	0-6 hrs	7-24 hrs	25-45 hrs	≥ 45 hrs
Participants Count, N	27	38	43	68
Passed CHI™, Count	14	25	28	48
Passed CHI™, %	52%	66%	65%	71%
EtoE score 60 & higher, Count	23	32	33	54
EtoE score 60 & higher, %	85%	84%	77%	79%
EtoE score 70 & higher, Count	19	22	14	35
EtoE score 70 & higher, %	70%	58%	33%	51%

Table 11. Examination scores and hours of training in simultaneous mode

	0-6 hrs	7-24 hrs	25-45 hrs	≥ 45 hrs
Participants Count, N	74	40	28	34
Passed CHI™, Count	42	30	20	23
Passed CHI™, %	57%	75%	71%	68%
EtoE score 60 & higher, Count	56	37	23	26
EtoE score 60 & higher, %	76%	93%	82%	76%
EtoE score 70 & higher, Count	41	23	9	17
EtoE score 70 & higher, %	55%	58%	32%	50%

Table 12. Examination scores and hours of training in sight translation

	3-6 hrs	7-24 hrs	25-45 hrs	≥ 45 hrs
Participants Count, N	83	40	26	27
Passed CHI™, Count	52	25	19	19
Passed CHI™, %	63%	63%	73%	70%
EtoE score 60 & higher, Count	68	32	21	21
EtoE score 60 & higher, %	82%	80%	81%	78%
EtoE score 70 & higher, Count	50	21	8	11
EtoE score 70 & higher, %	60%	53%	31%	41%

The study did not provide any evidence that deliberate exposure to additional content either in English or LOTE affects performance on either examination (Tables 13 and 14). This aspect of the interpreter experience would require further research.

Table 13. Examination scores and time watching or listening to content in English and LOTE

	≤1 hr/ week		2-7 hrs/ week		8-14 hrs/ week		≥ 15 hrs/ week	
	ENG	LOTE	ENG	LOTE	ENG	LOTE	ENG	LOTE
Participants Count, N	12	52	71	91	44	18	49	15
Passed CHI™, Count	6	38	50	58	28	10	31	9
Passed CHI™, %	50%	73%	70%	64%	64%	56%	63%	60%

	≤1 hr/ week		2-7 hrs/ week		8-14 hrs/ week		≥ 15 hrs/ week	
EtoE score 60 & higher, Count	9	47	59	70	36	14	38	11
EtoE score 60 & higher, %	75%	90%	83%	77%	82%	78%	78%	73%
EtoE score 70 & higher, Count	4	35	46	46	21	4	19	5
EtoE score 70 & higher, %	33%	67%	65%	51%	48%	22%	39%	33%

Table 14. Examination scores and reading time in English and LOTE

	≤1 hr/ week		2-7 hrs/ week		8-14 hrs/ week		≥ 15 hrs/ week	
	ENG	LOTE	ENG	LOTE	ENG	LOTE	ENG	LOTE
Participants Count, N	15	46	60	88	48	23	53	19
Passed CHI™, Count	10	32	41	58	32	13	32	12
Passed CHI™, %	67%	70%	68%	66%	67%	57%	60%	63%
EtoE score 60 & higher, Count	9	38	47	74	41	17	45	13
EtoE score 60 & higher, %	60%	83%	78%	84%	85%	74%	85%	68%
EtoE score 70 & higher, Count	4	28	31	46	32	10	23	6
EtoE score 70 & higher, %	27%	61%	52%	52%	67%	43%	43%	32%

Additional Confirmatory Analyses

Following recommendations of the *EtoE Study Follow-up Validation Report* by Prometric LLC (see *Part II*), CCHI has reviewed the factor loadings of the proposed three-factor model as well as the items statistics of other analyses. Based on the review, it appears that the item types of *Fill-in-the-Blank* and *Finish the Sentence* do not display strong performance within the EtoE test and may contain significant cognitive attributes beyond interpreting.

Thus, CCHI conducted confirmatory **three-factor analysis** of the data with these item types removed and consisting of:

- 1) Factor 1 – Meaning conversion: *Restate the Meaning, Meaning Equivalence*
- 2) Factor 2 – Meaning retention: *Memory Capacity*
- 3) Factor 3 – Meaning comprehension: *Shadowing, Listening Comprehension*

This three-factor model appears to be a somewhat better fit compared to the three-factor model that includes the item types of *Fill-in-the-Blank* and *Finish the Sentence*, and this reflects the practice as the “meaning prediction” functions are not widely utilized by interpreters in the consecutive mode. Taking into account the recent research cautioning against application of universal cutoff values of fit indices to determine adequate model fit¹³, it appears that this model reasonably reflects the construct of interpreting ability. The **four-factor**

¹³ Marsh HW, Hau K, Wen. Z. (2004) In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler's 1999 Findings. *Structural Equation Modeling* 2004; 11:320–41.
Chen F, Curran PJ, Bollen KA, Kirby J, Paxton P. (2008) An Empirical Evaluation of the Use of Fixed Cutoff Points in RMSEA Test Statistic in Structural Equation Models. *Sociol Methods Res.* 2008 January 1; 36(4): 462–494.

model that includes these item types as Factor 4 “Meaning prediction” does not provide a better fit. Table 9 presents the fit indices for both models.

Table 9. Fit indices for a three-factor and four-factor model

Model	SRMR	RMSEA	CFI	AIC	BIC	χ^2	χ^2 - change	df	df- change	p
Three Factor, 21 items	0.060	0.069 (p=0.003)	0.88	5743.2	5889.1	354.31		186		0.000
Four Factor, 29 items	0.068	0.059 (p=0.048)	0.86	7462.0	7665.6	600.05	245.74	371	185	0.000

Conclusion

The EtoE Study offers valuable information for developing a performance credentialing examination in a monolingual modality for interpreters of any language. The results present promising evidence of the potential efficacy of the EtoE examination in measuring interpreting skills in a monolingual format. They indicate that performance on the monolingual EtoE exam moderately correlates with the performance on the dual-language CHI™ exam. This correlation is stronger in the group of native speakers of the Language Other Than English (LOTE) ($r=0.62$, $n=111$) compared to native speakers of English ($r=0.29$, $n=43$). This is significant given the fact that the intended audience of the future ETOE™ examination are interpreters of languages of low incidence who are overwhelmingly native speakers of LOTE.

The study confirmed that the EtoE test included high-quality items, robust scoring rubrics, and consistent ratings by human raters. The study helped identify the item types that do not relate strongly to the overall measure of cognitive interpreting skills. Namely, *Medical Concepts*, *Fill-in-the-Blank*, and *Reading Comprehension* items did not relate well to other item types and the overall scores.

Review of several model fit indices combined with the analyses of the cognitive functions necessary to respond to different item types by subject matter experts suggests that the EtoE test measures interpreting abilities related to meaning comprehension, retention, conversion, and prediction to an extent significant to distinguish between persons with different interpreting ability in the healthcare context.

CCHI will take into account the EtoE Study findings to develop the future monolingual ETOE™ credentialing examination for healthcare interpreters. This future exam will allow to assess interpreting ability of candidates working in languages of low incidence, thus, ensuring equity and inclusiveness of CCHI's certification program. At the same time, CCHI will continue administering and developing bilingual interpreter performance exams for languages of high incidence (such as currently administered Arabic, Mandarin and Spanish CHI™ exams).

The information collected by the *EtoE Study Questionnaire* is also of importance to the interpreting profession overall and to interpreter educators. The observations gleaned from it display a significant diversity among the study participants. This diversity makes it possible to infer that this group is reasonably representative of the healthcare interpreting profession in general, and the results of the study could be applicable to interpreters of other languages. This report provides mostly description of the information; CCHI will continue analysis of the collected data in future publications.

Appendix A. Designing an English-to-English Interpreting Performance Test: Recommendations of the National Task Force (January 2019)

Acknowledgements

This project was led by **Margarita S. Bekker**, CoreCHI™, CCHI Chair, and **Natalya Mytareva**, M.A., CoreCHI™, CCHI Executive Director. This publication was prepared by Natalya Mytareva, the project's principal investigator.

CCHI expresses its gratitude to the EtoE National Task Force Panelists:

Enrica Ardemagni, Ph.D., CHI™-Spanish, (IN) Indiana University Purdue University Indianapolis, Professor Emerita of Spanish, President, NCIHC

Rosanna Balistreri, M.A., (CA) REACH-Reaching diversity, Cultural & Linguistic Consultant

Joy Connell, (MA) Massachusetts State Department of Mental Health, NCIHC

Esther Diaz, (TX), Translator and Interpreter Trainer, Austin Community College

Nora Goodfriend-Koven, M.P.H., (CA) Interpreter Educator & Consultant

Carola E. Green, B.S., (VA) Interpreter Educator & Consultant, FCCI (Federally Certified Court Interpreter English-Spanish)

Manuel Higginbotham, CHI™-Spanish, (TX) University of Texas Medical Branch, Manager of Language Access Services; Tex-Med Training Services, President

Jane Crandall Kontrimas, M.S., CoreCHI™, (MA) Beth Israel Deaconess Hospital, Russian Interpreter and Interpreter Training Coordinator

Katherine Langan, Ph.D., (IA) Mercy Medical Center – Des Moines, Interpreter Trainer, Sociolinguist

Elena Langdon, M.A., CT, CoreCHI™, (MA) Acolá Language Services & Consulting, Interpreter Educator

Gerardo Lazaro, A.B.D., CHI™-Spanish, (PA), The National Institute for Coordinated Healthcare, Interpreter Trainer

Jonathan Levy, M.A., (AZ) TransPerfect, Senior Director, Language Solutions

David McCoy-Galicia, CHI™-Spanish, CMI, (CA) Director of Medical Interpreting Training School (MITS)

Tim Moriarty, M.P.A., CMI, CHI™-Spanish, (MA) Baystate Health, Manager, Interpreter & Translation Services

Johanna Parker, M.A., USCCI, CHI™-Spanish, (CA) Stanford Health Care Lead Interpreter for Education and Training

Cynthia E. Roat, M.P.H., (WA) Language Access Consultant & Interpreter Trainer

Karin Ruschke, M.A., CoreCHI™, (IL) International Language Services, Inc., President

E. Zoe Schutzman, M.A., NYS-Certified Court Interpreter, CHI™-Spanish, (NM) Educator & Staff Development Specialist at The University of New Mexico Health Science Center's UNM Hospitals

Judy Shepard-Kegl, Ph.D., NIC-M, SC:L, ED:K-12, CoreCHI™, (ME) Linguistics Department, University of Southern Maine, ASL interpreter

Gabriela Espinoza Siebach, M.A., CHI™-Spanish, (TX) MasterWord®, Director of Business Development

Monica Thomasini, CoreCHI™, (CA) Quality Assurance Trainer at Language World Services

Melissa Wallace, Ph.D., (TX) University of Texas at San Antonio, Dept of Modern Languages and Literatures

CCHI Consultant:

James P. Henderson, Ph.D., Senior Vice President, Senior Psychometrician, Castle Worldwide, Inc./Scantron®

Contents

Healthcare Interpreter Competencies and Job Tasks	35
Test Item Types	36
Item content (scripts and texts)	59
Scoring Recommendations	59
Test Taker Questionnaire and Preparation Guide	60
Conclusion	60

Introduction

The Certification Commission for Healthcare Interpreters (CCHI) has administered the national certification programs for healthcare interpreters since 2010. The two currently available certifications – Core Certification Healthcare Interpreter™ (CoreCHI™) and Certified Healthcare Interpreter™ (CHI™ - Spanish, Arabic, Mandarin) – are aimed at the entry-level healthcare interpreter.

The entry-level certified healthcare interpreter is defined as:

A person who is able to perform the functions of a healthcare interpreter competently, independently, and unsupervised in any setting and in any modality where health care is provided, with the knowledge, skill, and ability required to relay messages accurately from a source language to a target language in a culturally competent manner and in accordance with established ethical standards.

In June 2018, as part of exploring feasibility of developing an interpreter performance examination in a monolingual (English-to-English, EtoE) modality, CCHI convened the **EtoE National Task Force Panel** of 22 experts. Some of the experts participated in the focus group discussions of this project in the Fall of 2017.¹⁴ The Panel's goal was to provide recommendations to CCHI about the types of items to include in the English only (EtoE) interpreter performance exam. The Panel met remotely twice under the guidance of CCHI's psychometric consultant Dr. James P. Henderson of Castle Worldwide/Scantron Corporation. The panelists also worked in small groups via email and conference calls. The final recommendations were reviewed via email discussion.

The present document reflects the discussions and opinions of the participants, CCHI Commissioners, and project principal.

CCHI will use this information for training Subject Matter Experts (SMEs) who will develop items for the EtoE test and for the preparation of instructions to candidates taking the exam.

Healthcare Interpreter Competencies and Job Tasks

Following the recommendations of the focus groups' discussion (see the previous publication "Assessing Healthcare Interpreting Performance Skills in an English-to-English Format"¹⁵), the EtoE National Task Force panelists identified test items that can potentially assess cognitive interpreting skills in an efficient manner in a monolingual, EtoE format.

¹⁴ See CCHI's publication about these discussions at http://cchicertification.org/uploads/CCHI_EtoE_Interpreter_Performance_Assessment.pdf.

¹⁵ *Ibid.*

The participating experts agree that a significant number of these cognitive-linguistic skills may be assessed to a relevant degree in a monolingual, EtoE format. Namely:

- Active listening/reading
- Anticipatory listening/reading
- Message analysis
- Comprehension of oral speech/written text
- Retaining and recalling information (short-term memory)
- Accurate reformulation of the source speech/text in the same language (fidelity to the message)
- Understanding of the concept of 'register'
- Attention-sharing skills
- Fluency (lexical and grammatical) in English (speech production)
- Speech quality in English (pronunciation, prosody, pace/speed)

The National Task Force panelists applied the following criteria to identifying item types recommended for this study:

- Item assesses a skill or subskill identified in the Job Task Analysis Study as part of the healthcare interpreter's skills.

For each proposed item type the panelists identified the interpreter's knowledge, skills and abilities/competencies (KSA/C's) needed to perform successfully. The item's KSA/C's have direct relation to the interpreter's KSA/C's defined in the 2016 Job Task Analysis Study by CCHI.¹⁶

- Such a skill/subskill can be assessed in English, i.e., in a monolingual format.
- Such a skill/subskill is perceived as critical to the interpreter's performance.
- Such a skill/subskill is expected of an interpreter at the entry level.

Test Item Types

The following test item types are proposed for consideration as they could potentially measure the interpreter's cognitive-linguistic skills in a monolingual, EtoE format.

1. Bilingual/interlingual reformulation item – Audio to Audio

Item description:

This item consists of two activities performed by candidates which are the first and last one on the test:

- a) Item #1 on the test: Candidates listen to the audio recording of a conversation between a provider and patient, presented in English, and record their consecutive interpretation of the conversation into the Language Other than English (LOE). The conversation is presented as eight audio recordings, and the candidates produce eight audio recordings in LOE.
- b) Last item on the test: Candidates listen to the eight audio recordings of their own interpretation in LOE and consecutively interpret them into English.

The reason for separating these two activities is to prevent candidates from memorizing the original English prompt instead of interpreting what they hear on their own LOE interpretation of that prompt.

¹⁶ Full text at http://cchicertification.org/uploads/CCHI_JTA2016_Report.pdf.

It is recommended to include at least one other test item using the same healthcare specialty topic between these two activities in order to further prevent candidates' direct recall of the original prompt.

Job tasks and interpreting skills measured:

The main skill assessed by comparing accuracy of the English output (interpreting from LOE) to the English input (source, exam prompt) is the skill of meaning-oriented reformulation, not of actual interpreting.

Knowledge of:

- English terminology, idioms, usage, and cultural significance
- Structure and grammar of English

Skill in:

- Retaining and recalling information from short-term memory
- Listening actively
- Anticipatory listening
- Communicating fluently in English
- Maintaining accuracy/ including register
- Sufficient mastery of English pronunciation/quality of speech to avoid impact on understanding
- Self-monitoring for accuracy in English
- Reformulation of message without additions, omissions, or substitutions (as a subskill of interpreting)

Item development recommendations:

- Develop an English script of a conversation between a provider and patient following the same parameters as for the consecutive interpreting item on the CHI™ examination: 265-380 words; up to 35 words per utterance; 4-6 terms; 2-4 colloquial or idiomatic expressions; approximately 100-135 seconds of the total recording – middle of the range preferred. For test construction convenience, limit the number of utterances to 8.
- *For Item # 1 on the test:* Each utterance of the conversation may be played twice before candidates record it, i.e., exactly as with a consecutive item on the CHI™ examination. For the last item on the exam (reformulation back into English): Each utterance is played once to avoid giving the candidate extra cues to remember the original English prompt of Item #1.
- Candidates are allowed to take notes as they are allowed to do so during the consecutive part of the CHI™ examination.
- Start the content of the conversation after the initial greetings are over, so that each utterance contains information pertinent to the patient's condition and/or treatment and avoid including courtesy phrases as a full utterance (e.g., thank you, how are you, goodbye). The purpose is to have enough linguistically meaningful content that can test the candidates' interpreting abilities. It is advisable to include a "background" scenario in the onscreen instructions before the item to orient the candidate to the situation: who are the speakers ("title" – doctor/nurse, etc. and patient/family member, etc.), specialty, and type of medical appointment.
- Avoid using medical terms and words that could be direct borrowings from English in other languages, especially terms with Latin roots or derivational morphemes (prefixes, suffixes). E.g., "television," "psychiatrist," "chemotherapy." Keep in mind, that languages such as Hindi, Nepali, i.e., countries with a history of the British influence on medicine have more English borrowings of medical terms than others. Try to include SMEs of such languages in the development of this item type. If an English word is borrowed by LOE candidates, then we are not testing the candidates' reformulation skills.

NOTE to SMEs: The reason for avoiding such "borrowing-prone" terms is to maintain a reasonably equal level of difficulty of the item for candidates of *any* language. For example, if a French candidate uses the English term in their response because they don't remember or know a correct French equivalent, but a Nepali candidate uses the English term because this is the only way to render the term in Nepali, then it

presents a problem to train raters (who are listening to the LOE recording to ascertain that candidates did not just repeat the prompt in English) to distinguish in which case the repeated English word is an error and in which it is not.

- Identify units of meaning that must be maintained by candidates when they interpret from LOE into English.
- Some specific instructions for other items on the test form are intended to keep intact the integrity of what we are trying to measure by this item.

Special suggestions on construction of dialogs:

1. Follow already established CCHI guidelines for dialogs development.
2. Have special vocabulary items—technical and cultural terms, idioms, etc.—appear in two utterances by the same speaker. These utterances do not need to be consecutive. This will allow raters to evaluate how the candidate handles the same item more than once.
3. Include cohesion between speakers and utterances (turns) so we can evaluate how the candidate maintains coherence throughout dialog as appropriate.
4. Be very specific and clear when choosing topics, terms and structures to avoid ambiguity which will become even more ambiguous in reformulation.
5. Avoid in the item inherently American English culturally determined expressions, e.g., idioms, expressions of politeness, etc. For example, in American English politeness may be expressed via lexical and grammatical means (modal verbs, tense, word order or syntax). Yet, in other languages it may be expressed via intonation or with a different grammatical means. It will be harder for raters to know in this case if, for example, a change of tense in the reformulated English response is an error or not. To sum up, keep the word/expression choices of the prompt as culturally neutral as possible.

Instructions to candidates:

Item #1 on the test:

Listen to the conversation between a provider and a patient presented in English only and interpret it consecutively into your non-English language. There are eight utterances in the conversation. You will listen to and record one utterance at a time. You can play each utterance in English two times before recording it.

Last item on the test:

Listen to the recording of your own interpretation of the first item on this test – it was a conversation between a provider and a patient. There are 8 utterances total. Record your consecutive interpretation of each utterance back into English. You can play each utterance only once.

Scoring recommendations:

- Decide whether raters will use a combination of analytic and anchored scales or just one type of scoring for this item.
- If using anchored scales – decide if the ones used on the CHI™ exam are appropriate: Lexical Content, Register, Grammar, and Quality of Speech.
- See the proposed 5-6 scoring scales **below. Consider using the same scales for the intralingual reformulation (paraphrase) item.**

Example of an item with Scoring Considerations:

Weight of the item on the test form: At this time the weight of this section of the exam has not been determined, but it will need to be reflected in the construction of the dialogs. This section of the exam has the potential to be the most important component for demonstrating cognitive skills. This section must be carefully constructed to achieve this potential. However, we recognize that the final utility and validity of this section will need to be determined after the pilot project is analyzed. If this item correlates best, then the actual (post-study) exam should probably include more items of this and fewer (or none) of any items that do not correlate.

Procedure of creating an item: One expert wrote the initial dialog. Another expert wrote out a Spanish interpretation with some planned changes for the sake of scoring and then reformulated the Spanish back to English after waiting a day. We propose to call the back interpretation “reformulation” because it reflects that we are evaluating a reformulation of the original English text after it was rendered into another language. After working with the original text, it was modified to allow for more specific types variants and rating recommendations.

Below is the original text in italic (blue), the reformulation is in normal font (green). The chart was the heuristic used to identify and evaluate differences between the original and reformulated version.

Scoring: After extensive discussion between using an analytical scale, anchored scale or both we opted for the anchored scale, in part because it would be easier to use already trained evaluators, and it would prevent issues arising from an intervening variable of a significant change in scoring between the current CHI™ tests and the EtoE exam.

We recommend five areas of evaluation: the three of the four areas already used by CCHI—Register, Lexical Content, and Grammar. The focus on Quality of Speech is more difficult to retain without additional verification. This will be discussed below. We recommend two new criteria—*Cohesion* and *Number of Ideas*.

Cohesion may seem to duplicate some aspects of grammar such as number/gender concordance as appropriate. However, it is also part of the strategies to maintain sense, not only within an utterance but between utterances by the same speaker and different speakers. These strategies include use of reference (deictics, anaphora, cataphora), conjunction (logical/spatial temporal), lexicon, ellipsis, and substitution. Ellipsis (leaving out) and substitution are not acceptable strategies in interpreting, but they occur in the source message. However, what is allowable in one language is not necessarily allowable in another. (Example of ellipsis: *I took 2 pills and have more*. With only this much text, we assume that *more* refers to the same type of pill. However, if the patient has been talking about all the medications prescribed, the two pills could refer to different medications and the *more* to even more types of meds. Ellipsis creates ambiguity rather than cohesion.)

In the example below, we show where there is a grammatical error by shifting from present to past tense. As a grammatical error it must receive a lower score. However, once used in the past tense, it remains in the past tense throughout the utterance maintaining temporal cohesion cognitively despite being grammatically incorrect. It will be important to avoid ambiguous referents for cohesion. For example, if an utterance includes both Metformin and Januvia, a subsequent utterance should not use “it” unless one of these has recently been specifically named.

Number of ideas is a count of the number of ideational units per utterance and so explicitly accounts for omissions and additions at a global level. Given that other languages will potentially restructure information, and that this is a cognitive not a linguistic test we consider it important to have an evaluative category above the phrase level. This is the same rationale for adding *cohesion* as an explicit evaluative category.

If this recommendation is adopted for the pilot, we recommend evaluations based on the four set categories used by CCHI, and two new ones to see if these significantly change outcomes and if so, in what direction. We recommend continuing to use the same point scale currently used by CCHI. Again, this is to maintain similarity between the tasks of specific language evaluation and the EtoE format. However, if during the pilot project the anchored scale proves to be insufficient to demonstrate the sought-after cognitive qualities of the candidates, an analytical scoring approach should be tried.

Since we worked from a written text we did not score for *quality of speech* (false start/ backtracking, pronunciation, rate of speech). In the languages for which an oral exam is offered, it is possible to score each candidate for both languages. In the EtoE format, though, native speakers of English will have an advantage.

The link between accent acquisition and cognition is still being investigated and is made more complex by factors including neurological plasticity, method of acquisition, auditory awareness (which may be as variable as perception of tonality), length and type of exposure to L2 (i.e., second, non-native language), etc. Therefore, in terms of scoring for these factors (if maintained), care must be taken not to give an advantage to candidates who are native speakers of English (including heritage speakers of other languages) while putting at a disadvantage those for whom English is a second language.

To sum up, the proposed scoring scales are:

- *Lexical content*
- *Register* (Have further discussions with SMEs: While maintaining register is ideal, it may be impossible in some languages. That is, when you convert certain terms into another language there may be only one register for that term. By the time candidates go back into English, they'd be maintaining the register of the non-English prompt and may end up lowering the register in English. This scale may work better for Spanish and other languages with good written representation of medical/healthcare knowledge than for other languages, i.e., this scale may work for this study since participants are interpreters of Spanish, Arabic and Mandarin, but may not work so well for interpreters of languages of lesser diffusion.)
- *Grammar*
- *Cohesion*
- *Number of ideas*
- *Quality of speech* (with caveat above)

Example's legend

The source message is always in italics. The chart (table) format was a heuristic to allow us to clearly see any variants and to make comments on how much or little core meaning was affected. Each row serves a different function. The source message is in italics with no background color, reformulation is in grey background and anchored scoring is in the row with a blue background. A colored font marks different types of variation. The reformulation row also shows our first impression of degree of change from the source and aspects we considered as we discussed how to evaluate the variation.

Scoring scale:

- 1 = Variation completely loses source meaning
- 2 = Variation strongly affects source meaning
- 3 = Variation affects source meaning but not significantly
- 4 = No variation or no effect on meaning.

Original dialog:

NP #1: *Because there is a difference between having a general malaise versus having the flu, before I can make any sort of accurate diagnosis, I will need a complete list of prodromes from you. [33 words in this utterance]*

Patient #1: *Well, at first, I thought I was just feeling travel fatigue, but I keep feeling light-headed, flushed, my skin crawls and I get the sensation of icepicks poking into my brain. [31 words in this utterance]*

NP #2: *Are you still having these symptoms now or have they gone away?*

Patient #2: *It comes and goes.*

Reformulation

NP #1: *Given that there is a difference between suffering from general malaise as opposed to having the flu I need you to give me a complete list of symptoms before I can make some kind of reliable diagnosis.*

Pt #1: Well, at first, I thought I was just experiencing tiredness after a trip, but I kept having dizziness, I felt flushed, my skin crawled, and I had the feeling that needles were stabbing my brain.

NP #2: You got problems?

PT #2: It arrives and departs.

Comparison of Nurse Practitioner utterance #1.

Score	Utterance	Comments
	1. <i>Because there is a difference between having a general malaise versus having the flu,</i>	
Reg 4 Lex 4 Gram 4 Coh 4 Ideas 4 Spch ql? Total 20	1. Because there is a difference between having general malaise as opposed to having the flu	one lexical change, no effect on register or meaning
	2. <i>before I can make any sort of accurate diagnosis,</i>	
Reg 4 Lex 3 Gram 4 Coh 4 Ideas 4 Spch ql? Total 19	3. before I can make some kind of reliable diagnosis	accurate/reliable slight difference
	3. <i>I will need a complete list of prodromes from you.</i>	
Reg 3 Lex 3 Gram 4 Coh 4 Ideas 4 Spch ql? Total 19	2. I need you to give me a complete list of symptoms	The tense change is debatable. Meaning is not affected but could make argument that register changed. Drop 1 Register point because lose the politeness of future tense for underlying command. Moved “you” from end to beginning. Also, can have effect of making command more direct. At the same time, how politeness is created may vary from language to language. So, it may be beneficial to avoid sentences like this one in the item. See below for “prodrome.”

Figure 1 Nurse Practitioner utterance #1: Source, reformulation, and scoring

Notes on scoring in Figure 1

Prodrome probably should not be used for the actual EtoE test, but we choose it for this example to raise the issue of how to score technical terminology. In the reformulation “symptom” was not counted as a change. While it could reflect a register change, it will have to be kept in mind that there may not be multiple levels of register for some lexical items in the candidate’s language. There may not even be a direct equivalent. We have to assume that if a way of referring to something has been negotiated between the interpreter and the speaker (patient or provider), the interpreter will render that back into English as close to the source message as possible.

Comparison of Patient utterance #1.

Score	Utterance	Comments
-------	-----------	----------

	1. Well, at first, I thought I was just <i>feeling travel fatigue</i>	
Reg 4 Lex 3 or 4? Gram 4 Coh 4 Ideas 4 Spch ql? Total 19	1. Well, at first, I thought I was just <i>experiencing tiredness after a trip</i>	shortness of breath, weakness, nausea, suffering – would be wrong. NB: concept of tiredness is analyzed, not anything else.—Lack of energy - would be okay. There are grammatical and lexical differences, but the meaning is preserved.
	2. <i>but I keep feeling light-headed, flushed,</i>	
Reg 4 Lex 3 Gram 2 Coh 3 Ideas 4 Spch ql? Total 16	2. but I <i>kept having</i> dizziness, I [<i>felt</i>] flushed,	First shift to past affects meaning, but subsequent maintenance of past tense shows internal cohesion—not constantly shifting tense so demonstrates cognitive sense of time once initial mistake is made. Restructuring of ideas from list of adjectives to list of verb phrases does not affect either number of ideas or meaning.
	3. <i>my skin crawls</i>	
Reg 4 Lex 4 Gram 2 Coh 3 Ideas 4 Spch ql? Total 17	3. my skin <i>crawled</i>	Again, shift from present to past, but internally cohesive.
	4. <i>and I get the sensation of icepicks poking into my brain.</i>	
Reg 4 Lex 4 Gram 2 Coh 3 Ideas 4 Spch ql? Total 17	4. and I <i>had</i> the <i>feeling</i> that <i>needles were stabbing</i> [omit] my brain.	Same issue with present to past; icepick may not be culturally salient so <i>something</i> for icepick which is pointy and sharp is acceptable.

Figure 2 Patient utterance #1: Source reformulation and scoring

Notes on scoring in Figure 2

This section shows issues relating to scoring for cohesion. As mentioned above, there are two aspects which can be considered in this type of exercise. Accurate rendition of the source text is certainly one, but as cohesion is an indicator of cognitive function, the question is, if a slip is made does the candidate maintain consistent, internal cohesion. In this case, the answer is yes.

The other point in this example is the acceptable range for lexical variation. Not all lexical challenges are related to technical terminology. Cultural artifacts such as *icepick*, *turkey baster*, *comal* (round, usually made of clay or some type of iron/steel, item like a griddle used for cooking tortillas), *suṛt* (square piece of cloth used for carrying babies or making bundles to be carried on the head), are equally challenging to render into a target language.

Reformulating these terms may demonstrate some variation. Additionally, range of meaning in the source language will not be equal in the target. A reformulation may produce a synonym rather than a repetition of the

exact word. For example, if a patient’s utterance is: “*I get my meds at Walgreens.*” The reformulation could be “*I **buy** my meds at Walgreens.*” This should not result in a lower score. (This last example also shows the importance of constructing very clear, unambiguous utterances.)

Comparison of Nurse Practitioner utterance #2.

Score	Utterance	Comments
	<i>NP: Are you still having these symptoms now or have they gone away?</i>	
Reg 1 Lex 1 Gram 1 Coh 1 Ideas 1 Spch ql? Total 5	You got problems?	Loss of duration, options, and specificity.

Figure 3 Nurse Practitioner utterance #2: Source, Reformulation and Scoring

Note on scoring in Figure 3

This utterance was constructed to show an example of a really bad reformulation. The entire weight of understanding this reformulation would fall on the LEP listener understanding the flow of a medical interview, knowing that duration of symptoms is a common question.

Comparison of Patient utterance #2.

Score	Utterance	Comments
	<i>Pt: It comes and goes</i>	
Reg 1 Lex 2-4?? Gram 3 Coh 1-4?? Ideas 2-4?? Spch ql? Total 9-16??	It arrives and departs.	Source is idiomatic which is lost; While pronoun “it” is correct, it lacks concordance with the plural symptoms from the NP’s question. The reformulation is stilted and awkward. It is an example of sticking literally to the original language (word for word conversion), but it is still easily understandable in English and covers the full meaning. Grammar: If you can say it comes and goes, (and we do) then surely you can say it arrives and departs: What is the ferry boat schedule? It arrives and departs every 2 hours.

Figure 4 Patient utterance #2: Source, Reformulation and Scoring

Notes on scoring Figure 4

This last utterance was an attempt to elicit a really low score. The reformulation is stilted and awkward indicating that it is overly literal although two motion verbs with close meaning were used. By definition, an idiom is a word or phrase that has a meaning beyond what the surface components mean. So, the cognitive task here is to provide the intended meaning of the source message. A reformulation like “Sometimes I have them” would score “4” all the way across even though the idiomatic expression is lost. A reformulation like: “They’re not persistent” would have a Register score of 3.

It should be noted that there is a lack of cohesion between the NP’s question “*Are you still having these symptoms now or have they gone away,*” and the patient’s response “*It comes and goes.*” However, the reformulation is faithful to the source message. This is an example of how important it is to construct a clear and cohesive prompt, so that any loss of cohesion in a candidate’s reformulation into English is truly and error and not a reflection of the poorly constructed prompt.

2. Paraphrase (Monolingual/intralingual reformulation) – Audio to Audio

Item description:

Candidates listen to an audio recording of a provider's or family member's utterance in English and record their paraphrasing (reformulation) of it in English.

Job tasks and interpreting skills measured:

Knowledge of:

- a) English terminology, idioms, usage, and cultural significance
- b) Structure and grammar of English
- c) The concept of equivalency of meaning
- d) The concept of register

Skill in:

- Retaining and recalling information in short-term memory
- Listening actively in English
- Anticipatory listening in English
- Communicating fluently in English
- Maintaining accuracy in English
- Sufficient mastery of English pronunciation/quality of speech to avoid impact on understanding
- Self-monitoring for accuracy in English
- Mental agility
- English comprehension depth & breadth
- English production depth & breadth

Item development recommendations:

- Include 8-10 items of this type on the test so that enough data can be gathered to assess this skill.
- Keep the length of each item the same as that of an utterance in a consecutive item on the CHI™ exam (up to 35 words per utterance).
- Each item may be played twice before candidate records it, i.e., exactly as with a consecutive item on the CHI™ examination.
- Use different healthcare specialty topics.
- Include at least one item on the same topic that is used in the bilingual reformulation item (the first one on the test).
- Create items of different types of sentences – statements (declarative), questions (interrogative), commands (imperative), and exclamations (exclamatory).
- Identify units of meaning that can be paraphrased.
- It is strongly recommended that the register of the speaker (i.e., of the prompt) be kept in a paraphrase. This requires a very thoughtful construction of the prompt so that it *is* possible to paraphrase it without changing the register. In many cases, changing the register is the method to arrive at an accurate paraphrase. It might be impossible to construct a prompt to meet the requirement of maintaining register. In this case, make sure to change the instructions to the candidate to allow register shifts.
- Evaluate the importance of the difference between a stand-alone word and its specific use in the context (e.g., the use of or absence of the definite article may be important to the meaning of the message).
- Create items of different paraphrasing types: word phrases, order of sentence parts, splitting sentences, etc.
 - Create four items focusing on the knowledge of a) English terminology, idioms, usage, and cultural significance; and d) The concept of register.

- Create four items focusing on the knowledge of b) Structure and grammar of English; and c) The concept of equivalency of meaning
- When creating an item, “try it out,” i.e., write down possible paraphrases (including incorrect ones), to see how much variation is possible. These model responses will help the raters create scoring parameters. Discard items that cannot be paraphrased or allow only for one plausible version.
- Candidates are allowed to take notes.

Example of an item:

You need to watch out for foods with high amounts of carbohydrates. The greater the number of carbs in your daily diet the more elevated your blood sugar level can get.

{8 units:

1. Monitor, be careful
2. High carb foods
3. The more you eat them the worse the consequences, i.e.:
4. increase
5. Presence/number of carbs
6. In daily intake
7. Possibility (can – not will/shall)
8. Blood sugar level}

Correct answer: Be careful with high carbohydrate foods. Consuming more carbs on a daily basis can increase your blood sugar level.

Instructions to candidates:

Listen to the speaker’s message and re-state (paraphrase) it in English using your own words (using synonyms, changing sentence structure, etc.). Your goal is to keep the same meaning of the whole message as much as possible. Do not just repeat what you heard word for word, you must use different English words and/or a different word order to convey the same meaning of the message. Make sure not to omit or add any information (units of meaning) and *keep the register of the speaker as much as possible*. {See above and adjust if the SMEs’ decision is to allow register shifts.}

Scoring recommendations:

- Decide whether raters will use a combination of analytic and anchored scales or just one type of scoring for this item.
- If using anchored scales – decide if the ones used on the CHI™ exam are appropriate: lexical Content, Register, Grammar, and Quality of Speech.
- Decide how to score a response when the candidate simply repeats everything verbatim, i.e., without paraphrasing.
- Decide how to score if a candidate simply summarizes everything in one sentence and as a result loses units of meaning. Consider adding a scale for “Following Instructions.”
- If they partially paraphrase - 30-40% - what score?
- How is keeping or not keeping the prompt’s register scored?
- Include a scale for Quality of Speech – unintelligible.
- Include raters rating not-their-own-language: a Spanish rater does not rate a Spanish candidate; diversity in raters’ English as native language

3. Shadowing – Audio to Audio

Item description:

Candidates listen to an audio recording of a provider's or patient family member's speech and repeat it in English simultaneously.

Job tasks and interpreting skills measured:

Skill in:

- Listening, processing and producing an English output simultaneously
- Retaining and recalling information in short-term memory
- Listening actively in English
- Anticipatory listening in English
- Communicating fluently in English
- Maintaining accuracy in English
- Sufficient mastery of English pronunciation/quality of speech to avoid impact on understanding
- Self-monitoring for accuracy in English

Item development recommendations:

- Because shadowing is an artifice and not an actual interpreting skill, the recommendation is that it be placed in the middle rather than at the beginning of the test; definitely do not make it item #1 or #2. This item could catch candidates by surprise and throw them off. Consider placing it towards the end of the test.
- Use only one Shadowing item on the test form.
- Develop an English script of a provider's or patient family member's speech following the same parameters as for the simultaneous interpreting item on the CHI™ examination: 170-220 words; 4-6 terms; 2-4 colloquial or idiomatic expressions; 82-90 seconds of audio prompt; recorded at 120-150 words per minute – somewhat in the middle of this range.
- Use a healthcare topic different from that of the bilingual reformulation item (the first one on the exam).

Example of an item:

See your doctor if you think you might be experiencing signs or symptoms of heart failure. Seek emergency treatment if you experience any of the following:

- Chest pain,
- Fainting or severe weakness,
- Rapid or irregular heartbeat associated with shortness of breath, chest pain or fainting,
- Sudden, severe shortness of breath and coughing up pink, foamy mucus.

Although these signs and symptoms may be due to heart failure, there are many other possible causes, including other life-threatening heart and lung conditions. Don't try to diagnose yourself. Call 911 or your local emergency number for immediate help. Emergency room doctors will try to stabilize your condition and determine if your symptoms are due to heart failure or something else. If you have a diagnosis of heart failure and if any of the symptoms suddenly become worse or you develop a new sign or symptom, it may mean that existing heart failure is getting worse or not responding to treatment. This may also be the case if you gain 5 pounds (2.3 kg) or more within a few days. Contact your doctor promptly.

Instructions to candidates:

This task is similar to simultaneous interpreting, except you will be doing it in English only. Listen to the English recording and start repeating what you hear in English simultaneously. You must start repeating within the first 10 seconds of starting to play the audio. Try to repeat everything exactly as you hear it (verbatim), without omitting, adding or changing any words.

Scoring recommendations:

- Decide whether raters will use a combination of analytic and anchored scales or just one type of scoring for this item.
- If candidate paraphrases accurately instead of repeating verbatim - do *not* penalize for this as long as all the meaning is kept because the purpose is to test ability to keep the meaning (by managing concentration, memory, split attention, speech production). However, in case of paraphrase, raters should truly discern how accurately the meaning is preserved by the candidate. Paraphrase instead of verbatim repetition may be indicative either of a highly skilled interpreter (i.e., beyond entry level) or of a candidate with poor concentration and memory skills. The latter would invariably change, omit, or add meaning. I.e., from a practical perspective, most likely paraphrased prompt will be penalized due to it being inaccurate.

4. Memory Capacity – Audio to Audio

Item description:

Candidates listen to an audio recording of a provider's or family member's utterance in English once, and repeat it exactly in English (verbatim) in a consecutive mode. The test form includes several items with incremental difficulty, i.e., longer and more complex sentences.

Job tasks and interpreting skills measured:

Skill in:

- Retaining and recalling information in short-term memory
- Sufficient mastery of English pronunciation/quality of speech to avoid impact on understanding
- Self-monitoring for accuracy in English

Item development recommendations:

- Include 8-10 items of this type on the test so that enough data can be gathered to assess this skill.
- Each item may be played only once before candidates records it, i.e., different from a consecutive item on the CHI™ examination because this is an assessment of candidate's memory, not interpreting.
- Candidates are not allowed to take notes. (See how this can be monitored by proctor?)
- Use different healthcare specialty topics.
- Include at least one item on the same topic that is used in the bilingual reformulation item (the first one on the test).
- Consider making items part of one or more dialogs so that the interpreter can use overall context of the conversation as a retention tool. While this approach may be closer to actual interpreting, it will not measure "pure" short-term memory capacity. Maybe, include 3-4 items contextualized as a dialogue, and 4 stand-alone items.
- Create items of different types of sentences – statements (declarative), questions (interrogative), commands (imperative), and exclamations (exclamatory).
- Create items of various lengths and sentence complexity:
 - First 2 items – about 12-15 words
 - 2 items – 16-25 words
 - 3 items 26-35 words
 - 1-2 items – 36-45 words
- Items should reflect a variety of topics, some of which should be "advanced" and from a wide range of medical areas. It is easier to remember information with which you are already familiar because you have the framework or grounding already established. If there are items like the ones here but some from specialties and sub-specialties, a correlation analysis could be added to develop a truer picture of

memory. We could add a question after each example like: “I have interpreted this material very often, occasionally, rarely, never”, which would allow weighting of the score on the basis of familiarity.

Example of an item:

Example 1. (12 words)

The onset of signs and symptoms of ear infection is usually rapid.

Example 2. (25 words)

You may want to talk to your doctor about osteoporosis if you went through early menopause or if either of your parents had hip fractures.

Example 3. (35 words)

Adenoids are two small pads of tissue high in the back of the nose believed to play a role in immune system activity. This function may make them particularly vulnerable to infection, inflammation and swelling.

Example 4. (45 words)

The symptoms of pneumonia vary from mild to severe, depending on factors such as the type of germ causing the infection, and your age and overall health. Mild signs and symptoms often are similar to those of a cold or flu, but they last longer.

Instructions to candidates:

In this section you will repeat in English every word you hear in English. The purpose is to assess your short-term memory capacity. Try to repeat everything exactly as you heard it, without omitting, adding or changing any words. Listen to the English recording. When the speaker is finished speaking, click the “Record” button and repeat what you heard in English, consecutively.

Scoring recommendations:

- Decide if raters will use a combination of analytic and anchored scales or just one type of scoring for this item.
- Decide how to score if candidate paraphrases accurately instead of repeating verbatim:
Option A: We *should* penalize for this because the purpose is to test memory capacity not accuracy.
Option B: We should *not* penalize when meaning/tone/intent are not lost by sentence reformulation or use of synonyms. However, omissions/additions/changes of meaning *should* be penalized.
Option C: An exact replica of the prompt is scored as correct (full score) but partial score (decide how much) is given for getting the same idea across.
- Consider adding a scale for “Following Instructions.”

5. “Cloze” – Text to Audio

Item description:

Candidates read a text in English with a gap in it and record in English the word or phrase that fills that gap and makes the utterance complete and logical.

Definition: A *Cloze Test* (also called the “cloze deletion test”) is an exercise, test, or assessment consisting of a portion of text with certain words (usually every 5th word; the higher the number the easier the test) removed (cloze text), where the participant is asked to restore the missing words. Cloze tests require candidates to understand context and vocabulary to identify the correct words that belong in the deleted passages of a text. The Cloze Test is commonly administered for the assessment of native and second language learning, especially

reading comprehension. Wilson L. Taylor¹⁷ introduced the term "cloze procedure" in 1953 and thoroughly researched the value of closure tasks as predictors of reading comprehension. Basic to the procedure is the idea of closure wherein the reader must use the surrounding context to restore omitted words. Comprehension of the total unit and its available parts (including the emerging cloze write-ins) is essential to the task.

This item, though, is intended to measure interpreter's comprehension and anticipatory reading ability as subskills of interpreting. The item is an adaptation of the "true" cloze test, with the focus on semantic and syntactic elements. The item's "gap" can be more than one word.

Job tasks and interpreting skills measured:

Knowledge of:

- English terminology, idioms, usage, and cultural significance
- Structure and grammar of English

Skill in:

- Reading and comprehending written text in English
- Anticipatory reading in English
- Self-monitoring for accuracy in English

Item development recommendations:

- Include 2-3 items of this type on the test.
- Develop an English script (an utterance by a provider or patient's family member) or select an excerpt from a document typical for healthcare settings following the same parameters as for the sight translation item on the CHI™ examination: 41-45 words; 3-4 terms; if not a document – also 1-3 colloquial or idiomatic expressions.
- It is important to give enough context prior to the gaps to fill in, so the candidates have a reasonable chance to know what is missing.
- Use different healthcare specialty topics.
- When creating a script, include different types of sentences that have the gap – statements (declarative), questions (interrogative), commands (imperative), and exclamations (exclamatory).
- Create items of various length and sentence complexity:
 - First two items – the gap to be filled consists of 1-3 words (including articles, particles, and prepositions)
 - Remaining items – the gap to be filled consists of 4-8 words (including articles and prepositions)

Example of an item:

Example 1.

Prompt:

Unmanaged diabetes can lead to uncontrolled ...[blank]... levels which can damage the body's organs, including the kidneys.

Correct answer:

blood sugar; glucose

Scoring comment: any other words would be 0 points; blood sugars = 1 point because of the grammatical error – plural.

Example 2.

¹⁷ Taylor, W.L. (1953) "Cloze Procedure": A New Tool for Measuring Readability. – In: Journalism & Mass Communication Quarterly. Volume: 30 issue: 4, page(s): 415-433.

Prompt:

When the thyroid produces too much hormone, the body uses energy faster than it should. When the thyroid doesn't produce enough hormone, ...[blank]... it should.

Correct answer:

the body uses energy slower than

Scoring comment: energy is used slower than = 1 point because the syntactic structure does not match the end of the prompt sentence, but the key meaning is preserved.

Instructions to candidates:

Read the following text with one or more missing words in it; the missing part is marked as "...[blank]...". Think which word(s) or phrase would make the text logical and complete and record that word(s) or phrase. Record just the missing word or words, do not read the whole sentence or text.

Scoring recommendations:

- Decide whether raters will use a combination of analytic and anchored scales or just one type of scoring for this item.
- Traditionally, teachers award one point per each missing word. A percentage of the correct answers is determined by dividing the number of points by the number of blanks. A common scale for determining reading comprehension is:
 - 61% or more correct replacements – independent reading level
 - 41-60% correct replacements – instructional level
 - Less than 40% correct replacements – frustration level
- Scoring should not focus on the number of words produced by the candidate but rather on the completeness of meaning and correct grammar.
- Scoring could be analytic:
 - 0= incorrect meaning
 - 1=close meaning but not quite correct or correct meaning but incorrect grammar which does not affect the meaning
 - 2=correct meaning and grammar (if applicable to the specific item)
- Another approach would be to use two scales – Lexical Meaning and Grammar.

6. "Finish the sentence" – Audio to Audio

Item description:

Candidates listen to an audio recording of an unfinished utterance in English and finish it with a logical ending.

Note: Because this is an audio-input item, a Cloze Test concept is adapted to require candidates to finish a sentence.

Job tasks and interpreting skills measured:

Knowledge of:

- English terminology, idioms, usage, and cultural significance
- Structure and grammar of English

Skill in:

- Listening and comprehending oral speech in English
- Anticipatory listening in English
- Retaining and recalling information in short-term memory
- Sufficient mastery of English pronunciation/quality of speech to avoid impact on understanding

- Self-monitoring for accuracy in English

Item development recommendations:

- Include 2-3 items of this type on the test.
- Develop an English script of an utterance by a provider or patient’s family member following the same parameters as for the consecutive item on the CHI™ examination: up to 35 words; 1-2 terms; one colloquial or idiomatic expression.
- It is important to give enough context prior to the ending, so candidates have a reasonable chance to identify what is missing.
- Construct an item in such a way that the missing part is fairly straight-forward as far as understanding the speaker’s intended (unspoken) meaning.
- Decide if the item should be played once or twice (adjust *Instructions to candidate* accordingly).
- Candidates are allowed to take notes.
- Use different healthcare specialty topics.
- Include at least one item on the same topic that is used in the bilingual reformulation item (the first one on the test).
- When creating a script, include different types of sentences that have the ending missing – statements (declarative), questions (interrogative), commands (imperative), and exclamations (exclamatory).
- Create items of various lengths and sentence complexity:
 - First item – the missing ending consists of 1-3 words (including articles, particles, and prepositions)
 - Remaining items – the missing ending consists of 4-8 words (including articles and prepositions)

Example of an item:

Prompt:

Chickenpox is a common illness caused by a virus called varicella zoster. People often get the virus as young children if they have not...

Correct answers:

been vaccinated against it;
 received a vaccine against it;
 had chickenpox (it) before;
 had it (chickenpox) before or have not had (received) a vaccine against it

Instructions to candidates:

Listen to an unfinished utterance by a provider or patient family member and record how you think it is best to finish so that it sounds complete and logical. Record just the missing ending, do not repeat the whole sentence.

Scoring recommendations:

- See comments for item type #5.

7. Synonyms – Multiple-choice item

Item description:

Candidates read an English sentence which contains a key high-register medical term and select, out of four options, the option that has the closest meaning.

Job tasks and interpreting skills measured:

Knowledge of:

- English terminology, idioms, usage, and cultural significance
- Structure and grammar of English
- The concept of register

Skill in:

- Analytical reading in English
- Evaluating equivalency of meaning in English

Item development recommendations:

- Include 2-3 items of this type on the test.
- Keep the length of each prompt no longer than that of an utterance in a consecutive item on the CHI™ exam (up to 35 words per utterance).
- Include one key high-register term of entry-level difficulty.
- The prompt can be either an utterance by a provider or patient's family member's speech or an excerpt from a healthcare document.
- The four options are various *paraphrases* of the prompt; some may use a synonym of a different register, others – explicitation without change or loss of meaning, the absolutely wrong option will have a change or loss of meaning.
- Use different healthcare specialty topics.
- Make sure NOT to use any terms that are used in the bilingual reformulation item (the first one on the test).

Example of an item:

Prompt:

If you have congestive heart failure, your outlook depends on the cause and the severity, your overall health, and other factors, such as your age.

Options:

a) If your heart cannot pump blood as well as it should... {correct, best answer, 2 points}

b) If you have a heart attack... {wrong, 0 points}

c) If you have a heart defect from birth... {wrong, 0 points}

d) If your heart stops pumping blood... {not quite correct – this may happen in acute heart failure but not in chronic heart failure, 1 point}

Instructions to candidates:

Read the statement carefully. Then select a synonym of the underlined medical term. Choose the option that has the closest meaning.

Scoring recommendations:

- It is recommended to develop a weighted scoring rubric to reward for better choices. For example, 2 points – best correct, 1 point – somewhat correct or correct in some cases but not in this specific context, 0 points – incorrect.

8. Equivalence of medical terminology – Text to Audio

Item description:

Candidates read an English sentence or short passage which contains a key medical term of entry-level difficulty, yet relatively complex, and record how they would re-state (paraphrase) it, possibly using neutral- or lower-register words and terms, in English.

Job tasks and interpreting skills measured:

Knowledge of the concept of register can be assessed.

Knowledge of:

- English terminology, idioms, usage, and cultural significance
- Structure and grammar of English
- The concept of register

Skill in:

- Analytical reading in English
- Evaluating equivalency of meaning in English

Item development recommendations:

- Include 2-3 items of this type on the test.
- Keep the length of each item the same as that of an utterance in a consecutive item on the CHI™ exam (up to 35 words per utterance). Could be one sentence.
- Include 2-5 key high-register terms, of the entry-level difficulty.
- The text can be either a script of a provider's or patient family member's speech or an excerpt from a healthcare document.
- Use different healthcare specialty topics.
- Make sure NOT to use any terms that are used in the bilingual reformulation item (the first one on the test).
- More focus on paraphrase of syntax than simply words.

Example of an item:

Prompt 1: [Microscopic particles in ambient smoke may cross the pulmonary circulatory barrier.](#)

Comment to SMEs: This item is too hard for an entry-level interpreter.

Prompt 2: [Many elderly patients present with circulatory problems. It is really important for them to manage hypertension carefully as it could lead to stroke, renal failure and even myocardial infarction.](#)

Answer: Lots of older patients have blood circulation issues. High blood pressure may cause stroke, kidney failure and even heart attack. This is why it must be carefully handled.

Instructions to candidates:

Read the English passage and record how you would say it in English using other words (i.e., re-state/paraphrase it), keeping the meaning of the original as accurate and complete as possible. Do not repeat the medical term, instead use other words to describe its meaning. You may use neutral or lower register words if needed.

Scoring recommendations:

- Decide whether raters will use a combination of analytic and anchored scales or just one type of scoring for this item.
- Decide how to score if candidates make English grammar or quality of speech errors.

- Possibly develop a weighted scoring rubric to reward for better choices.
- Decide what score is given if the key medical term is repeated without re-statement/explicitation. Consider adding a scale for “Following Instructions.”
- If paraphrase changes the meaning – penalize or lower score
- If vulgar or offensive – inappropriate change, penalize or lower score
- Include a scale for Quality of Speech – e.g., “unintelligible”

9. Reading comprehension and production of speech in English – Text to Audio

Item description:

Candidates read an English text and an open-ended question based on it and record their answer in English.

Job tasks and interpreting skills measured:

Knowledge of:

- English terminology, idioms, usage, and cultural significance
- Structure and grammar of English

Skill in:

- Reading and comprehending written text in English
- Communicating fluently in English
- Sufficient mastery of English pronunciation/quality of speech to avoid impact on understanding
- Self-monitoring for accuracy in English

Item development recommendations:

- Include one item on the test.
- Text 170-220 words.
- Allow up to two minutes for candidates’ recorded response.
- Decide on the number of questions posed: 1-3. If more than one question is included, develop a scoring rubric for “number of questions answered.”
- Consider placing the item towards the end of the form so that the time the candidate has to read the text is managed by the candidate. Or consider limiting the time the candidate has to read the text – in this case make sure the allotted time gives reasonable opportunity for success to every candidate.
- Use a healthcare topic different than that of:
 - the Bilingual Reformulation item (the first one on the exam),
 - the Shadowing item (#3),
 - the Listening Comprehension item (#10).
- Do NOT use topics or questions related to interpreter’s role, code of ethics or standards.
- Use health- or healthcare-related content. Avoid emotive content.
- To assess the text complexity, consider using the *Flesch-Kincaid Grade Level* and establish the complexity threshold.
- Think of this item as a typical English-language class exercise. Its intent is to determine whether reading comprehension skills of a text written at a “professional” level of difficulty correlate to interpreting skills. This is not a sight translation item.

Note: The ability to read and analyze information is a goal all interpreters should aspire to. At the same time, there are other ways to learn in addition to reading. This is especially important for candidates from “less common” languages. Some of those languages are spoken in areas where there is less formal written schooling. (There is also a theory that dyslexic people (20% of the population) may be drawn to interpreting because it is predominantly oral, which is their strength.) With this caveat in mind, it may be more relevant and fairer to include on the test only texts that interpreters would be expected to sight

translate (or interpret explanation of) on the job rather than texts for professional development or continuing education.

- Use complex logical structure or complex, yet common, healthcare concept in the prompt. For example, co-insurance or deductible, discharge instructions with multi-step care, multi-step prep instructions (e.g., for colonoscopy).
- The question should be asked in such a way that it would require candidates to show their understanding of the text and give candidates an opportunity to display their spontaneous English speech production. Avoid asking questions that may be misunderstood in some cultures, that test medical knowledge, that are beyond entry-level experience. Keep in mind that candidates may come from cultures where doctor's orders are not questioned. Their answer may be simply, "No, because the doctor knows best." Make sure to create questions that elicit an answer.

Example of an item:

Example 1: Text typical for sight translation

Prompt:

During the test you may hear the tech turn the machine on and off several times. Some patients worry that either the machine is malfunctioning, or the test has found something wrong with the patient. Please do not worry! Restarting the machine is a completely normal part of the procedure.

Questions:

- 1) What is the main idea of this text?
- 2) Why do some patients worry? What are their concerns? (And are these concerns warranted?)

Answer:

1) Variant A: The goal is to calm and reassure patients, so they do not worry.

Variant B: The communication goal of the author is to forestall anxiety, allay concerns, and reassure patients.

2) When they hear the machine turn on and off, they think it is broken or the technician has found something wrong in their pictures.

Comments for scoring the sample answer:

- Key medical information of the prompt must be correct.
- Variant B covers same points using higher register vocabulary. Should higher register get extra points? No. Both answers are correct. This speaks to the question of what if the candidate only speaks for one out of two minutes—if they include all information (more efficiently) they should not be penalized.

Example 2: reading and discussion (continuing education) type of text

Prompt:

One of many causes of congestive heart failure is high blood pressure. If your heart is too stiff to pump the blood as well as it should, the blood flow slows down. This can lead to a decrease in kidney function. A no-added-salt diet may help keep your blood pressure under control. The more salt in the body the harder it is for the kidneys to excrete water. As less water is removed via the kidneys, water volume in the blood increases raising blood pressure. The heart has more difficulty pumping against higher pressure. Fluid begins seeping out of the blood vessels into the surrounding tissues leading to swelling of the feet and ankles. When fluid accumulates in the lungs there is less room for air and patients feel short of breath. Water build up puts additional pressure on the kidneys further decreasing the urine output.

NOTE to SMEs: The text contains quite complex information and, therefore, might be too difficult for the exam as it requires candidates to grasp the logic and info quickly. At the same time, the Flesch-Kincaid Grade Level is 6.9 (which is not too high).

Questions:

- 1) Is it important for patients to understand the reason why the doctors recommend what they recommend? Why or Why not?
- 2) Explain how salt intake & blood pressure & kidneys are interrelated.
- 3) Explain how congestive heart failure leads to swollen ankles.

Answer:

Patients need to know because they will probably follow direction more if they understand why it matters. (or express and support an alternate opinion with a reason that supports the statement.)

--An understanding of the self-exacerbating nature of the cycle. (a vicious circle)

--more salt intake raises blood pressure, by decreasing how well kidneys excrete excess water.

--Meanwhile, higher blood pressure decreases kidney function,

--and makes it harder for the heart to pump blood

--so, fluid backfills into ankles and lungs and stomach.

Instructions to candidates:

Read the text carefully. Read the question(s) following the text. Record your detailed answer in English. Demonstrate how you understand the text and express your own opinion on the topic. Keep in mind that you have two (2) minutes to record your complete answer. Demonstrate your command of English to the best of your ability.

Scoring recommendations:

- Decide whether raters will use a combination of analytic and anchored scales or just one type of scoring for this item. It is recommended to use a rubric similar to the one for scoring AP[®] Spanish exams that accounts for task completion, accuracy and completeness, and language use. (See *Appendix C*) The rubric describes accuracy and completeness as behavioral anchors rather than in percentages, which might be more user-friendly and intuitive for raters. Consider adapting further and replacing subjective qualifiers (“very good”) with quantifiable (e.g., no more than two critical errors OR less than 10% of the message was obstructed) or more specific anchors.
- The examples above include points the answers should include. It will be necessary to decide for the specific test item: 1) what the points are; 2) how many points should be/can be included in a one-minute prompt; 3) how many points should be included for top marks, middle range, and so on.
- Decide how to score lexical complexity (breadth of English vocabulary usage), grammar and quality of speech. See the proposed rubric below; Level 3 or above is adequate to pass. Grammar must be correct; vocabulary should be used to mean what it is supposed to mean.
NOTE to SMEs: Think about candidates’ ability to do the job: clear, simple speech should be adequate to succeed. Additional ability (grammatical complexity and finesse, large vocabulary) is very nice to have, but not critical.
- If more than one question is included, develop a scoring rubric for “number of questions answered.”
- Decide how to score if the candidates spoke for one minute or less (i.e., 50% of the allotted time). The number of words is not relevant. What matters is conveying the key ideas in two minutes or less. In fact, covering everything in one minute is/may be a bonus.
- Decide how to score if the candidates spoke for two minutes but did not express their opinion or did not show that they understood the text. What matters is conveying the key ideas. NO ideas = no points.
- What is the correct reading comprehension answer? The ‘best’ answer includes all the main points, and the speech is clear, easy to understand, and grammatically correct.
- Consider candidates’ ability of self-monitoring for accuracy in English as a speech production parameter.

Example of a rubric for self-monitoring (Determine the value of x = the number of self-corrections that does not distract the listener):

# of self-corrections	# of points added to the score	Reasoning
0-1(or 2?)	3 points	because it is best to need no corrections
1(or2) -x	2 points	because it is better to correct than leave an error
More than x	1 point	because too many self-corrections are distracting and annoying to the listener
	0 points	Consider whether adding "0" points is helpful

10. Listening comprehension and production of speech in English – Audio to Audio

Item description:

Candidates listen to an audio recording of a provider's speech in English and record their summary of it.

Job tasks and interpreting skills measured:

Knowledge of:

- English terminology, idioms, usage, and cultural significance
- Structure and grammar of English

Skill in:

- Listening actively in English
- Identifying key points in an oral speech in English
- Summarizing key points of an oral speech in English
- Communicating fluently in English
- Sufficient mastery of English pronunciation/quality of speech to avoid impact on understanding
- Self-monitoring for accuracy in English

Item development recommendations:

- Include one item on the test.
- Allow candidates to take notes for this item.
- Develop an English script of a provider's speech following the same parameters as for the simultaneous interpreting item on the CHI™ examination: 170-220 words; 4-6 terms; 2-4 colloquial or idiomatic expressions; 82-90 seconds of audio prompt; recorded at 120-150 words per minute – somewhat in the middle of this range.
- Allow two minutes for candidates' recorded response.
- Use a healthcare topic different than that of:
 - the bilingual reformulation item (the first one on the exam),
 - the Shadowing item (#3),
 - the Reading Comprehension item (#9).
- Use complex logical structure or complex, yet common, healthcare concept in the prompt.
- Identify how many key points are in the prompt that candidates must convey in their summary. Include the number of key points that would make scoring easier (see the 1st bullet in the Scoring Recommendations).
- Consider using a video input rather than an audio; although, it will increase the cost of the test production.

Example of an item:

{Based on the dialog from the 90-second YouTube video “wrist catheterization” at <https://www.youtube.com/watch?v=RRngelW3k3s>}

Scenario: MD speaks to the patient {Provide this context before candidates listen to the audio prompt.}

Prompt:

I understand you are concerned that you may have something going on with your heart. You do have multiple risk factors that put you at higher risk for heart disease including high blood pressure, diabetes, excess weight. When we have high risk patients who we want to make sure don't go on to have a heart attack later, we bring them in to do a more definitive test. The test in the office did not show any clear blockages but the function of the heart was a little weaker than normal, so what we'll do, we will try to do your angiogram through your arm, if that does not work, and it doesn't always work through the arm for some people, we will go in through the artery in the groin.

Answer:

Include the following key points in any order:

pt is worried about heart.

pt has multiple risk factors.

risk factors put her at more risk for heart disease.

risk factors include HBP, diabetes, excess weight.

Doctors want to prevent heart attacks,

so, if patients are high risk

doctors do more (definitive) tests in the hospital.

This patient had a test in the office (did not specifically show blocks)

but did show the heart was weaker (than normal).

The (catheter for the) angiogram will go in at the wrist,

if possible and if not,

through the groin.

Comments to raters scoring this answer:

This example has 12 ideas. (This could be argued.) Some ideas may be more important than others to remember, depending on the context. (The patient already knows s/he is worried about his/her heart. So that may be less important to recall in this case)

If we keep this example, we should

- decide how central to the picture each part is, and then
- decide how many parts the candidates should reasonably be able to remember,
- decide how many points to give for each part.

The items in the prompt are not in chronological order, which makes it more challenging. Yet, it is still acceptable as real speech is not always in chronological order, either.

Instructions to candidates {Because this task is different than most candidates' expectation of an interpreting test, it is strongly recommended to offer a really thorough Practice Exam which will demonstrate the expectations. Otherwise, we may get an almost word for word recreation of the original text};

Listen to the provider's speech in English and state in your own words all the key points the speaker makes. Do not repeat the whole speech word-for-word, because this is not an interpreting task.

Scoring recommendations:

- Decide whether raters will use a combination of analytic and anchored scales or just one type of scoring for this item. It is recommended to use the modified AP® Spanish rubric (see item type #9, *Appendix C*).

- Decide on the weighted scoring rubric, e.g., 3 – all key points mentioned, 2 – at least 75% of key points mentioned, 1 – 74-50% key points mentioned, 0- less than 50% of key points mentioned. decide how many points to give for each score. Consider the point about the scoring rubric above.
- Count the number of ideas in the text, decide whether some are more important in context than others.
- Decide how to score lexical complexity (breadth of English vocabulary usage), grammar and quality of speech. See the proposed rubric below; Level 3 or above is adequate to pass. Grammar must be correct; vocabulary should be used to mean what it is supposed to mean.

NOTE to SMEs: Think about candidates’ ability to do the job: clear, simple speech should be adequate to succeed. Additional ability (grammatical complexity and finesse, large vocabulary) is very nice to have, but not critical.

- Consider adding a scale for “Following Instructions” to score candidates who repeated the prompt word-for-word instead of summarizing it.
- Consider candidates’ ability of self-monitoring for accuracy in English as a speech production parameter. Same as in item type #9:

Example of a rubric for self-monitoring (Determine the value of x = the number of self-corrections that does not distract the listener):

# of self-corrections	# of points added to the score	Reasoning
0-1(or 2?)	3 points	because it is best to need no corrections
1(or2) -x	2 points	because it is better to correct than leave an error
More than x	1 point	because too many self-corrections are distracting and annoying to the listener
	0 points	Consider whether adding “0” points is helpful

Item Content (Scripts and Texts)

It is important to maintain the same level of difficulty and diversity of the speeches and texts in the EtoE exam as are present in the CHI™ performance exams, i.e., appropriate for the entry-level interpreter.

However, to test some cognitive skills, it may be beneficial to use speeches/texts with confusing or complicated logic, speaker’s backtracking, tangential comments, etc., in order to test certain cognitive skills. Candidates on the EtoE exam are not actually required to interpret such speeches/texts. Therefore, it is important to develop a scoring scale to assess how well candidates preserve the core logic and concepts of the speeches and texts.

The content of the *audio* English prompts should represent the speech typical of healthcare providers and of patient’s English-speaking family members.

The content of the *text* English prompts should represent the documents typical for various healthcare settings.

It may be beneficial to include texts/recordings intended to test understanding of scientific (e.g., biomedical) concepts.

Scoring Recommendations

Ideally, scoring of the EtoE examination should be similar to the scoring scales and procedures of the dual-language CHI™ performance exam.

Therefore, the following elements should be the same as in scoring of the current CHI™ performance exams:

- All raters are interpreters with a minimum of 5-years' experience in healthcare interpreting.
- All raters undergo a similar training.
- Every audio response is scored by two raters independently of one another.
- Raters do not see each other's scores nor the total score of a candidate.

However, the actual rating scales may need to be adapted since the EtoE test items are different than those of the dual-language CHI™ performance exam. For example, the scales of Speech cohesion/Fluidity and Number of ideas" may be considered for reformulation or speech production items.

It is recommended for Listening and Reading Comprehension items to use a rubric similar to the one used for scoring AP® Spanish exams that accounts for task completion, accuracy and completeness, and language use. Consider adapting further and replacing subjective qualifiers ("very good") with quantifiable (e.g., no more than 2 critical errors OR less than 10% of the message was obstructed) or more specific anchors.

Further review of the existing language proficiency scales (e.g., ILR, ACTFL, CEFR, IELTS, TOEFL) might provide guidance for defining the EtoE examination scales.

The National Task Force panelists agree that it is important to develop clear scoring parameters, especially for assessing EtoE reformulation accuracy, with special attention to circumlocution and explicitation related to how different languages treat ambiguity.

For the bilingual reformulation item, it is important to include a step when raters listen to the candidates' non-English recording and assess insertion of any English words, with a corresponding penalty.

Special consideration should be given to ensure fair assessment of candidates of languages without a written or Western biomedicine tradition where during actual interpreting very little word-level transcoding is possible.

The National Task Force panelists recommend collecting data about the study participants' native language and second language acquisition and utilizing this data to analyze whether native-English speakers have higher scores of passing the EtoE exam. If such correlation is ascertained, it would be important to account for this unintended advantage either via adjusting eligibility requirements or other means.

Test Taker Questionnaire and Preparation Guide

It is critical to clearly define the purpose of such an exam and its limitations, as well as educate all stakeholders about its value. The EtoE exam is not the same as a dual-language CHI™ performance exam, but it is a step closer to the latter than a written knowledge exam alone.

Because many tasks on this examination are different from most candidates' expectation of an interpreting test and because many candidates without formal education may have never been exposed to such tasks, it is strongly recommended to offer a thorough Practice Exam which will demonstrate the expectations of the candidate and provide examples of model responses.

It is recommended to ask EtoE test takers (in addition to CCHI's regular eligibility determination process) to complete a questionnaire that will allow for taking into account candidates' educational background and interpreting experience at a more granular level.

Conclusion

CCHI Commissioners are grateful for all the ideas and suggestions the panelist shared during the in-depth discussions. The next step in this project is to design and conduct a concurrent validity study to determine if the EtoE examination (or any of its specific types of test items), in fact, measures interpreting skills and abilities. The study will involve candidates of the Arabic, Mandarin and Spanish languages who will take both the EtoE exam and the corresponding CHI™ dual-language performance exam.

If the study results conclusively prove that scores on the English output portion of the EtoE examination are not significantly different from the scores on the dual-language CHI™ exam, then the EtoE examination will be developed in accordance with the NCCA accreditation standards, for the purpose of enhancing the CoreCHI™ certification.

If the study results are inconclusive or negative, the profession and educators will have evidence-based proof that only dual-language interpreter performance examination can reliably assess interpreting skills.

Regardless of the results, the study will benefit interpreter educators and employment recruiters providing them evidence-based information for training and job preparedness of interpreters of less common languages.

Appendix B. Performance Item Template for the EtoE Examination

SME's name: _____

Item Number (will be assigned by CCHI): _____

Type:

Please put an X or underline the appropriate classifications for each of the following considerations.

Key Variables:

Input method: __Audio __Text

Output method: __Audio __Text

Type of discourse: __Dialog __Monolog __Utterances
 __Text: Phrase(s) __Text: Cohesive Passage

Healthcare Provider: __Physician __Nurse __Allied Health __Laboratory __EMT Other: _____

Patient Type: __Patient __Family Member

Situation: __Emergency __Non-Emergency

Medical Condition:

__Cardiology __Dentistry __Endocrinology __Family Medicine/General Practice
__Imaging __Internal Medicine __Mental Health __Pediatrics
__OB/GYN __Oncology __Orthopedics __Respiratory

__Other (specify): _____

Your final submission must include (See Meeting #1 Handouts):

1. The *script* of the item in English = what the candidate will hear or see on the test.
2. Three *model responses* to the item by candidates of these levels of proficiency:
 - a. experienced/skilled interpreter (highest score)
 - b. minimally competent interpreter (passing score)
 - c. not-yet competent interpreter or non-interpreter (failing score)

What do we intend to assess?

{For each item type, specific knowledge and skills were provided as defined in the EtoE National Task Force *Recommendations on Designing the English-To-English Interpreting Performance Test.*}

Item specifications:

{For each item type, specifications were provided based on the EtoE National Task Force *Recommendations on Designing the English-To-English Interpreting Performance Test.*}

Resources to support the accuracy of content (medical info): {link to 1-3 reputable webpages}

1. Write your ITEM below; use as much space as needed.

2. Write Model Response #1 (what candidate with the highest/best score would respond)

3. Write Model Response #2 (what candidate with the minimal passing score would respond)

4. Write Model Response #3 (what candidate with a failing score would respond)

Appendix C. Initial EtoE Item Type General Review Form



Date/Time

May 29 2019

09 : 12 PM

EtoE Item Type General Review Form

Item Type - Pick from the Dropdown list*

Bilingual reformulation

Reviewer's name *

Reviewer's non-English language of interpreting (primary)*

Importance for Mode of Interpreting *

	Consecutive	Simultaneous	Sight Translation
1. For what mode of interpreting are the Knowledge/ Skill/ Ability/ Competency (KSA/C)'s measured by this type of item needed? (Check all that apply)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Performance Expectation *

	0- No (These KSA/Cs are not important)	1- Minimal importance (These KSA/Cs might somewhat contribute to performance quality)	2- Moderate Importance (These KSA/Cs are helpful to have, they add to performance quality)	3- Yes, any/every interpreter must have these KSA/Cs.
2. Must the entry-level interpreter possess the Knowledge/ Skill/ Ability/ Competency (KSA/C) measured by this type of item to interpret accurately? E.g., Must the interpreter be able to paraphrase/ have short term memory/ distinguish synonyms in order to interpret accurately?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Frequency*

	0- Never	1- These KSA/Cs are used only in 1 mode of interpreting	2- These KSA/Cs are used in 2 modes of interpreting	3- These KSA/Cs are used in 3 modes of interpreting
3. How often do interpreters need to utilize the KSA/C(s) measured by this type of item?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Consequence*

	0- No effect (accuracy is possible)	1- Minimal effect (accuracy may be slightly affected)	2- Moderate effect (accuracy will be affected to some degree)	3- Substantial effect (accuracy cannot be maintained)
4. To what degree would the inability of the interpreter to perform the item's task affect accuracy of interpreting?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Item ID # of the item you rank as the best (#1) for this type*

Read the proposed items in the corresponding MS Word file and select the item that you consider the best (for various reasons that you do not need to explain at this point).

Item ID # of the item you rank as second choice (#2) for this type*

Read the proposed items in the corresponding MS Word file and select the item that is your second choice (for various reasons that you do not need to explain at this point).

Item ID # of the item you rank as third choice (#3) for this type*

Read the proposed items in the corresponding MS Word file and select the item that is your third choice (for various reasons that you do not need to explain at this point).

Any comments/ suggestions about this TYPE of items

You don't have to provide any item-specific comments at this time. We'll have another form for editing specific items.

©2019, CCHI. All rights reserved.

[Save and Resume Later](#)

Submit Form

Appendix D. Final EtoE Item Type General Review Form



Date/Time

May ▾ 29 ▾ 2019 ▾ 
 09 ▾ : 40 ▾ PM ▾

EtoE Item Type General Review Form

Reviewer's name *

Reviewer's non-English language of interpreting (primary) *

Item Type - Pick from the Dropdown list *

Shadowing ▾

Importance for Mode of Interpreting *

	Consecutive	Simultaneous	Sight Translation
1. For what mode of interpreting are the Knowledge/ Skill/ Ability/ Competency (KSA/C)'s measured by this type of item needed? (Check all that apply)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Performance Expectation *

	0- No (These KSA/Cs are not important)	1- Minimal importance (These KSA/Cs might somewhat contribute to performance quality)	2- Moderate Importance (These KSA/Cs are helpful to have, they add to performance quality)	3- Yes, any/every interpreter must have these KSA/Cs.
2. Must the entry-level interpreter possess the Knowledge/ Skill/ Ability/ Competency (KSA/C) measured by this type of item to interpret accurately? E.g., Must the interpreter be able to paraphrase/ have short term memory/ distinguish synonyms in order to interpret accurately?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Frequency*

	0- Never	1- These KSA/Cs are used only in 1 mode of interpreting	2- These KSA/Cs are used in 2 modes of interpreting	3- These KSA/Cs are used in 3 modes of interpreting
3. How often do interpreters need to utilize the KSA/C(s) measured by this type of item?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Consequence*

	0- No effect (accuracy is possible)	1- Minimal effect (accuracy may be slightly affected)	2- Moderate effect (accuracy will be affected to some degree)	3- Substantial effect (accuracy cannot be maintained)
4. To what degree would the inability of the interpreter to perform the item's task affect accuracy of interpreting?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Difficulty*

	0-30% - The item type is too difficult for an entry-level interpreter.	31-50% - The item type is somewhat difficult for an entry-level interpreter..	51-80% - The item type is of difficulty expected of an entry-level interpreter	81-100% - The item type is easy for an entry-level interpreter.
5. How many candidates would complete this type of items correctly, i.e., at least at the minimally competent level? I.e., how many candidates would get a minimal passing score or higher (include the percentage of all candidates who would "pass" this task).	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

6. Weight compared to other items types: Considering that there are 9 items type on the test and that the total possible weight (score) is 100%, what weight would you assign to this item type. See the sample response below (it's meant just to show you the minimal answer expected of you, but you may add an explanation of your suggestion).*

Example: Shadowing - 10% of the test.

Any comments/ suggestions about this TYPE of items

You don't have to provide any item-specific comments at this time. We'll have another form for editing specific items.

©2019, CCHI. All rights reserved.

[Save and Resume Later](#)

Submit Form

Appendix E. EtoE Study Participation Questionnaire

Name

Email

Language of interpreting (drop down list): Arabic, Mandarin, Spanish

Your CCHI ID (Eligibility ID)

Is interpreting or translation your main profession (means of earning a living)?

Yes/No

What is your current interpreter certification status? Select ALL that apply

- Not certified in interpreting at this time
- passed the CoreCHI™ exam
- CHI™-Arabic
- CHI™-Mandarin
- CHI™-Spanish
- Other medical interpreter certification
- Other interpreter certification (e.g., RID, court)
- ATA certification

1. What is your age?

- 18 to 20 years
- 21 to 30 years
- 31 to 40 years
- 41 to 50 years
- 51 to 60 years
- 61 years and over

2. How do you identify yourself?

- Male
- Female
- Other
- Do not wish to share

3. Which of the following most closely describes the highest level of formal education (from any country) that you have completed?

- Did not complete high school
- High school diploma/GED or equivalent
- Associate degree (any major/specialization)
- Bachelor's degree (any major/specialization)
- Master's degree (any major/specialization)
- Doctoral degree (any major/specialization)
- Post-doctoral degree (any major/specialization)

4. How did you acquire your non-English interpreting language?

- Native speaker
- Second language learner: formal learning (college, etc.)
- Second language learner: informal learning (self-taught)
- Heritage speaker (The person who speaks the non-English language most exclusively at home with family and friends, while growing up and living in an English-speaking country.)

5. Did you have any university-level training in interpreting (regardless of the setting type) (e.g., an interpreting course at a community college, college, or university)?

Yes/No

[If Yes:](#)

5a. How much formal training did you have in interpreting that is not related to healthcare settings?

- Less than 45 instructional hours (3 credits in U.S.)
- 45 instructional hours
- 46-100 instructional hours
- over 100 instructional hours
- Bachelor's degree in interpreting
- Master's degree in interpreting

6. Did you have any university-level training in translation?

Yes/No

[If Yes:](#)

6a. How much formal training did you have in translation (including continuing education and conferences)?

- Less than 45 instructional hours (3 credits in U.S.)
- 45 instructional hours
- 46-100 instructional hours
- over 100 instructional hours
- Bachelor's degree in translation
- Master's degree in translation

7. Did you have any university-level training in linguistics?

Yes/No

8. How much formal (academic or non-academic) training do you have in healthcare interpreting (including continuing education and conferences)?

- Less than 40 instructional hours
- 40 instructional hours
- 41-65 instructional hours
- 66-100 instructional hours
- over 100 instructional hours
- Associate degree in healthcare interpreting
- Bachelor's degree in healthcare interpreting
- Master's degree in healthcare interpreting

9. How did you receive the MAJORITY of training in healthcare interpreting?

- an academic program in medical interpreting of 45 hours (3 credits in U.S.) in duration (any country)
- an academic program in medical interpreting of more than 45 hours in duration (any country)
- a non-college program of 40-100 hours with in-person instruction (e.g., *Bridging the Gap*, *The Community Interpreter*, *The Art of Interpretation*, etc.)
- a non-college program of 40-100 hours with online instruction
- a combination of in-person workshops and conferences
- a combination of online courses and webinars
- on-the-job training

10. What type of training in healthcare interpreting in terms of its content do you have? Select ALL that apply.

- general knowledge of the profession (ethics, role, cultural awareness)
- medical terminology
- interpreting skills: consecutive, simultaneous and sight translation modes
- none of the above

11. Estimate how many hours of training (with an instructor) you have had in consecutive interpreting (regardless of the setting):

- 0-2 hours
- 3-6 hours
- 7-12 hours
- 13-24 hours
- 25-45 hours
- more than 45 hours

12. Estimate how many hours of training (with an instructor) you have had in simultaneous interpreting (regardless of the setting):

- 0-2 hours
- 3-6 hours
- 7-12 hours
- 13-24 hours
- 25-45 hours
- more than 45 hours

13. Estimate how many hours of training (with an instructor) you have had in sight translation (regardless of the setting):

- 0-2 hours
- 3-6 hours
- 7-12 hours
- 13-24 hours
- 25-45 hours
- more than 45 hours

14. In what settings do you interpret regularly? Select ALL that apply

- healthcare
- workers' compensation
- court (incl. immigration)
- non-court legal (police, FBI, attorney)
- conference
- schools
- military
- social services
- business
- other community
- telephone/video (all settings)
- I'm mostly a translator
- none of the above

15. How many years of experience do you have in interpreting (overall, any setting)?

- Less than 2 years
- 2 to 5 years
- 6 to 10 years

- 11 to 15 years
- 16 to 20 years
- 21 or more

16. How many years of experience do you have in healthcare interpreting specifically?

- Less than 2 years
- 2 to 5 years
- 6 to 10 years
- 11 to 15 years
- 16 to 20 years
- 21 or more

17. How much healthcare interpreting experience do you have? Please choose the option that is closest to describing you:

- Novice: You may interpret regularly and/or frequently or not in healthcare settings. You have not interpreted full time or nearly full time for at least a year, or for the equivalent of a year doing part-time work.
- Early career: You may interpret regularly and/or frequently or not in healthcare settings. You have interpreted full time or nearly full time for at least a year, or for the equivalent of a year doing part-time work.
- Experienced: You work fairly regularly and/or frequently in healthcare settings and have interpreted full time or nearly full time for at least five years, or for the equivalent of that amount of time doing part-time work.
- Very experienced: You work regularly and/or frequently in healthcare settings. You have interpreted full time or nearly full time for ten years or more, or the equivalent of that amount of time doing part-time work.

18. What is your current employment status in relation to healthcare interpreting?

- I am a staff interpreter
- I am a freelancer (contractor)
- I am a volunteer
- I'm a dual-role interpreter, with interpreting as a secondary responsibility
- I don't interpret in healthcare settings

19. How many hours do you interpret per week in any setting?

- Less than 2 hours
- 3-20 hours
- 21 – 40 hours
- 41 hours and over

20. How many hours do you interpret per week in healthcare settings specifically?

- Less than 2 hours
- 3-20 hours
- 21 – 40 hours
- 41 hours and over

21. How much time do you spend reading in English (any type of content)?

- I don't spend time reading in English
- less than 30 minutes a week
- 1 hour a week
- 2-7 hours a week
- 8-14 hours a week
- more than 15 hours a week

If "I don't" checked – do not show next question

22. What do you read in English? Select ALL that apply

- social media posts, emails
- news, blogs
- fiction books
- non-fiction articles and books
- professional publications

23. How much time do you spend watching or listening to English-language programs?

- I don't spend time listening to or watching English-language programs
- less than 30 minutes a week
- 1 hour a week
- 2-7 hours a week
- 8-14 hours a week
- more than 15 hours a week

[If "I don't" checked – do not show next question](#)

24. What type of English-language programs do you listen to or watch? Select ALL that apply

- news on TV, radio or online
- radio or TV programs other than news
- movies
- podcasts
- YouTube videos
- music

25. How much time do you spend reading in your non-English language (any type of content)?

- I don't spend time reading in my non-English language
- less than 30 minutes a week
- 1 hour a week
- 2-7 hours a week
- 8-14 hours a week
- more than 15 hours a week

[If "I don't" checked – do not show next question](#)

26. What do you read in your non-English language? Select ALL that apply

- social media posts, emails
- news, blogs
- fiction books
- non-fiction articles and books
- professional publications

27. How much time do you spend watching or listening to programs in your non-English language?

- I don't spend time listening to or watching programs in my non-English language
- less than 30 minutes a week
- 1 hour a week
- 2-7 hours a week
- 8-14 hours a week
- more than 15 hours a week

[If "I don't" checked – do not show next question](#)

28. What type of programs in your non-English language do you listen to or watch? Select ALL that apply

- news on TV, radio or online

- radio or TV programs other than news
- movies
- podcasts
- YouTube videos
- music



ETOE™ Examination Guide



Contents

Introduction	2
EtoE Study Participation Requirements.....	2
CCHI Application Guidelines.....	3
Non-Discrimination Policy.....	3
Confidentiality and Non-Disclosure Agreement.....	4
ETOE™ Examination Description	4
Preparing for the ETOE™ Exam	5
Logistics: ETOE™ Exam Administration.....	6
Admission to Testing Center	6
Identification	6
Testing Procedures	6
Examination Results/ Scores.....	7
Appendix A. Sample ETOE™ Exam Items (Tasks).....	8

Introduction

This guide is intended for use by individuals who are interested in participating in CCHI's ***English-to-English (EtoE) Research Study***. CCHI is seeking **300 Spanish, Arabic and Mandarin interpreters**, who are eligible for the CHI™ certification, **to participate in the *EtoE Research Study* as a volunteer in January – March 2020**.

The interpreter's competencies are very complex and start with the language proficiency in two languages. These competencies also include skills that either have no direct correlation to language proficiency or are not exclusive to language proficiency, – skills responsible for a successful conversion of meaning from one language into another. CCHI decided to find out if such cognitive interpreting skills can be measured via a standardized oral performance test in English so that this test can be used for interpreters of any language.

The EtoE study is conducted with volunteer candidates applying for the CHI™ certification. The study participants will take two exams – the regular, dual-language CHI™-exam in Arabic, Mandarin or Spanish AND the English only performance (ETOE™) exam. The comparison of the results will inform us if there is a correlation between the two tests. If a valid correlation is found, the English only performance exam will enhance the existing CoreCHI™ certification by providing performance testing to interpreters of any language.

This document is ONLY INTENDED AS A GUIDE and only applicable to individuals participating in CCHI's certification programs. CCHI's information, procedures, and fees detailed in this publication may be amended, revised, or otherwise altered at any time, and the most current information is available on CCHI's website (www.cchicertification.org).

All correspondence and requests for information concerning the administration of CCHI examinations should be directed to info@CCHICertification.org.

EtoE Study Participation Requirements

Participation in the EtoE Research Study is voluntary and does not confer any special rights or exemptions beyond the benefits described below.

All study participants must comply with all CCHI's policies (<http://cchicertification.org/about-us/policies/>).

CCHI selects participants at its own discretion. CCHI reserves the right to end registration at any time at its sole discretion.

Who qualifies to participate:

Spanish, Arabic and Mandarin interpreters who meet **ALL** of the following criteria:

- have ***received approval for their CCHI application*** but have ***not yet taken the CHI™*** oral performance exam
- are ***willing to take both skills performance exams*** –
 - the CHI™ certification exam and
 - the ETOE™ Interpreting Skills exam in English
 - are able to ***complete both exams on one testing date between January 24 and March 14, 2020***.

What are the benefits of participating:

- **Help advance the profession by providing critical data that will allow CCHI to definitively determine if it is possible to test *fundamental interpreting skills in an English-only format*.** If the study determines that this is possible, then CCHI will be able to offer a *valid and reliable performance-based oral test for interpreters of all languages without delay*. This is especially significant for interpreters of languages of lesser diffusion for whom it will not be feasible to develop a language-specific interpreting exam within the foreseeable future.
- **\$100 off the CHI™ exam fee.** The cost of the exam becomes \$175 instead of the regular \$275. (Those who have already paid the full fee will receive a \$100 check within two weeks of taking the ETOE™ exam.)
- **The opportunity to take the CHI™ exam first, before the CoreCHI™ exam.** The certification will still be granted only after passing both exams, but the order of taking the exams is switched. This allows candidates to get the “hardest” step (in some candidates’ opinion) of the certification process out of the way first. An added bonus is that in case a candidate does not pass the exam, they may “pause” their process at an earlier step, thus, saving the cost of the CoreCHI exam (\$175).

How to register and participate:

1. **Send us an email to solutions@cchicertification.org** and indicate “**I want to participate in the EtoE Study**” in the subject line. Your email indicates your agreement to comply with all CCHI’s policies (available at <http://cchicertification.org/about-us/policies/>).
2. **Complete the *EtoE Study Questionnaire* online.** We’ll send you the link with the instructions. Only candidates who complete this *Questionnaire* will be allowed to schedule.
3. **Be prepared to schedule your test appointment for about 2.75-3 hours between January 24 – March 31, 2020.** The total duration for the appointment includes: 60 min for the CHI exam, 75 min for the ETOE exam, and about 30 min registration time.

CCHI Application Guidelines

All applicants must submit and pay for their applications online at the CCHI website at: <https://cchi.learningbuilder.com>

All applicants must upload supporting documentation with the application as pdf or image (jpg, png) files.

All questions pertaining to CCHI application for certification should be directed to: apply@CCHICertification.org.

Non-Discrimination Policy

CCHI endorses the principles of equal opportunity and non-discrimination. CCHI does not discriminate with regard to age, gender, national origin, race, religion, ethnicity, disability, marital status, veteran status, sexual orientation, or any other category protected by federal or state law.

Confidentiality and Non-Disclosure Agreement

All exam content is strictly confidential. By participating in the EtoE Research Study and taking the examinations, candidates agree that they shall not disclose, reproduce, or distribute examination content or otherwise compromise the security of the examinations.

CCHI respects the privacy of all applicants and candidates. All materials submitted or received in connection with applications and all test scores are held in confidence, except upon permission for disclosure from the applicant or candidate or except as required by law, including disclosure to governmental licensing bodies upon appropriate written request. The full text of *Confidentiality Policy* is available at <http://cchicertification.org/about-us/policies/>.

ETOE™ Examination Description

The purpose of the ETOE™ Exam is to collect data about how an interpreter performs various cognitive tasks related to the meaning of a message in English. It is not a certification exam. It is an exam that includes different types of items that the study will help us identify as pertaining to interpreting cognitive skills or not. Candidates taking this exam will not receive any score or pass/fail decision. The examination is administered via a computer-based application in a proctored environment at a test center and scored by independent raters.

CCHI convened a group of subject matter experts (SMEs) to develop the ETOE™ exam based on *the EtoE National Task Force Recommendations* and in accordance with the national Job Task Analysis Studies conducted by CCHI in 2010 and 2016.

The ETOE™ examination consists of the following tasks:

Order	Task	Input Type*	Weight, %	# of items
1	Reading Comprehension	Text-to-Audio	10%	1
2	Shadowing	Audio-to-Audio	12.5%	1
3	Finish the Sentence	Audio-to-Audio	8%	5
4	Restate the Message	Audio-to-Audio	12.5%	8
5	Listening Comprehension	Audio-to-Audio	15%	1
6	Memory Capacity	Audio-to-Audio	15%	8
7	Equivalence of Medical Terminology	Text-to-Audio	10%	3
8	Medical Concepts	Multiple-choice question	9%	3
9	Fill-in-the-Blank	Text-to-Audio	8%	3
	TOTAL		100%	33

We ask candidates to:

- Follow the directions for each task precisely – each task has a different purpose and different instructions
- Perform the task to the best of your ability.
- The tasks on the EtoE exam are NOT interpreting. Your responses must be in English.

The exam appointment consists of administering two (2) exams and is 3 hours long. First, candidates will take the CHI™ certification exam in their language (60 minutes), and after a 10-minute break – the ETOE™ Exam. The appointment starts with the usual procedures of registration and launch of the exam. Before each examination itself starts, candidates have 15 minutes to test the headset and audio controls and read the directions in order

to familiarize themselves with the exam interface and ascertain that the equipment is working properly. This introductory time is not counted towards the examination time.

The exam design is the following:

- Candidates listen to the recorded audio prompts or read text prompts on the screen and record their oral responses via a headset.
- All items are delivered in a firm sequential order and require a response, i.e., candidates cannot skip any items and return to them later.
- Candidates are allowed to take notes (paper and pen/pencil are provided by the proctor).
- Candidates cannot pause the prompt's playback.
- Some audio prompts can be played two times, some only once – please read the directions before each item carefully.
- The ETOE™ exam is time-limited and is 75 minutes long. Candidates must manage their time. A time in the top center of the screen displays how much time is left. Playing a prompt twice takes away from the overall exam time.
- Candidate can record their oral responses only once. The time allocated to the recording of each response has been established as sufficient by the subject matter experts.
- Candidates cannot pause their recording. Candidates cannot listen to their recording once its completed.

For the description of the CHI™ oral performance exams (Arabic, Mandarin and Spanish), see our webpage at <http://cchicertification.org/certifications/preparing/chi-description/>.

Preparing for the ETOE™ Exam

Review **Appendix A “Sample ETOE™ Exam Items (Tasks)”** of this *Guide*. It explains the **purpose** of each task we are evaluating in this study and the **exact directions** that candidates will see on the screen during the exam. The *Appendix* also contains examples for each item type with possible correct answers to illustrate how the tasks should be performed.

Watch the **recording of the webinar about the ETOE™ exam** at <https://youtu.be/LAvf6STVsPM>. The video contains the screenshots of the exam interface. The presenter also explains the expectations of the study participants and provides examples for each item type with the possible correct responses.

To prepare for the CHI™ exam, view the recording of the webinar “Nail the Exam: Tips for Taking the CHI™ Oral Performance Exam” at <https://youtu.be/Mpps5zEr3jM>. Also see the screenshots of the exam at <https://youtu.be/hJ-IT4J9MbE> (for a pdf file [click here](#)). More resources are at <http://cchicertification.org/additional-resources/>.

Logistics: ETOE™ Exam Administration

Admission to Testing Center

Information on admission to the testing site will be provided in your **CHI™/ETOE™ Notice to Schedule** from CCHI and in the **Exam Scheduling Confirmation email** from CCHI's testing vendor *Prometric*. You must comply with all information required by CCHI and its designated test delivery vendors.

CCHI examinations are administered at test centers contracted by CCHI and supervised by trained proctors. The proctor's responsibilities are to provide a secure standardized environment for administering CCHI exams in compliance with CCHI's and test center's policies and procedures. The proctor does not have knowledge of nor may comment on the content of CCHI's exams. If the candidate has any questions, concerns or suggestions related to the content or structure of CCHI's exams, they must communicate them in writing directly to CCHI at info@cchicertification.org.

On the day you are scheduled to take your examination, please arrive at the test center at about 30 minutes before your scheduled examination time to prepare for any eventualities.

IF YOU ARRIVE MORE THAN 10 MINUTES AFTER THE SCHEDULED TESTING TIME (if you are late), YOU WILL NOT BE ADMITTED AND WILL FORFEIT YOUR EXAMINATION FEES.

Identification

At the test center, you must present the identification required by CCHI's testing vendor and outlined in your *Notice to Schedule*. It must be a current, valid government-issued photo identification with signature (Driver's license, immigration card, passport, State ID card, or military ID card). Your name on the *Notice to Schedule* must match your valid photo ID document.

Testing Procedures

Report to your designated test center location on the day of the examination at the time you were instructed when your appointment was scheduled. If you arrive more than 10 minutes late you will not be admitted, will forfeit your examination fee, and must reregister for the examination by contacting CCHI.

Candidates are expressly prohibited from bringing certain items to the testing site. Please review the information provided in your scheduling notice about what items are and are not permitted.

The test center will provide paper and pen/pencil for note taking. You may not bring your own paper and pen/pencil. Please make sure you have enough paper before the exam starts. You must leave your notes at the test center at the conclusion of the examination, or your examination will be voided.

Once the candidate's identity is confirmed by the proctor, the proctor will show the candidate to the computer station and log them in to the CHI™ exam. The proctor will help ascertain that the equipment is working properly, instruct the candidate to read the Directions, and answer any procedural questions. This time does not count towards the actual examination time. The proctor will monitor the testing room during the exam. After the exam is submitted by the candidate or the allocated time elapses, the proctor will log the candidate out.

The candidate will be asked to leave the testing room and may have an optional 10-minute break. Then the candidate will be shown to the computer station again, and the proctor will log them in to the ETOE™ exam. The proctor will help ascertain that the equipment is working properly, instruct the candidate to read the Directions, and answer any procedural questions. This time does not count towards the actual examination time. The proctor will monitor the testing room during the exam. After the exam is submitted by the candidate or the allocated time elapses, the proctor will log the candidate out.

If the candidate experiences any issues (including technical issues) during the testing that were not resolved at the test center AND that they feel will affect the outcome of the CHI™ exam, the candidate must notify the proctor before they leave the test center. The candidate must ALSO contact CCHI separately at

info@CCHICertification.org within 24 hours of taking their exam to report the issue(s). All communication with CCHI about testing experience must be in writing.

Examination Results/ Scores

Candidates who take the CHI™ oral performance examination will not receive preliminary results upon completion of the CHI™ examination since this examination requires human scoring. Candidates who take the CHI™ oral performance examination will receive official results within approximately six to eight weeks from the last date of the corresponding testing window via email.

Candidates will not receive any results for the ETOE™ examination. The data collected during this examination will be analyzed by CCHI and published in the *EtoE Study Report* (all personal information will be removed for the analysis and reporting).

Appendix A. Sample ETOE™ Exam Items (Tasks)

1. Reading Comprehension

Purpose of this activity: To assess your understanding of an English written text on a healthcare topic and to assess your ability to speak English.

Directions:

First, read the text carefully. Then read the questions following the text. When you are ready to answer the questions, click the “Record” button and record your answers to all three questions in your own words in English. You have up to 3 minutes to record all your answers. Your answers should demonstrate your understanding of the text in English.

Read this text:

One of many causes of congestive heart failure is high blood pressure. If your heart is too stiff to pump the blood as well as it should, the blood flow slows down. This can lead to a decrease in kidney function. A no-added-salt diet may help keep your blood pressure under control. The more salt in the body the harder it is for the kidneys to excrete water. As less water is removed via the kidneys, water volume in the blood increases raising blood pressure. The heart has more difficulty pumping against higher pressure. Fluid begins seeping out of the blood vessels into the surrounding tissues leading to swelling of the feet and ankles. When fluid accumulates in the lungs there is less room for air and patients feel short of breath. Water build up puts additional pressure on the kidneys further decreasing the urine output.

Answer these questions:

- 1) Explain how salt intake & blood pressure & kidneys are interrelated.
- 2) Explain how congestive heart failure leads to swollen ankles.
- 3) Why is it important for patients to understand the reason why the doctors recommend what they recommend?

2. Shadowing

Purpose of this activity: To assess your ability to focus, to understand a speech in English, and to repeat it in English as accurately as possible.

Directions:

This task is similar to simultaneous interpreting, except you will be doing it in English only. Click the “Play/Record” button, and as you listen to the English recording, **REPEAT** what you hear in **ENGLISH** simultaneously. You will not be able to replay or pause the recording. You must **REPEAT** what you hear within the first 10 seconds of starting the audio. Try to repeat everything exactly as you hear it (verbatim), without omitting, adding, or changing any words. If you stop for any reason, start repeating again whenever you can.

For the first 24 hours, you may experience common side effects such as: sleepiness, headache and dizziness; an upset stomach; warm and dry skin, and flatulence. If we end up performing a biopsy or removing polyps from your gastrointestinal tract, the results will be mailed to you. If you do not receive the results in 2-3 weeks, please contact our office.

3. Finish the Sentence

Purpose of this activity: To assess your ability to complete a sentence in English based on the logic of the beginning of the oral message.

Directions:

First, click the “Play” button to listen to an unfinished statement. You may take notes. You can play the audio a total of 2 times. When you are ready, click the “Record” button and record the WORD or PHRASE in ENGLISH that would best complete the statement.

Chickenpox is a common illness caused by a virus called varicella zoster. People often get the virus as young children if they have not...

Possible response:

had chickenpox before or have not been vaccinated against it.

4. Restate the Message

Purpose of this activity: To assess your ability to re-state (paraphrase) the oral message accurately in English, without adding or omitting any information and without changing the meaning. Re-statement can be achieved by explaining a term or providing its synonym.

Directions:

First, click the “Play” button to listen to the speaker’s message. Your goal is to re-state (paraphrase) the message in English using your own words (using synonyms, changing sentence structure, etc.) and to keep the same meaning of the entire message as much as possible. You may take notes and replay the recording again. You can play the audio a total of two times. When you are ready, click the “Record” button and record your retelling of the message. Do not just repeat what you hear word for word: use your own words in ENGLISH to convey the same meaning as the message. At the same time, not every word can be changed, so some words in your answer will be the same as in the original. Do not omit or add any information (units of meaning) and match the register as much as possible.

You need to watch out for foods with high amounts of carbohydrates. The greater the number of carbs in your daily diet the more elevated your blood sugar level can get.

Possible response:

Be careful with high carbohydrate foods. Consuming more carbs on a daily basis can increase your blood sugar level.

5. Listening Comprehension

Purpose of this activity: To assess your understanding of an English oral speech on a healthcare topic, your ability to identify key points of a speech and restate the message.

Directions:

First, click the “Play” button to listen to the provider’s speech. Your goal is to identify and remember the key points of the speech. You will need to state in your own words all the KEY points the speaker makes. You may take notes and replay the recording again. You can play the audio a total of two times. When you are ready, click the “Record” button and record your retelling of the speech in ENGLISH. Do not repeat the whole speech word-for-word because this is NOT an interpreting task.

Your EKG looks ok, but your symptoms are still a little concerning. I don’t think this is your heart, but I would like to order a stress test to be sure. If it turns out normal, I think we can wait to see if your symptoms continue, get better or worse before ordering any other tests. There are lots of things that can cause headaches, dizziness, a pounding heart, and fatigue. Please pay attention to the stress in your life and try to do something relaxing for a little while each day. Eating a healthy diet with lots of fruit and vegetables, low in fat and salt is also good for your heart. Regular exercise is great for the cardiovascular system, and also helps with stress management. Try to do something light that you enjoy, go for a walk or dancing, at least 3 times a week for 30 minutes.

Possible response:

There are three key points in this speech:

- The provider informs the pt that she is ordering a cardiac stress test even though the EKG was good.
- The pt’s symptoms {of headaches, dizziness, a pounding heart, and fatigue} are still concerning, and she wants to keep an eye on them before ordering other tests.
- The provider gives advice about what the pt can do to alleviate his symptoms. {For example, monitor stress, eat healthy, exercise regularly (something light) at least 3 times a week for 30 min.}

6. Memory Capacity

Purpose of this activity: To assess your short-term memory capacity, i.e., how long a message you can repeat word-for-word.

Directions:

In this section you will repeat exactly every word you hear (i.e., this is NOT interpreting). Click the “Play” button to listen to the message. You may take notes, but you can only play the audio ONCE (you cannot re-play or pause the audio). When you are ready, click the “Record” button and repeat in ENGLISH what you hear. You need to repeat everything exactly as you hear it, without omitting, adding, or changing any words.

Example 1:

The onset of signs and symptoms of ear infection is usually rapid.

Example 2:

You may want to talk to your doctor about osteoporosis if you went through early menopause or if either of your parents had hip fractures.

7. Equivalence of Medical Terminology

Purpose of this activity: To assess your ability to re-state (paraphrase) the underlined medical terms accurately in English, without adding or omitting any information and without changing the meaning. Re-statement can be achieved by explaining a term or providing its synonym.

Directions:

First, read the passage and make sure you fully understand its meaning and the underlined medical terms. When you are ready, click the “Record” button and record how you would rephrase (re-state) it in ENGLISH using your

own words. You must find alternatives for the underlined terms. Words that are not underlined may be changed or kept the same as necessary to maintain the meaning of the original text.

Many elderly patients present with circulatory problems. It is really important for them to manage hypertension carefully as it could lead to stroke, renal failure and even myocardial infarction.

Possible response: Lots of older patients have blood circulation issues. High blood pressure may cause stroke, kidney failure and even heart attack. This is why it must be carefully handled.

8. Medical Concepts

Purpose of this activity: To assess your ability to evaluate equivalency of meaning in English.

Directions:

Read the statement paying special attention to the underlined medical term. Read the four options and choose the option that has the closest meaning to the underlined term in the statement.

If you have congestive heart failure, your outlook depends on the cause and the severity, your overall health, and other factors, such as your age.

- A. If you have a heart defect from birth, your outlook depends on the cause and the severity, your overall health, and other factors, such as your age.
- B. If you have a heart attack, your outlook depends on the cause and the severity, your overall health, and other factors, such as your age.
- C. If your heart cannot pump blood as well as it should, your outlook depends on the cause and the severity, your overall health, and other factors, such as your age.
- D. If your heart stops pumping blood, your outlook depends on the cause and the severity, your overall health, and other factors, such as your age.

(Answer: C)

9. Fill-in-the-Blank

Purpose of this activity: To assess your ability to supply a missing word or phrase based on the logic of the written message.

Directions:

Read the passage below. It has one or more words removed. The removed portion is marked as "...[blank]...". Provide a word or phrase that would make the text logical, correct, and complete. When you are ready, click the "Record" button and record the missing portion in ENGLISH.

Unmanaged diabetes can lead to uncontrolled ...[blank]... levels which can damage the body's organs, including the kidneys.

Possible response:

blood sugar; glucose

Appendix G. Sample Rubric: Reading Comprehension

QUALITY OF SPEECH	Quality of speech focuses on the physical characteristics of the speech produced. Physical characteristics include false starts, self-corrections, repetitions, pronunciation, articulation, volume control, pace, and intonation.			
	0 Unqualified	1 Limited	2 Competent	3 Accomplished
	The response includes major changes in quality of speech that render the meaning of the message largely inaccurate or incomprehensible, or the response is left blank, or is spoken in a language other than English.	The response includes significant changes in quality of speech, which change the meaning of the message.	The response includes slight changes in quality of speech, which do not affect the meaning of the message.	The response sustains a high quality of speech, which maintains the meaning of the message.
TASK COMPLETION	Task completion refers to completion of the task in a relevant manner, from the point of view of following the instructions.			
	0 Unqualified	1 Limited	2 Competent	3 Accomplished
	The response does not provide an answer to any of the questions or is inappropriate.	The response provides an answer to only one of the questions or answers the questions verbatim.	The response provides answers to two of the questions using spontaneous speech.	The response provides thorough answers to all three questions using spontaneous speech.

ACCURACY AND COHESION/COHERENCE	Accuracy focuses on relevance (logical response) and correctness (medical concepts) of the information. Cohesion focuses on the degree to which sentences (or different parts of one sentence) are connected so that the flow of ideas is easy to follow. Coherence is the quality of being understandable. Here, errors include omissions of information, additions of information, and disorganized flow of ideas.			
	0 Unqualified	1 Limited	2 Competent	3 Accomplished
	The response provides an inaccurate and completely incohesive and incoherent message, or the response may be irrelevant, or spoken in a language other than English.	The response provides a partially inaccurate and somewhat incohesive and incoherent message, which may contain incorrect information.	The response provides a generally accurate and mainly cohesive and coherent message, which contains correct information.	The response provides an accurate and thoroughly cohesive and coherent message, which contains correct information.
LEXICAL CONTENT	Lexical content refers to accurate rendition of “units of information” and maintaining register (when possible/applicable). Units of information can be individual words, groups of words, or phrases that communicate a single concept. Register is a variety of language used for a particular purpose or in a particular social setting. Here, errors include inaccurate re-statements of a unit of information, incorrect usage of words/phrases, and unjustified changes of register.			
	0 Unqualified	1 Limited	2 Competent	3 Accomplished
	The response includes completely inaccurate lexical errors and complete changes to meaning, or the response is left blank or is in a language other than English.	The response includes significant lexical errors and significant changes to meaning.	The response includes some minimal lexical errors and minimal changes to meaning.	The response includes virtually no lexical errors or changes to meaning.

GRAMMAR	Grammar includes the set of rules that govern how sentences, phrases, and words are put together in a given language (keeping in mind generally accepted speech patterns). Examples of errors in grammar include verb tense, number, gender, word order, and incomplete thoughts.			
	0 Unqualified	1 Limited	2 Competent	3 Accomplished
	The response includes grammatical errors that render the meaning of the message completely inaccurate, or the response is left blank, or is spoken in a language other than English.	The response includes grammatical errors that shift the meaning and clarity of the message.	The response includes minor grammatical errors that do not affect the meaning and clarity of the message.	The response maintains correct grammar throughout the message.

Key Elements by Scale

1. **Quality of Speech.** This scale is used in considering all EtoE item types. The key elements to consider are Fluency, Expression, Self-Repairs and Repetitions, Pronunciation, and Intonation and Pace.

Key Elements	0	1	2	3
Fluency	Insufficient language to evaluate	Minimal	Good	Excellent
Expression		Labored	Occasional Hesitation	No hesitation
Self-Repairs + Repetitions		Frequent	Occasional, appropriate	Few, Appropriate
Pronunciation		Poor	Good	Excellent
Intonation + Pace		Poor	Good	Excellent

2. **Task Completion.** This scale is used across five of the item types, including Reading Comprehension and Speech Production, Shadowing, Restate the Meaning, Equivalence of Medical Terminology, and Memory Capacity. Because of the variability in the requirements of each task, this scale has variable key elements across item types (tasks).

Task Completion - Reading Comprehension				
Key Elements	0	1	2	3
Speech	Does not complete the task	Some statements may be made/answered by reading from the text verbatim	Spontaneous speech	Spontaneous speech
Length	Insufficient language to evaluate	Too short/long	Somewhat unnecessarily short/long	Adequate length

3. **Accuracy and Cohesion/Coherence.** This scale is used across all item types, with variable key elements across item types.

Accuracy and Cohesion/Coherence - Reading Comprehension				
Key Elements	0	1	2	3
Relevance	Completely irrelevant	Somewhat Irrelevant	Relevant, some tangential/irrelevant information	Relevant and well-developed
Completion	Insufficient language to evaluate	Incomplete: Omits critical/key information	Complete: Omits some minor details	Complete and thorough
Factual information		Incorrect	Partially correct	Correct
Cohesion/Coherence		Mostly incohesive (incorrect coordination of sentence parts, lack of logic/ non sequiturs)	Generally cohesive	Cohesive
Organization		Mostly disorganized	Generally organized	well-organized

Accuracy and Cohesion/Coherence - Reading Comprehension				
Key Elements	0	1	2	3
Cultural and Social References		Includes inaccurate cultural and/or social references	Includes generally accurate cultural and/or social references	Completely accurate cultural and/or social references

4. **Lexical Content.** This scale is applied to the six of the nine EtoE item types: Reading Comprehension and Speech Production, Restate the Meaning, and Listening Comprehension and Speech Production, Equivalence of Medical Terminology, and Finish the Sentence.

Lexical Content - Reading Comprehension				
Key Elements	0	1	2	3
Interference from Another language	Constant	Frequent	Occasional	Virtually none
Vocabulary	Extremely limited	Very limited range	Good range	Rich
Accuracy of Use	Inaccurate	Inaccurate	Accurate	Precise
Register	N/A	Minimal or no attention to	Generally maintained	Appropriately maintained

5. **Grammar.** Grammar key elements are considered differently across the item types 1) Reading Comprehension and Speech Production, and 2) Restate the Meaning, Listening Comprehension and Speech Production, and Fill in the Blank.

Grammar - Reading Comprehension				
Key Elements	0	1	2	3
Errors	Many errors (patterns)	Some errors	Minor errors	No errors
Meaning	Errors render the meaning completely inaccurate	Errors shift the meaning	Errors do not affect the meaning	Meaning is preserved
Structures	Insufficient language to evaluate	Mostly simple	Some control	Good control
Sentence Completion		Incomplete or one-word sentences	Complete sentences	Complete sentences
Interference from Another language	Constant	Frequent	Occasional	Virtually none

Appendix H. Item Analysis Indices Description

©Prometric LLC

DEFINITIONS OF SCALE STATISTICS	
Statistic	Definition
Cases	<i>Cases</i> is the number of valid examinees included in each analysis. Invalid scores are deleted on a score-wise basis, so the number of cases may vary from score to score.
Items	<i>Items</i> is the number of test items included in the analysis of each scale. Items may be, and usually are, included in several scales, causing the sum across scales to exceed the length of the test.
Maximum	The <i>maximum</i> is the highest score encountered for each scale.
Median	The <i>median</i> is the middle-most score of the distribution of scores for each scale. If the number of valid examinees is even, the median is taken as the midpoint of the two middle-most scores.
Minimum	The <i>minimum</i> is the lowest score encountered for each scale.
Mean	The <i>mean</i> is the arithmetic average for each scale across all examinees. Mathematically, $\bar{X} = \sum_i X_i / N$.
SD	The <i>SD</i> or standard deviation is a standard measure of dispersion of scores around the mean of the scores. Mathematically, $S = \sqrt{\left(\sum_i (X_i - \bar{X})^2 \right) / N}$. Note that the descriptive (biased) formula is used in calculation of the standard deviation.
Alpha	Coefficient Alpha, a measure of the internal consistency or statistical homogeneity of a scale, provides an estimate of the scale's reliability. Alpha is the generalization of the KR-20 reliability formula. Mathematically, $\alpha = \frac{k}{k-1} \left(1 - \frac{\sum S_g^2}{S_x^2} \right)$.
SEM	The <i>SEM</i> or standard error of measurement, in classical test theory, is the standard deviation of error scores around true scores. It is also interpreted as the standard deviation of scores that would be obtained on repeated measures of an individual with constant ability. Mathematically, it is computed as $SEM = S_x \sqrt{1 - r_{xx}}$, or here as $SEM = S_x \sqrt{1 - \alpha_{xx}}$.
Mean P+	The <i>Mean P+</i> is the average of the proportions of candidates answering the items correctly, averaged across all items included in the score.
Mean Pearson	The <i>Mean Pearson</i> is the average of the Pearson product-moment (i.e., point-biserial) item-criterion correlations averaged across all items included in the score. Note that this correlation is the correlation of the item with the particular scale only if that scale is selected as the criterion. Please note that the "Own Score" criterion option can produce Mean Pearson results that are difficult to interpret if items are assigned to multiple scores.
Passing	If a passing score is specified for a scale, <i>Passing</i> is the proportion of examinees with scores at or above the passing score.

Appendix I. Standardized Factor Loadings from CFA Models 1, 2, 3

Model 1 Standardized Factor Loadings

Item	Lambda
SH133	0.48
R134	0.62
R135	0.59
R136	0.61
R137	0.49
R138	0.57
R139	0.48
R140	0.61
R141	0.58
B142	0.40
B143	0.34
B144	0.20
F145	0.42
F146	0.33
F147	0.40
F148	0.38
F149	0.43
E150	0.67
E151	0.68
E152	0.67
LC154	0.44
M158	0.46
M159	0.53
M160	0.57
M161	0.65
M162	0.56
M163	0.59
M164	0.62
M165	0.62

Model 2 Standardized Factor Loadings

Item	Lambda1	Lambda2
SH133	0.48	
R134	0.62	
R135	0.59	
R136	0.62	
R137	0.50	
R138	0.58	

Item	Lambda1	Lambda2
R139	0.47	
R140	0.61	
R141	0.58	
E150	0.67	
E151	0.68	
E152	0.66	
LC154	0.44	
M158	0.46	
M159	0.52	
M160	0.57	
M161	0.66	
M162	0.57	
M163	0.59	
M164	0.63	
M165	0.63	
B142		0.38
B143		0.32
B144		0.18
F145		0.62
F146		0.52
F147		0.61
F148		0.57
F149		0.49

Model 3 Standardized Factor Loadings

Item	Lambda1	Lambda2	Lambda3
R134	0.64		
R135	0.59		
R136	0.60		
R137	0.51		
R138	0.58		
R139	0.48		
R140	0.61		
R141	0.59		
E150	0.68		
E151	0.70		
E152	0.70		
B142		0.39	
B143		0.32	
B144		0.18	
F145		0.62	
F146		0.52	

Item	Lambda1	Lambda2	Lambda3
F147		0.61	
F148		0.57	
F149		0.49	
SH133			0.48
LC154			0.42
M158			0.48
M159			0.53
M160			0.58
M161			0.68
M162			0.60
M163			0.60
M164			0.67
M165			0.67