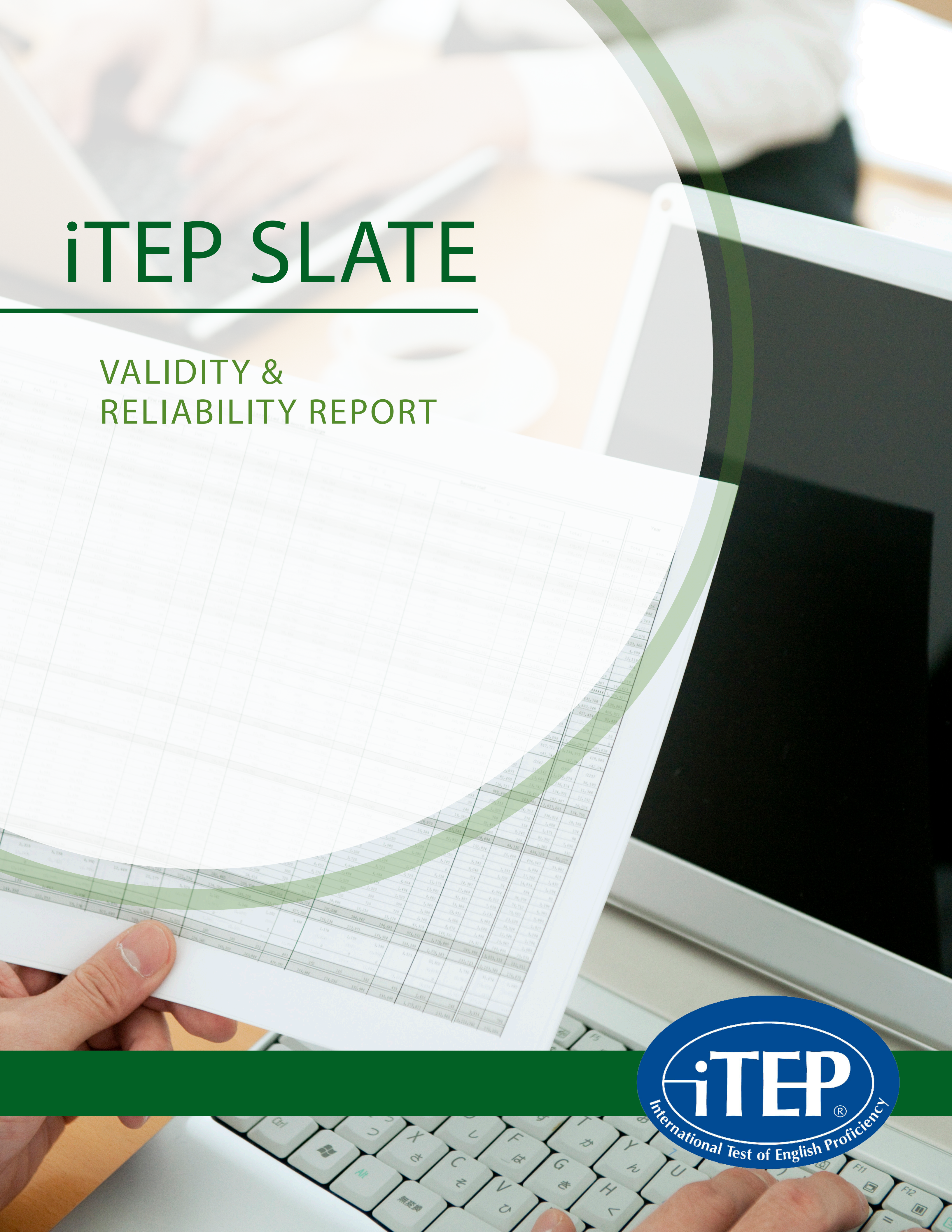
iTEP SLATE

VALIDITY & RELIABILITY REPORT



iTEP®
International Test of English Proficiency

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABOUT THIS REPORT

Purpose

The purpose of this report is to describe the technical details and rationale for the development of iTEP SLATE and to summarize the reliability and validity of the assessment.

Acknowledgments

The data analyzed for this report were provided in raw form by iTEP International to Dr. Stephanie Seiler, Independent Consultant. The data were not manipulated in any way, unless the data manipulation was done according to accepted statistical procedures and/or was done according to other rationale as described in this report. Dr. Seiler conducted all analyses and wrote the report. Dr. Jay Verkuillen and Dr. Rob Stilson provided statistical consultation.

About the Author

Stephanie Seiler holds a PhD in Industrial and Organizational Psychology from the University of Illinois at Urbana-Champaign. Stephanie worked in the personnel selection industry for 5 years as a Selection and Assessment Manager and then as the Director of Research and Development for a leading personnel selection company. In these roles, she was responsible for customer research, consulting, implementation of assessments, and development of new and innovative assessments. Stephanie has published and presented in the areas of educational and personnel assessment, innovative assessment methodologies, and research ethics. She is currently operating as an independent assessment consultant and is a Lead Research Associate and product developer with the University of Illinois at Urbana Champaign with the National Center for Professional and Research Ethics (NCPRE).

About iTEP International

iTEP International (iTEP), founded by career international educators, developed the iTEP suite of examinations to provide institutions and individual test-takers with an efficient, secure, accurate, and affordable on-demand language proficiency assessment.

There are currently seven versions of iTEP available: iTEP Academic, iTEP SLATE (Secondary Level Assessment Test of English), iTEP Business, iTEP Au Pair, iTEP Intern, iTEP Hospitality, and iTEP Conversation. All seven exams have the same basic structure, standardized rubric scoring, and administration procedures. The exams assess all or some of the following five components of English language proficiency: Grammar, Listening, Reading, Speaking, and/or Writing.

CHAPTER 1

INTRODUCTION

The International Test of English Proficiency – Secondary Level Assessment Test of English (iTEP SLATE), developed and published by iTEP International (iTEP) is a multimedia assessment that evaluates the English language proficiency of English as a Second Language (ESL) middle school and high school students.

iTEP SLATE is commonly used for:

- Making high school admissions decisions
- Placing students within language programs
- Guiding course instruction and curriculum development
- Evaluating pre- and post-course progress
- Determining eligibility for exchange programs

iTEP SLATE is also used to assess the proficiency of English language teachers.

In order to target the level and type of English proficiency needed to be a successful student, the content of iTEP SLATE is tailored to reflect the academic and life experiences of individuals who are in middle school and high school. iTEP SLATE does not require any specialized academic or cultural knowledge, so it is well-suited for testing in any academic discipline. The assessment evaluates examinees' ability to apply their English knowledge and skill to process, learn from, and respond appropriately to new information that is presented in English. iTEP SLATE is delivered over the Internet at secure Certified iTEP Test Centers around the world. Examinees can schedule a testing date within three business days of contacting the test center.

There are two versions of iTEP SLATE:

- **iTEP SLATE-Core**: assesses Grammar, Listening, and Reading and is 50 minutes in length, with an additional 10 minutes for pre-test preparation. Results are available immediately.
- **iTEP SLATE-Plus**: assesses Grammar, Listening, Reading, Speaking, and Writing and is 80 minutes in length, with an additional 10 minutes for pre-test preparation. Results are available within 24 hours.

iTEP automatically emails the examinee's official score report to the client. An online iTEP client account provides a variety of tools for managing results.

Approach and Rationale for the Development of iTEP SLATE

School is a social and communicative experience. Whether the student is listening to a lecture, writing a paper, reading exam instructions, working on a group project, or making a purchase at the school store, the ability to understand and use the school's primary language is a fundamental prerequisite for the student to succeed. Though success in school can also depend on factors that have little direct link to language, such as intelligence, motivation, self-discipline, and physical and emotional health, these will have little use for the student if he/she is unable to process, learn from, and respond to information.

iTEP SLATE was designed and developed to provide English language proficiency scores that are valid for many types of educational decision making. The developers of iTEP SLATE recognized that in order to thoroughly evaluate English proficiency, the assessment needed to include items that evaluated both written and spoken language, as well as the examinee's grasp of English grammar. In addition, iTEP developers made the distinction between receptive language skills (i.e., listening and reading) and expressive language skills (i.e., writing and speaking). Assessment items that measure an examinee's ability to express ideas in English were developed for inclusion in iTEP SLATE – Plus.

When language proficiency is measured accurately, reliably, and comprehensively, educators or administrators can use examinees' scores on the assessment to make more rigorous, evidence-based decisions. iTEP SLATE was developed with these goals in mind. Furthermore, iTEP SLATE uses the best technology available and on-demand support to help ensure an engaging, user-friendly examinee and administrator experience.

iTEP is recognized by the Academic Credentials Evaluation Institute (ACEI) and Accrediting Council for Continuing Education and Training (ACCET), as an approved internationally regarded English proficiency exam that meets institutional standards. In addition, iTEP is committed to actively engaging with the international education community through memberships and affiliations with NAFSA, EnglishUSA, TESOL, ACEI, ACCET, and AISAP.

CHAPTER 2

DETAILED DESCRIPTION OF iTEP SLATE

Theoretical Model for Language Assessment

Traditionally, language researchers and educators have grouped language skills into four distinct categories (Listening, Reading, Speaking, and Writing), and from a commonsense perspective this categorization is no surprise, as each of these elements of communication refers to a distinct set of activities and knowledge used for distinct purposes. In addition, it is common for a distinction to be made between language *skills* and language *knowledge* (e.g., grammar and vocabulary) (Bachman, 1990).

On the surface, the Listening, Reading, Speaking, and Writing sections of iTEP SLATE align with the traditional categorization of language *skills,* and the Grammar section aligns with the notion of language *knowledge*. Listening scores reflect the ability to comprehend spoken language, Grammar scores reflect the knowledge of correct grammar, and so on. Additionally, practical considerations clearly warrant testing across multiple competency areas. In the case of admissions, use of multiple measures helps ensure content coverage (measurement breadth) across the most critical elements of language; in the case of placement or program evaluation, multiple measures help pinpoint different areas of examinee strengths and weaknesses.

The traditional categorization of language into skills and knowledge domains may seem to suggest that each iTEP SLATE scale measures an isolated language capability; however, modern theories of language emphasize the interrelatedness of language knowledge and skill and the practical fact that any attempt to measure a single component of language will likely be confounded by other language skills that are necessary to answer the question (for example, an evaluation of reading proficiency requires knowledge of grammar, sentence structure, vocabulary, etc.). In addition, these theories emphasize that one must consider the context in which the communication occurs; communication in a casual setting is likely to involve a different set of competencies—and a different judgment of effectiveness—than communication in an academic or business setting. These modern theories suggest that in practice, language effectiveness must be evaluated in the situational context for which the assessment is to be used (Association of Language Testers in Europe (ALTE), 2011; Bachman, 1990). Plainly stated, a language assessment should represent the real-world use of language. The component parts of language are still relevant to language assessment, but they must be interpreted in context.

iTEP SLATE aligns with best practices in language assessment by evaluating one's ability to communicate effectively in the context of common scenarios that are encountered in school settings.

Description of iTEP SLATE Scales

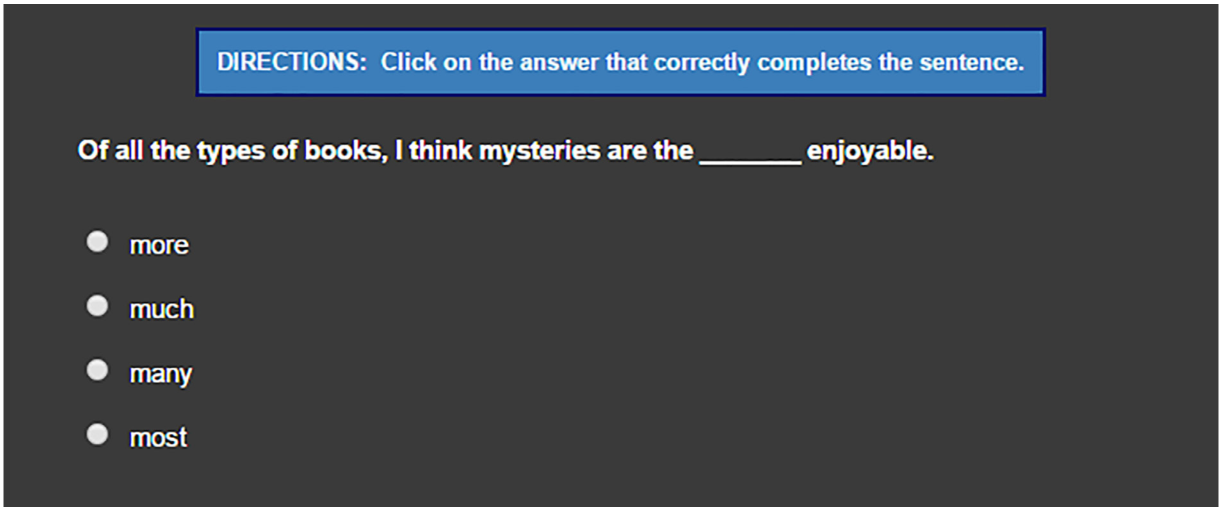
Grammar Section

The ability to understand and use a language's grammar rules correctly is an important component of effective communication. Grammar does not need to be perfect in order for someone to comprehend the meaning of a statement, yet as the number of grammatical errors increases, the likelihood that the information will be conveyed incorrectly also increases. Still higher standards for grammatical correctness are present within most academic settings.

The iTEP SLATE: Grammar section evaluates an examinee's understanding of and ability to use proper English grammar. It is comprised of twenty-five multiple-choice questions, each of which tests the examinee's familiarity with a key feature of English structure (e.g., use of the correct article, verb tense, modifier, or conjunction; identifying the correct sentence structure, pronoun, or part of speech). The Grammar section includes a range of sentence structures from simple to more complex, as well as both beginning and advanced vocabulary. The first 13 questions require the examinee to select the word or phrase that correctly completes a sentence, and the next 12 questions require the examinee to identify the word or phrase in a sentence that is grammatically incorrect. Each of the two question types is preceded by an on-screen example.

The Grammar section takes 10 minutes to complete.

Sample Grammar Item



DIRECTIONS: Click on the answer that correctly completes the sentence.

Of all the types of books, I think mysteries are the _______ enjoyable.

○ more

○ much

○ many

○ most

Listening Section

The ability to comprehend spoken information is of central importance within an academic setting—as well as for navigating the social aspects of academic life. The iTEP SLATE: Listening section evaluates an examinee's proficiency in understanding spoken English information. In this section, the examinee listens to two types of spoken information: (1) a short conversation between two speakers; and (2) a brief lecture on an academic topic. After listening to the conversation or lecture, the examinee is presented with a question (orally and in writing) that measures several key indicators of whether the information was understood. These indicators include: identifying the primary subject of the conversation or lecture (Main Idea), recalling important points (Catching Details), understanding why a particular statement was made (Determining the Purpose), inferring information based on contextual information (Making Implications), and determining the relationship between key pieces of information (Connecting Content).

To ensure realism in the Listening section, item writers take steps to ensure that the content reflects a conversational tone. In addition, while the examinee listens to each audio file, a static image of the speaker(s) is presented onscreen.

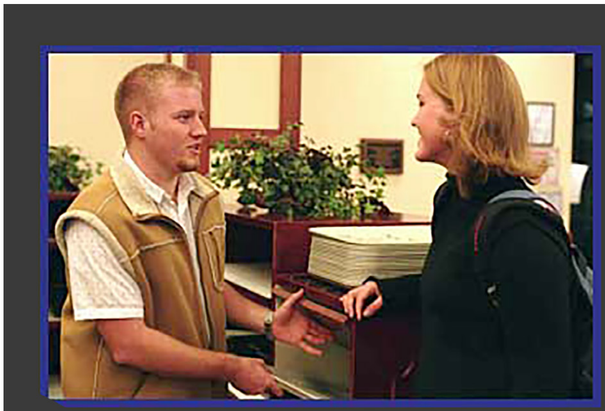
The Listening section takes 20 minutes to complete and consists of three parts:

Part 1: Four short high-beginning to low-intermediate difficulty level conversations of 2-3 sentences, each followed by 1 multiple-choice question

Part 2: One 2- to 3-minute intermediate difficulty level conversation followed by 4 multiple -choice questions

Part 3: One 4-minute low-advanced difficulty level lecture followed by 6 multiple-choice questions

Sample Listening Item



Transcript of audio played to examinee [text is for demonstration in this report and is not presented to the examinee]

Male Student

“Some of us are going to town tomorrow after school. Do you want to come?”

Tara

“Thanks, but I have to write my speech for English class, and then I have to work. I take care of my neighbor’s children on Wednesdays.”

Male Student

“Oh yeah, I saw you in the park with them last week.”

What will the girl do tomorrow after school?

○ Go to town

○ Meet friends in the park

○ Apply for a job

○ Write a speech

Reading Section

Along with Listening, the ability to comprehend written information is critical for effective learning in an academic setting—as well as for navigating academic life in general. Course lectures are typically paired with required textbooks or other reading materials, and students are frequently evaluated on their recall and understanding of both the lectures and the readings. Additionally, a typical course examination involves responding to written materials such as multiple-choice questions or essays; it is also common for students to be assigned themes or essays in which they are expected to write one or more pages of well-reasoned, readable arguments or ideas.

The iTEP SLATE: Reading section evaluates an examinee's level of reading comprehension by measuring several key indicators of whether a written passage was understood. These indicators include: identifying the significant points and main focus of the written passage (Catching Details and Main Idea, respectively), determining what a word means based on its context (Vocabulary), and understanding why a particular statement within a larger passage was written by connecting together relevant information (Synthesis). In addition, the Reading section evaluates the examinee's understanding of how a paragraph should be constructed in order to properly convey information (Sequencing). Sequencing items require the examinee to read a paragraph and determine where a new target sentence should be placed based on the surrounding content.

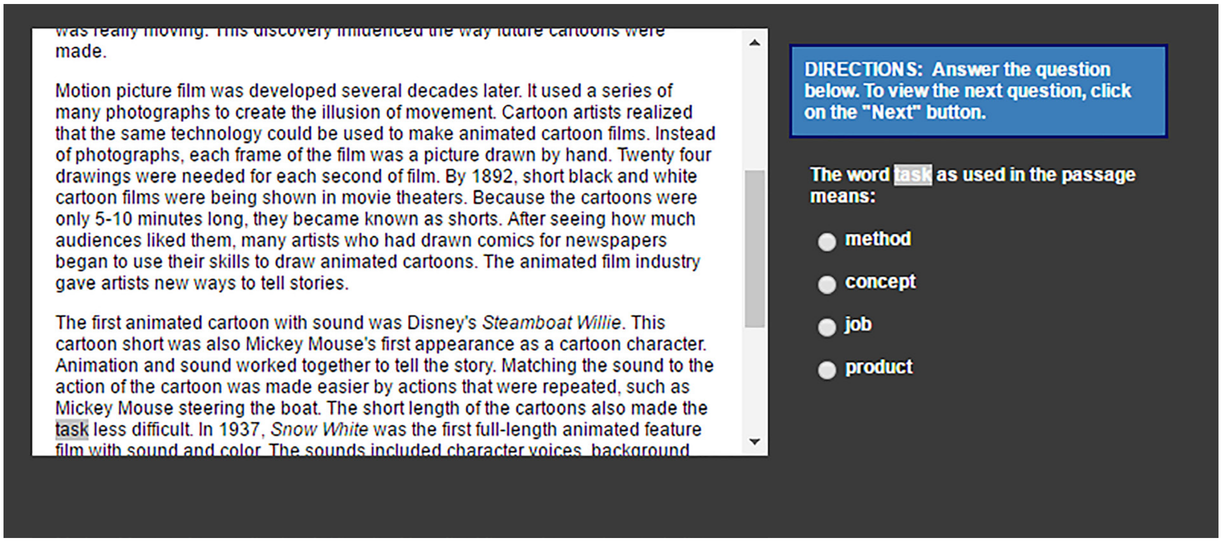
The Reading section takes 20 minutes to complete and consists of three parts:

Part 1: Two intermediate reading level passages of approximately 50 words in length, followed by 2 multiple-choice questions

Part 2: One intermediate reading level passage of approximately 200 words in length, followed by 4 multiple-choice questions

Part 3: One low-advanced passage of approximately 500 words in length, followed by 6 multiple -choice questions

Sample Reading Item



was really moving. This discovery influenced the way future cartoons were made.

Motion picture film was developed several decades later. It used a series of many photographs to create the illusion of movement. Cartoon artists realized that the same technology could be used to make animated cartoon films. Instead of photographs, each frame of the film was a picture drawn by hand. Twenty four drawings were needed for each second of film. By 1892, short black and white cartoon films were being shown in movie theaters. Because the cartoons were only 5-10 minutes long, they became known as shorts. After seeing how much audiences liked them, many artists who had drawn comics for newspapers began to use their skills to draw animated cartoons. The animated film industry gave artists new ways to tell stories.

The first animated cartoon with sound was Disney's *Steamboat Willie*. This cartoon short was also Mickey Mouse's first appearance as a cartoon character. Animation and sound worked together to tell the story. Matching the sound to the action of the cartoon was made easier by actions that were repeated, such as Mickey Mouse steering the boat. The short length of the cartoons also made the task less difficult. In 1937, *Snow White* was the first full-length animated feature film with sound and color. The sounds included character voices, background

DIRECTIONS: Answer the question below. To view the next question, click on the "Next" button.

The word task as used in the passage means:

○ method

○ concept

○ job

○ product

Speaking Section

The speaking and writing in a new language are often considered more advanced skills, developed after the individual has acquired a basic grasp of the language's grammar and vocabulary and learned to apply this knowledge to comprehend written and spoken information. The longer version of iTEP SLATE, iTEP SLATE – Plus, evaluates the examinee's English Speaking ability (along with Writing ability as described next).

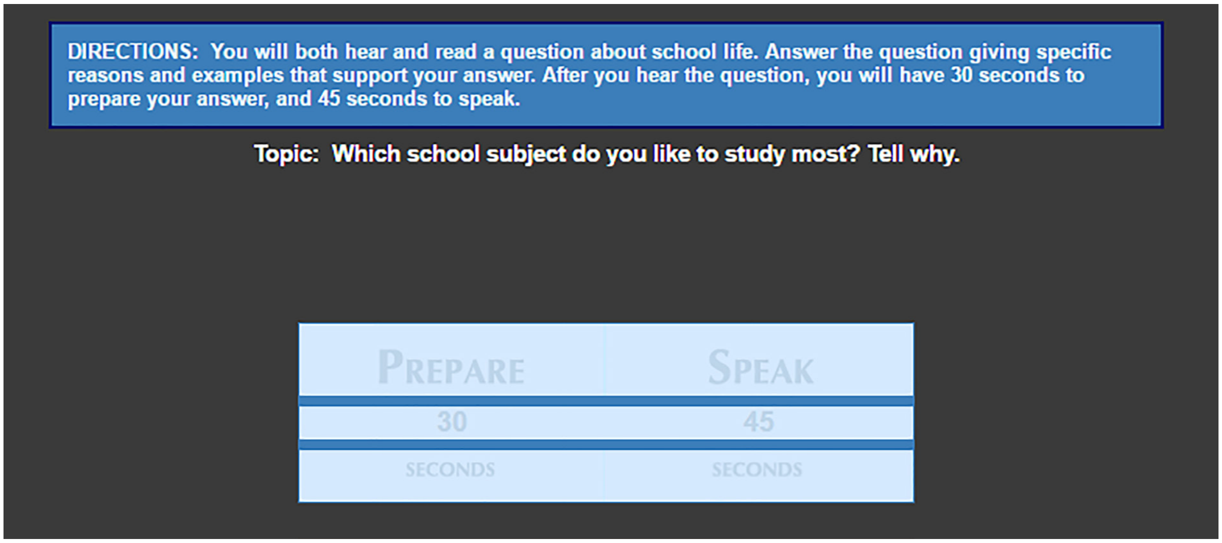
During the Speaking section of the assessment, the examinee listens to and reads a prompt (either a question or a brief lecture), and then prepares an oral response. The examinee then records his/her response for later evaluation by a trained iTEP rater.

The Speaking section takes 5 minutes to complete and consists of two parts:

Part 1: The examinee hears and reads a short question geared at the low-intermediate level, then has 30 seconds to prepare a spoken response, and 45 seconds to speak.

Part 2: The examinee hears a brief upper-level statement presenting two sides of an issue, then is asked to express his or her thoughts on the topic, with 45 seconds to prepare, and 60 seconds to speak.

Sample Speaking Item



DIRECTIONS: You will both hear and read a question about school life. Answer the question giving specific reasons and examples that support your answer. After you hear the question, you will have 30 seconds to prepare your answer, and 45 seconds to speak.

Topic: Which school subject do you like to study most? Tell why.

Writing Section

In addition to the Speaking section, iTEP SLATE – Plus evaluates the examinee's English Writing ability.

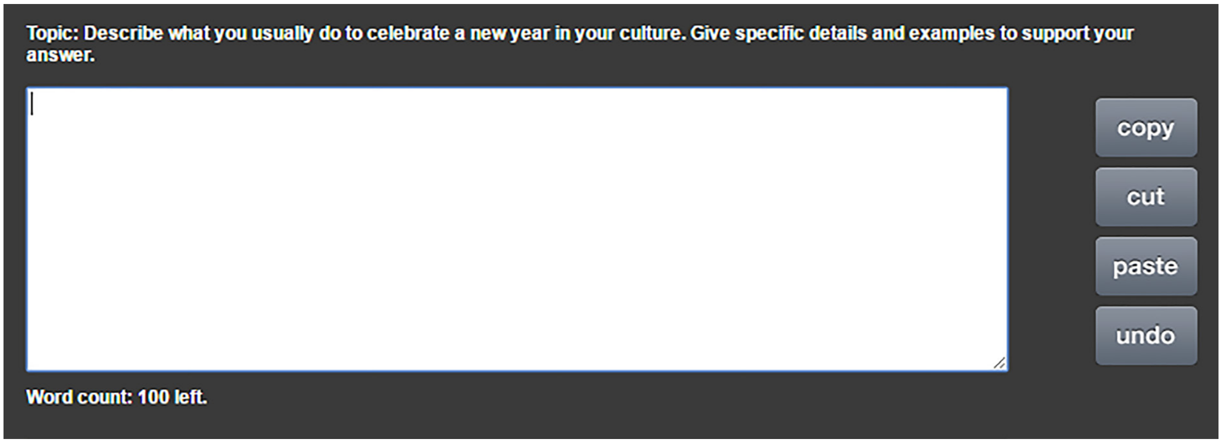
During the Writing section of the assessment, the examinee reads a question and then writes a response. The responses are submitted for later evaluation by a trained iTEP rater.

The Writing section takes 25 minutes to complete and consists of two parts:

Part 1: The examinee is given five minutes to write a 50-75 word note, geared at the low-intermediate level, on a supplied topic

Part 2: The examinee is given 20 minutes to write a 175-225 word piece expressing and supporting his or her opinion on an upper-level written topic

Sample Writing Item



Topic: Describe what you usually do to celebrate a new year in your culture. Give specific details and examples to support your answer.

Word count: 100 left.

Assessment Administration

Delivery Method

iTEP SLATE is administered via the Internet. Items are administered to examinees at random from a larger item bank, according to programming logic and test development procedures that ensure each examinee receives an overall examination of comparable content and difficulty to other examinees.

A static paper-and-pencil version of iTEP SLATE is also available.

iTEP SLATE must be administered at a secure location or a Certified iTEP Test Center.

The examinee inputs responses to the test in the following manner:

- During the Reading, Listening, and Grammar sections, the examinee selects from a list of multiple choice options for each question
- Writing samples are keyboarded directly into a text entry field
- Speaking samples are recorded with a headset and microphone at the examinee's computer

Examinee Experience

Prior to the start of the test, the examinee logs in and completes a registration form. The system guides the examinee through a series of steps to ensure technical compatibility and to prepare him/her for the format of the assessment.

Each section/scale has a fixed time allotted to it. In the Reading and Grammar sections, examinees can advance to the next section if there is time remaining, or they are free to use any extra time to review and revise their answers. In the Listening section, the prompts each play only once and once submitted, an item response cannot be reviewed or changed. In the Writing section, there are fixed time limits for each part, but examinees may advance to the next section before time expires. In the Speaking section, there are fixed time limits for each part and examinees cannot advance until time expires.

The directions for each section are displayed for a set amount of time, and are also read aloud. The amount of time instructions are displayed varies according to the amount of text to be read. If an examinee needs more time to read a particular section's directions, he or she can access them by clicking the Help button, which displays a complete menu of directions for all test sections.

Following each section of the test, examinees see a transition screen indicating which section will be completed next. These transition screen provides a 15-second break between sections, and displays a progress bar showing completed and remaining test sections. After the last test section is completed, examinees see a final screen telling them to wait for further directions from the administrator.

Screenshots of the examinee experience, including pre-assessment modules and instructions, are shown in Appendix A.

Scoring/Grading

iTEP SLATE computes an overall proficiency level from 0 (Beginner) to 6 (Mastery), as well as individual proficiency levels from 0 to 6 for each scale. Sub-scale scores are also computed (e.g. parts of speech, synthesis, main idea), in order to give a more detailed picture of the examinee's skill level. The Overall score represents the combination of scores across each scale; for greater accuracy, Overall scores are reported to one decimal point (e.g., 0.0, 0.1, 0.2, ... , 5.9, 6.0).

iTEP SLATE is graded as follows:

- The Grammar, Listening, and Reading scales are scored automatically by the computer. Each response is worth 1 point. There is no penalty for guessing.
- The Speaking and Writing scales are evaluated by native English-speaking, ESL-trained professionals, according to a standardized scoring rubric (see Appendix B and Appendix C). Raters attend refresher training sessions throughout the year to ensure continued adherence to the rubric.
- For computing the Overall score, each test scale is weighed equally.
- The official score report presents an individual's scoring information in both tabular and graphical formats. The graphical format, or skill profile, is particularly useful for displaying an examinee's strengths and weaknesses in each of the skills evaluated.

Proficiency Levels

The seven iTEP SLATE proficiency levels may be expressed briefly as follows:

Level 0: Beginning

Level 1: Elementary

Level 2: Low Intermediate

Level 3: Intermediate

Level 4: High Intermediate

Level 5: Low Advanced

Level 6: Advanced

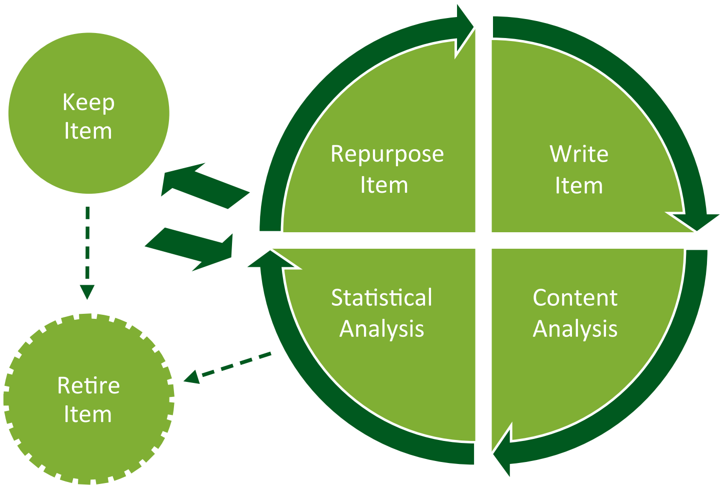
iTEP has mapped iTEP SLATE Proficiency Levels to the levels described in the Common European Framework of Reference for Languages (CEFR; See Appendix D).

CHAPTER 3

iTEP SLATE DEVELOPMENT PROCESS, RELIABILITY, AND VALIDITY[1]

Development Process

iTEP International adheres to a continuous cycle of item analysis (see Figure 1) to ensure the content of the assessment adheres to the reliability and validity goals of the assessment. The cycle begins with item writing, enters an expert review and content analysis stage, and then works through a number of statistical analyses to evaluate the difficulty level and other psychometric properties of the item. Items that do not meet quality standards during the content analysis and/or statistical analysis phase are either removed from further consideration, or repurposed if it is determined that minor adjustments will improve the item. Items that meet quality standards during the content analysis and statistical analysis phases are retained in the assessment; in order to maintain a secure assessment and minimize the likelihood of an item being shared among examinees over time, all items used in the assessment are retired after a certain length



[1] All analysis and evaluation of iTEP SLATE as described in Chapter 3 was conducted in accordance with the Standards for Educational and Psychological Testing (hereafter *Standards;* American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), Uniform Guidelines on Employee Selection Procedures (Equal Employment Opportunity Commission, 1978), and the Principles for the Validation and Use of Personnel Selection Procedures (Society for Industrial and Organizational Psychology, 2003).

of time. Items may also be identified as having “drifted” in difficulty over time, indicating that the item may have been compromised; these items are retired immediately upon identification.

Figure 1. Continuous Cycle of Item Development

Reliability

The reliability of an assessment refers to the degree to which the assessment provides stable, consistent information about an examinee. Demonstrating reliability is important because if a test is not stable and consistent—whether across the items in the assessment, across repeated administrations of the assessment, or based on performance scores provided by trained raters—then the results cannot be relied upon as accurate. Moreover, the reliability of an assessment theoretically sets a maximum limit for its validity; when an assessment is not consistent, it is less effective as an indicator of a person’s true ability and will therefore demonstrate lower correlations with relevant outcomes (such as grades, academic adjustment, or attrition).

Internal Consistency Reliability

Internal consistency reliability refers to the stability of the items within a particular assessment, or in this case, within each assessment scale. When it can be shown that the items are statistically related to each other, the case can be made that the assessment is consistent in its measurement. Cronbach’s alpha (Cronbach, 1951) is a commonly-used and accepted classical test theory (CTT) statistic that is used to estimate internal consistency reliability. The statistic reflects the average correlation between all items within an assessment or assessment section. Values of .70 or above have traditionally been considered desirable, with some scholars stating that test developers should aim to develop tests with values of at least .80 or even .90 and higher. These benchmarks are general rules and do not take into account other desirable characteristics of an assessment, such as assessment brevity to minimize testing time (Gatewood & Field, 2001), the breadth of content coverage within the assessment (to ensure a large domain of the characteristic being measured is represented) (Loevinger, 1954), and the validity of the assessment (Nunnally & Bernstein, 1994). Test developers must think critically about the interrelated factors influencing test reliability and validity and use their best judgment when deciding what should be considered acceptable (Gatewood & Field, 2001).

Because the calculation of internal consistency reliability requires that the assessment scale contain multiple items, this class of statistics is appropriate for the Grammar, Listening, and Reading scales of iTEP SLATE; calculation of internal consistency reliability is not possible for the Speaking and Writing scales, as trained raters provide only one summary score for each of these sections based on the examinee’s overall Speaking or Writing performance.

Within the Grammar, Listening, and Reading sections of iTEP SLATE, the set of items administered to each examinee are selected at random from a larger item bank; therefore, the traditional CTT calculation of Cronbach’s alpha it is not possible. In order to compute an internal consistency reliability estimate for each scale, the following procedure was used to derive an estimate that can be interpreted in a manner similar to Cronbach's alpha. The procedure relies on statistics derived from item response theory (IRT), a class of statistical models that are particularly suited to handling randomly-administered items.

$$ICC = \frac{a^2}{a^2 + \pi^2 / 3}$$

$$a = \frac{K \cdot ICC}{1 + (K - 1) \cdot K \cdot ICC}$$

[2] All examinee data provided by iTEP was included in the analysis, with the exception of the following: (1) when a unique identifier indicated the data was for an examinee re-testing, only the examinee's first testing occasion was included; or (2) if the examinee timed-out on any scale without seeing one or more of the items, the examinee was removed; or (3) examinees younger than 10 or older than 19 years of age were removed. Examinee non-responses to items that were seen but not answered were scored as incorrect.

1. For each scale, compute the IRT common discrimination parameter using the 1-Parameter Logistic model (1PL). The common *a* parameter reflects the average extent to which each item provides statistical information that distinguishes lower-performing examinees from higher-performing examinees. The *a* parameter is in concept most similar to an item-total correlation from classical test theory.

2. Use the a parameter estimate to compute an intraclass correlation coefficient (ICC). This formula is:

3. The resulting value of the ICC reflects the average internal consistency reliability for any one item in the scale, and therefore the final internal reliability estimate (α) must be "stepped up" using the Spearman-Brown prophecy formula to reflect the reliability of the total scale. The Spearman-Brown prophecy method is the same method that would be used to examine the impact of shortening or lengthening a test (for example, cutting a 50-item test in half). The Spearman-Brown prophecy formula is:

   Where K is a scaling factor reflecting the proportional increase or decrease in the number of test items. In the current case, K is the number of items in the scale.

The internal consistency reliability results, which can be interpreted as conceptually similar to Cronbach's alpha estimates, were computed for a sample of over 11,000 examines who completed iTEP SLATE between 2014 and 2016.[2] The results are provided in Table 1. As shown, the Reading estimate exceeds the .70 benchmark, the Listening estimate meets the .80 benchmark, and the Grammar estimate exceeds the .80 benchmark.

Table 1. Internal Consistency Reliability Estimates for Relevant iTEP SLATE Scales

Scale	Number of Items	Discrimination (a)	Intraclass Correlation (ICC)	Internal Consistency Reliability (α)
Grammar	25	.96	.23	.88
Listening	14	.96	.23	.80
Reading	12	.81	.20	.75

Note: The sample size for the analysis was N = 11,405. The Internal consistency reliability estimates are not Cronbach's alpha values, but can be interpreted in a similar manner to Cronbach's alpha.

Test-Retest Reliability

Test-retest reliability refers to the stability of test scores across repeated administrations of the test. A high level of test-retest reliability indicates that the examinee is likely to receive a similar score every time he or she takes it—assuming the examinee's actual skill in the domain being measured has not changed. Test-retest reliability estimates for all iTEP SLATE scales, and the Overall score, were computed using a sample of 126 examinees who took iTEP SLATE twice in an operational environment (i.e., at a testing center). Analyses were restricted to examinees with at least 5 days and less than 2 months between testing occasions (average time elapsed: 24.7 days).

The test-retest values shown in Table 2 reflect the correlation between the Time 1 and Time 2 scores for the sample. Values can range from -1.0 to 1.0, with values at or exceeding .70 typically considered desirable. As can be seen, only the Overall score exceeds this threshold. However, it should be noted that the sample used to compute the test-retest correlations was an operational sample, and it could reasonably be assumed that at least some of the sample had worked diligently to improve their performance between Time 1 and Time 2 testing occasions; given the number of days between test administrations for the sample (up to 2 months; 24.7 days on average), this seems very likely. Had the test-retest estimates been computed on a research sample and/or if the sample size of available data allowed for the analysis of a shorter time period between testing occasions, the correlations would likely be higher. Therefore, the values given in Table 2 can be considered lower-bound estimates of the true test-retest reliability of iTEP SLATE.

Table 2. Test-Retest Reliability Estimates for iTEP SLATE

Scale	Test-Retest Reliability
Grammar	.80
Listening	.59
Reading	.53
Speaking	.79
Writing	.80
Overall	.87
Overall – Core	.82

Note: The sample size for the analysis was N = 126. The OVERALL – Core score was approximated by removing the Speaking and Writing section scores from the Overall scores of examinees who completed the longer iTEP SLATE – Plus.

Rater Agreement

The iTEP SLATE Speaking and Writing sections are evaluated by a trained rater and as such, it is necessary to estimate the accuracy of these judgments—specifically, the extent to which the scores given by rater are interchangeable with the scores of another. Evaluations of rater *agreement*, as opposed to rater *reliability*, are more appropriate in cases where the examinee's absolute score is of interest rather than the examinee's rank order position relative to other examinees (LeBreton & Senter, 2008).

Tables 3 and 4 summarize a raw investigation of rater agreement using a sample of Speaking and Writing ratings from six examinees obtained from eight raters during a training exercise. The examinees completed either iTEP SLATE or iTEP Academic.

It should be noted that the results in Tables 3-6 likely reflect a lower-bound estimate of rater agreement, as the cases used for the training exercise were purposely selected to be more challenging to rate than a typical case.

Table 3. Raw Rater Agreement Analysis - Speaking Scale

		Rater Deviations from Average Score									
Examinee	**Average Score**	**R1**	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**	**Average Deviation**	**Max. Deviation**
E1	1.79	.04	.54	.29	.04	.29	.46	.71	-	.34	.71
E2	4.88	.63	.38	.63	.38	.13	.38	.63	.88	.50	.88
E3	2.53	.28	.97	.03	.22	1.53	.22	.72	.28	.53	1.53
E5	4.03	.22	.72	.53	.53	.22	.03	.53	.47	.41	.72
E6	3.50	.50	.00	1.50	-	-	-	.25	.75	.60	1.50
Average	**3.34**	**.33**	**.52**	**.59**	**.29**	**.54**	**.27**	**.57**	**.59**	**.47**	**1.07**

Note: No Speaking scale ratings were provided for Examinee 4 due to a technical issue with the audio recording. The missing values occurred because the rater(s) did not provide a rating. Average Score: the examinee's average rating across all eight raters. Rater Deviations from Average Score: the absolute value of the difference between each rater's score and the Average Score for each examinee. Average Deviation: average Rater Deviation for each examinee. Max Deviation: highest Rater Deviation value that was observed across all eight raters.

Table 4. Raw Rater Agreement Analysis – Writing Scale

		Rater Deviations from Average Score									
Examinee	Average Score	R1	R2	R3	R4	R5	R6	R7	R8	Average Deviation	Max. Deviation
E1	1.79	.04	.71	.29	.79	.04	.29	.71	-	.41	.79
E2	3.81	.56	.69	.06	.31	.44	.56	.06	.44	.39	.69
E4	3.43	.32	-	.07	.18	.18	.32	.18	.18	.20	.32
E5	4.41	.16	.66	.59	.09	.09	.09	.16	.09	.24	.66
E6	3.60	.15	.40	.85	-	-	-	.15	.15	.34	.85
Average	**3.41**	**.25**	**.61**	**.37**	**.34**	**.19**	**.32**	**.25**	**.21**	**.32**	**.66**

Note: No Writing scale ratings were provided for Examinee 3. The missing values occurred because the rater(s) did not provide a rating. Average Score: the examinee's average rating across all eight raters. Rater Deviations from Average Score: the absolute value of the difference between each rater's score and the Average Score for each examinee. Average Deviation: average Rater Deviation for each examinee. Max Deviation: highest Rater Deviation value that was observed across all eight raters.

As seen in Table 3, in all but 2 instances the raters' Speaking scores for each examinee deviated less than 1 point from the average rating across all raters (as a reminder, scale scores can range from 0 to 6). Across all raters and examinees, the average deviation was .47 points, and the average maximum deviation was 1.07 points. These results suggest a moderately strong agreement across raters.

As seen in Table 4, all of the raters' Writing scores for each examinee deviated less than 1 point from the average rating across all raters. Across all raters and examinees, the average deviation was .32 points, and the average maximum deviation was .66 points. These results suggest a strong agreement across raters.

Using the same data that were used for Tables 3 and 4, rater agreement was also estimated using a version of the r_{WG} agreement statistic (James, Demaree, and Wolf, 1984). The value of r^{WG} can theoretically range from 0 to 1, and represents the observed variability in scores among raters relative to the amount of variability that would be present if all raters had assigned scores completely at random. The formula for r_{WG} is:

$$r_{WG} = 1 - \frac{S^2{}_{x}}{\sigma^2{}_{E}}$$

Where $S^2{}_{X}$ is the observed variance of ratings on the variable across raters and $\sigma^2{}_{E}$ is the variance expected if the ratings were completely random.

The specific version of r_{WG} chosen for the analysis uses a value for $\sigma^2{}_{E}$ that would occur if the raters' completely random scores came from a triangular (approximation of normal) distribution (see LeBreton & Senter, 2008).

The closer an r_{WG} value is to 1, the higher the agreement. There is no agreed-upon minimum value that is considered acceptable for r_{WG}, but as a benchmark, test developers might consider .80 or .90 to be a minimally acceptable value for an application such as assigning ratings based on a score rubric. To put these values in perspective, an r_{WG} of .80 would suggest that 20% (1 - .80) of an average rater's score across examinees was due to error, or factors other than the examinee's "true score" on the exercise.

The r_{WG} agreement statistics are presented in Tables 5 and 6.

Table 5. r_{WG} Rater Agreement Statistics – Speaking Scale

Examinee	Observed Variance	Error Variance	r_{WG}
E1	.20	2.1	.91
E2	.34	2.1	.84
E3	.58	2.1	.72
E5	.24	2.1	.89
E6	.78	2.1	.63
		Average	**.80**

Note: No Speaking scale ratings were provided for Examinee 4 due to a technical issue with the audio recording.

Table 6. r_{WG} Rater Agreement Statistics – Writing Scale

Examinee	Observed Variance	Error Variance	r_{WG}
E1	.30	2.1	.86
E2	.23	2.1	.89
E4	.06	2.1	.97
E5	.12	2.1	.94
E6	.24	2.1	.89
		Average	**.91**

Note: No Writing scale ratings were provided for Examinee 3.

The results in Table 5 indicate moderately strong agreement amongst the raters. The minimum r_{WG} was observed for Examinee 6, with a value of .63. The r_{WG} average across all examinees was .80, indicating that 20% of the average rater's score across examinees was due to factors other than the examinee's "true score" on the exercise.

[3] The modern conception of construct validity refers not just to convergent and discriminant validity, but to the accumulation of all forms of evidence in support of an assessment's validity (AERA et al., 2014).

The results in Table 6 indicate strong agreement amongst the raters. The minimum r_{WG} was observed for Examinee 1, with a value of .86. The average r_{WG} across all examinees was .91, indicating that only 9% of the average rater's score across examinees was due to factors other than the examinee's "true score" on the exercise.

Overall, the results of the rater agreement analyses suggest that ratings provided by any one iTEP rater are likely to be a reliable indication of an examinee's actual proficiency on the Speaking and Writing scales.

Validity

The iTEP SLATE examination was designed and developed to provide English language proficiency scores that are valid for many types of educational decision making. The *Standards* define validity as "the degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test" (AERA et al., 2014, p. 184). In other words, the term validity refers to the extent to which an assessment measures what it is intended to measure. Evidence for validity can, and should, come from multiple lines of investigation that together converge to form a conclusion regarding the relative validity of the assessment, including:

1. Expert judgments regarding the extent to which the content of the assessment reflects the real-world knowledge, skills, characteristics, or behaviors the assessment is designed to measure (Content Validity)

2. An examination of the degree to which the assessment (or assessment scale) is correlated with theoretically similar measures and un-correlated with theoretically unrelated measures (Convergent and Discriminant Validity; traditionally conceived of as the main contributors to Construct Validity[3])

3. An examination of the degree to which the assessment is correlated with the real-world outcomes it is intended to measure, for example: adjustment to school, grades, or improvement in language proficiency (Criterion Validity)

Content Validity

Content validity, or content validation, refers to the process of obtaining expert judgments on the extent to which the content of the assessment corresponds to the real-world knowledge, skill, or behavior the assessment is intended to measure. For example, an assessment that asks questions about an examinee's knowledge of cooking techniques may be judged by experts to be content valid for measuring that aspect of cooking skill, but it would not be content valid for measuring the examinee's athletic ability—*even if* it turned out that cooking assessment scores were correlated with athletic ability.

According to the *Standards* (AERA et al., 2014), evidence for assessment validity based on test content can be both logical and empirical and can include scrutiny of both the items/prompts themselves as well as the assessment's delivery method(s) and scoring.

Content-related validity evidence for iTEP SLATE, for the purposes of academic decision-making, can be demonstrated via a correspondence between the assessment's content and relevant educational and social experiences. To ensure correspondence, developers conducted a comprehensive curriculum review and met with educational experts to determine common educational goals and the knowledge and skills emphasized in curricula across the country. This information guided all phases of the design and development of iTEP SLATE.

Content validity evidence for iTEP SLATE is also demonstrated through the use of trained item writers who are experts in the field of education and language assessment and who have substantial experience in item-writing. The content and quality of items submitted by item-writers is continually supervised, and feedback is provided in order to ensure ongoing adherence to the content goals of the assessment and to avoid content-irrelevant test material. Some of the critical steps taken to achieve this objective are summarized in Appendix E.

Finally, content validity evidence for iTEP SLATE is shown via its correspondence with the CEFR framework. iTEP mapped iTEP SLATE to the CEFR framework through a process of expert evaluation and judgment on the content of the assessment and associated scores.

Convergent and Discriminant Validity

Convergent and discriminant validity evidence is demonstrated through a pattern of high correlations among scales that measure concepts that are known to be closely related, and lower correlations among scales measuring unrelated concepts (AERA et al., 2014). The intercorrelations among iTEP SLATE scales are shown in Table 3. The examinee data analyzed are the same as described in the Reliability section.

Table 7. iTEP SLATE Scale Intercorrelations

Scale	Listening	Reading	Speaking	Writing	Overall
Grammar	.60	.58	.58	.65	.83
Listening	–	.55	.59	.60	.82
Reading	–	–	.49	.55	.78
Speaking	–	–	–	.82	.83
Writing	–	–	–	–	.86

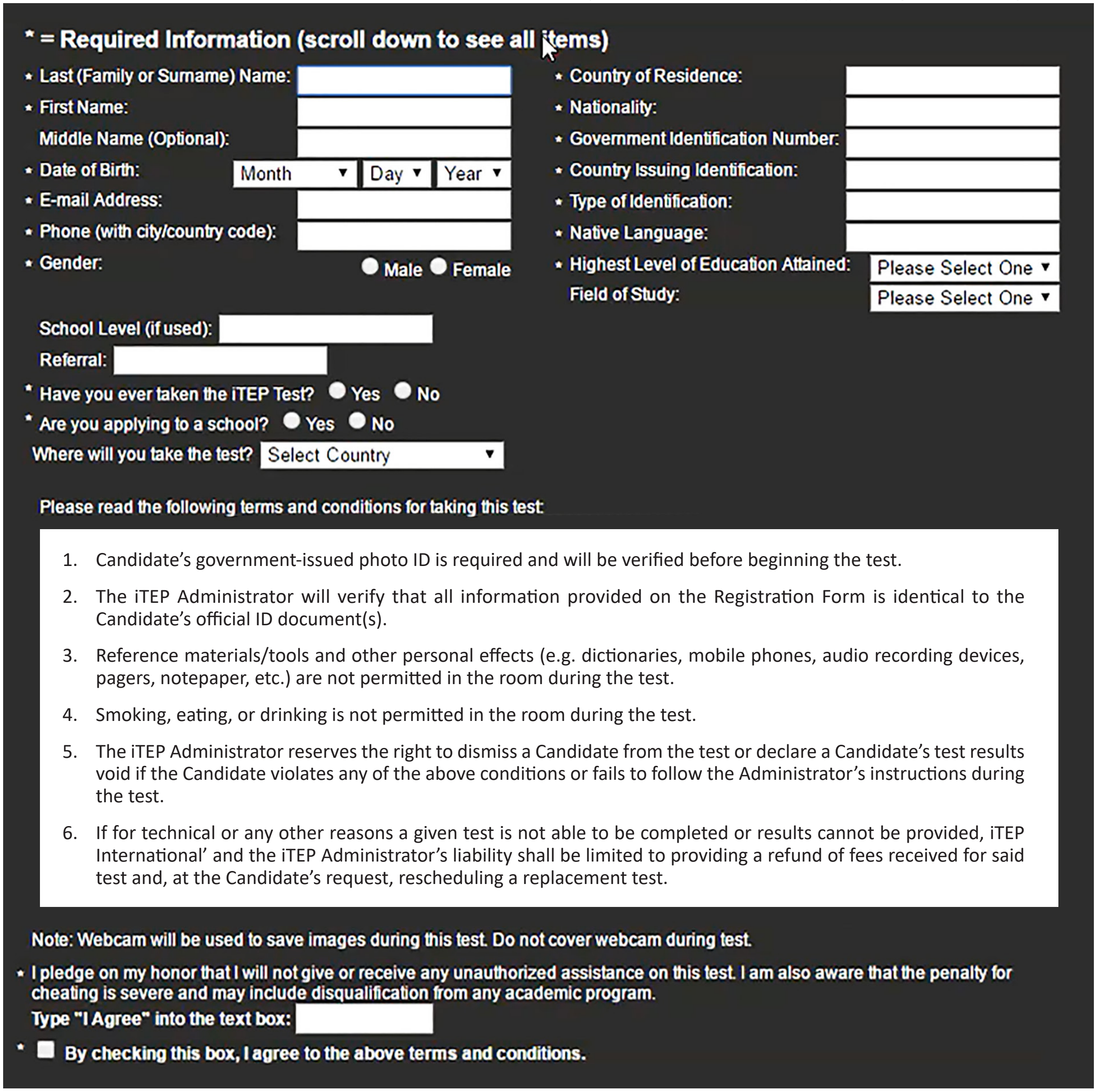
Note: N = 11,105 for correlations involving Speaking; N = 11,400 for correlations involving Writing; N = 11,405 for all other correlations.

The pattern of correlations within iTEP SLATE provides preliminary evidence for the convergent and discriminant validity of the assessment. Overall, the relatively strong correlations between the majority of scales (i.e., in the .50-.60 range) indicates that each scale is likely measuring related components of language proficiency, and the fact that the correlations do not approach 1.0 indicates that each scale likely measures a distinct element of proficiency. Compared with the Grammar/Speaking correlation, the higher correlation between Grammar and Writing is conceptually logical given more weight is placed on grammar, by design, when iTEP raters evaluate examinees' writing ability than when evaluating their spoken ability. The strong correlation between Speaking and Writing is also to be expected, given these skills are considered more advanced demonstrations of language proficiency that require expressive, as opposed to receptive, language skills.

In addition to the internal examination of convergent and discriminant validity within the iTEP SLATE scales, preliminary analyses conducted by a iTEP partner suggested a .93 correlation between iTEP Academic scores and TOEFL® scores. Given the strong similarity between iTEP Academic and iTEP SLATE, the correlation indicates that iTEP SLATE scores are likely to be closely aligned with those of other language proficiency tests.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*.



*** = Required Information (scroll down to see all items)**

* Last (Family or Surname) Name: ▭

* First Name: ▭

Middle Name (Optional): ▭

* Date of Birth: Month ▾ Day ▾ Year ▾

* E-mail Address: ▭

* Phone (with city/country code): ▭

* Gender: ○ Male ○ Female

* Country of Residence: ▭

* Nationality: ▭

* Government Identification Number: ▭

* Country Issuing Identification: ▭

* Type of Identification: ▭

* Native Language: ▭

* Highest Level of Education Attained: Please Select One ▾

Field of Study: Please Select One ▾

School Level (if used): ▭

Referral: ▭

* Have you ever taken the iTEP Test? ○ Yes ○ No

* Are you applying to a school? ○ Yes ○ No

Where will you take the test? Select Country ▾

Please read the following terms and conditions for taking this test

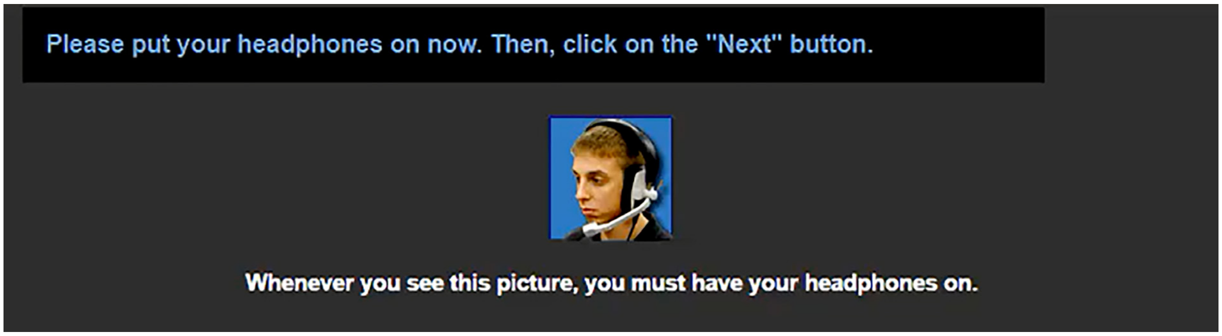
1. Candidate's government-issued photo ID is required and will be verified before beginning the test.
2. The iTEP Administrator will verify that all information provided on the Registration Form is identical to the Candidate's official ID document(s).
3. Reference materials/tools and other personal effects (e.g. dictionaries, mobile phones, audio recording devices, pagers, notepaper, etc.) are not permitted in the room during the test.
4. Smoking, eating, or drinking is not permitted in the room during the test.
5. The iTEP Administrator reserves the right to dismiss a Candidate from the test or declare a Candidate's test results void if the Candidate violates any of the above conditions or fails to follow the Administrator's instructions during the test.
6. If for technical or any other reasons a given test is not able to be completed or results cannot be provided, iTEP International' and the iTEP Administrator's liability shall be limited to providing a refund of fees received for said test and, at the Candidate's request, rescheduling a replacement test.

Note: Webcam will be used to save images during this test. Do not cover webcam during test.

* I pledge on my honor that I will not give or receive any unauthorized assistance on this test. I am also aware that the penalty for cheating is severe and may include disqualification from any academic program.

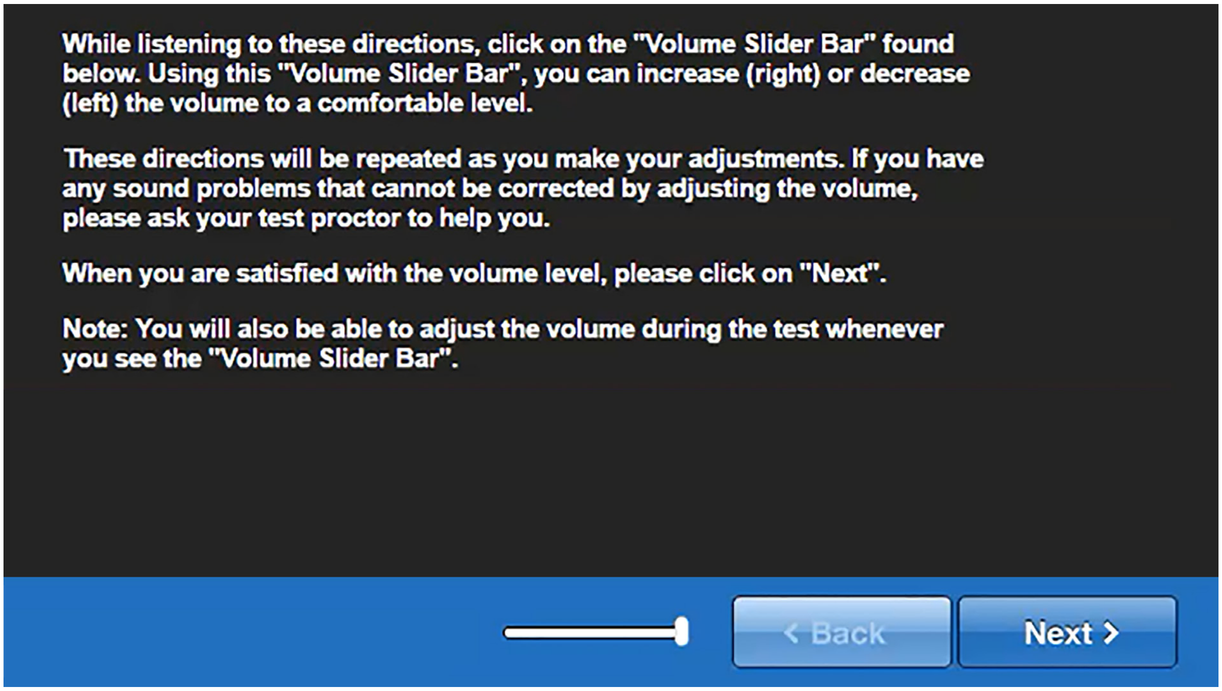
Type "I Agree" into the text box: ▭

* □ By checking this box, I agree to the above terms and conditions.



Please put your headphones on now. Then, click on the "Next" button.

Whenever you see this picture, you must have your headphones on.

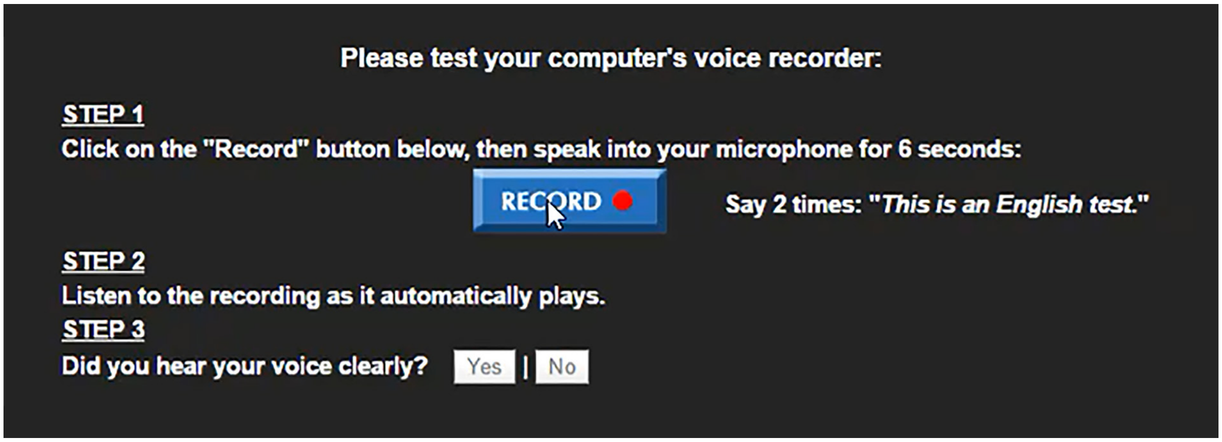


While listening to these directions, click on the "Volume Slider Bar" found below. Using this "Volume Slider Bar", you can increase (right) or decrease (left) the volume to a comfortable level.

These directions will be repeated as you make your adjustments. If you have any sound problems that cannot be corrected by adjusting the volume, please ask your test proctor to help you.

When you are satisfied with the volume level, please click on "Next".

Note: You will also be able to adjust the volume during the test whenever you see the "Volume Slider Bar".



Please test your computer's voice recorder:

STEP 1

Click on the "Record" button below, then speak into your microphone for 6 seconds:

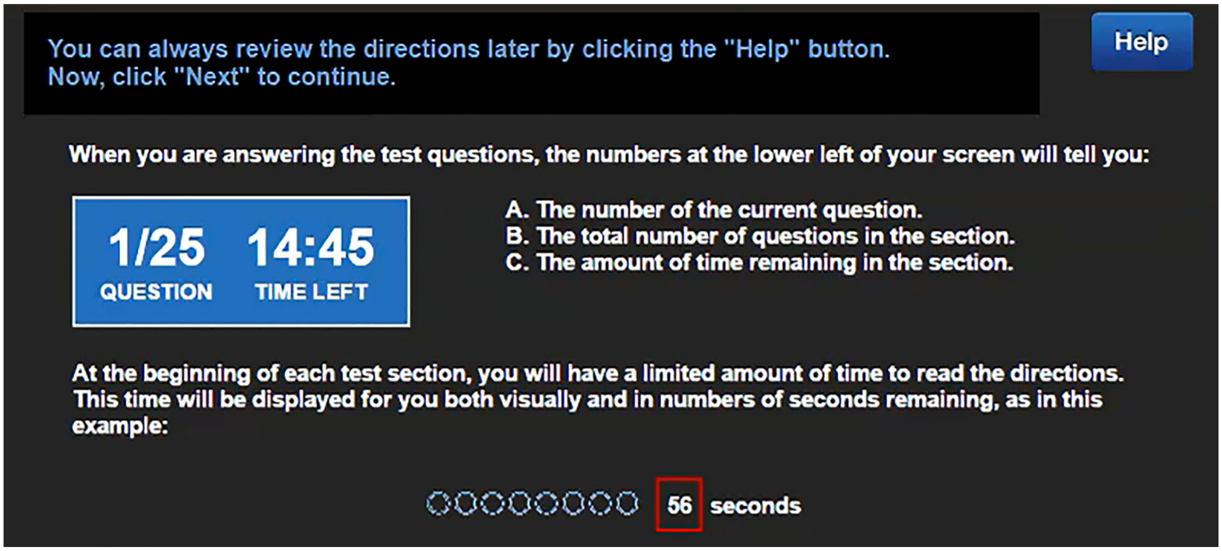
Say 2 times: "*This is an English test.*"

STEP 2

Listen to the recording as it automatically plays.

STEP 3

Did you hear your voice clearly?



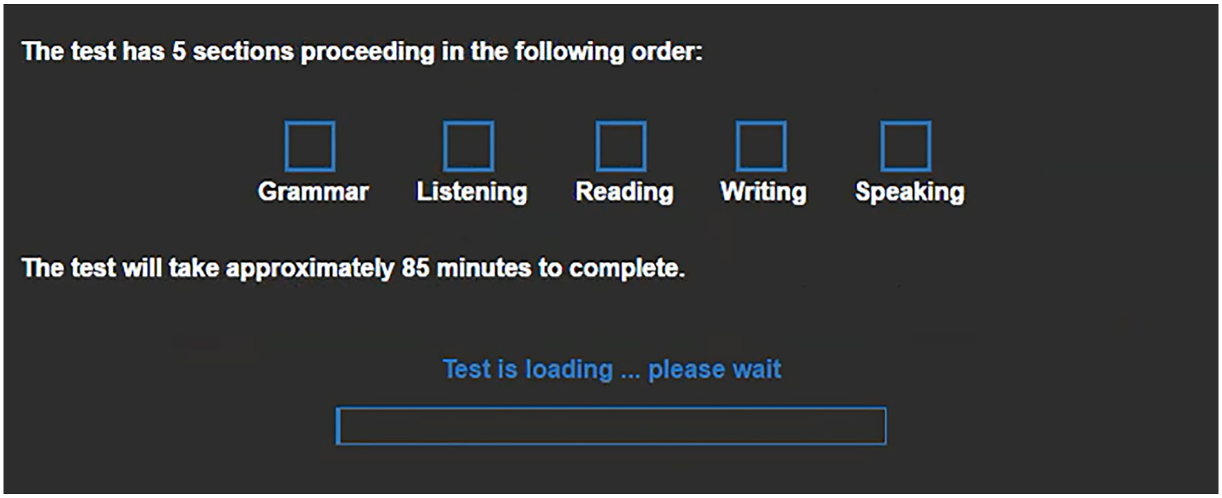
You can always review the directions later by clicking the "Help" button.
Now, click "Next" to continue.

When you are answering the test questions, the numbers at the lower left of your screen will tell you:

A. The number of the current question.
B. The total number of questions in the section.
C. The amount of time remaining in the section.

At the beginning of each test section, you will have a limited amount of time to read the directions. This time will be displayed for you both visually and in numbers of seconds remaining, as in this example:

○○○○○○○○ 56 seconds



The test has 5 sections proceeding in the following order:

□ Grammar □ Listening □ Reading □ Writing □ Speaking

The test will take approximately 85 minutes to complete.

Test is loading ... please wait

APPENDIX B: PEAKING SCALE RATER SCORING RUBRIC

Rating criteria for each Level are discussed in terms of the following:

- General statement of ability and control of language, fluency
- Syntax and grammar
- Lexicon: sophistication of vocabulary
- Degree of elaboration in content, cultural/stylistic appropriateness
- Intelligibility and required listener/reader effort (includes mechanics such as spelling, punctuation, and capitalization)

Level	Rating Criteria
6 **ADVANCED**	• Highly effective control of the language; high degree of fluency; pauses and self-correction highly similar to those of native speakers • High degree of syntactic variety and sophistication; rare minor errors in grammatical usage • Fairly high degree of variety and sophistication in vocabulary for age group, including idiomatic expressions; rare minor errors in word usage • Content elaboration is detailed and relevant to the task; high degree of cultural and stylistic appropriateness; high degree of organizational markers and coherence • Requires only rare effort by listener to determine intended meaning; pronunciation and intonation are highly intelligible with slight non-native influence
5 **LOW ADVANCED**	• Mostly effective control of the language; fairly strong fluency; pauses and self-correction mostly similar to those of native speakers; occasional hesitation and false-starts • Fairly strong degree of syntactic variety and sophistication; occasional minor errors and awkwardness in grammatical usage, more frequent errors in complex structures • Fairly strong degree of variety and sophistication in vocabulary for age group, including occasional idiomatic expressions; occasional errors in word usage • Content elaboration is detailed and relevant to the task; some degree of cultural and stylistic appropriateness; fairly high degree of coherence and organizational markers • Requires occasional effort by listener to determine intended meaning; pronunciation and intonation are mostly intelligible with some non-native influence

Level	Rating Criteria
4 HIGH INTERMEDIATE	• Fairly effective control of the language; adequate fluency; some hesitation and false-starts • Some syntactic variety and sophistication; fairly frequent significant errors and awkwardness in grammatical usage, especially in complex structures • Fair variety and sophistication in vocabulary; rare use of idiomatic expressions; fairly frequent errors in word usage • Content elaboration is somewhat detailed and mostly relevant to the task; some cultural and stylistic appropriateness; some degree of organizational markers and coherence • Requires fair degree of effort by listener to determine intended meaning; pronunciation and intonation are fairly intelligible with moderate non-native influence
3 INTERMEDIATE	• Emerging control of the language; some degree of fluency; frequent hesitation and false-starts • Occasional syntactic variety and sophistication; fairly frequent errors and awkwardness in grammatical usage, even in simple structures • Attempts at variety and sophistication in vocabulary; rare use of idiomatic expressions; frequent errors in word usage • Content elaboration is minimally detailed and fairly relevant to the task; occasional cultural and stylistic appropriateness; attempts at organizational markers and coherence • Requires significant degree of effort by listener to determine intended meaning; pronunciation and intonation are somewhat intelligible with considerable non-native influence
2 LOW INTERMEDIATE	• Weak control of the language; little fluency; considerable hesitation and false-starts • Little syntactic variety and sophistication; significantly frequent errors and awkwardness in grammatical usage, even in simple structures • Little variety and sophistication in vocabulary; little use of idiomatic expressions; significantly frequent errors in word usage • Content elaboration is very minimally detailed and parts may be irrelevant to the task; little cultural and stylistic appropriateness; few attempts at organizational markers and coherence • Requires sustained effort by listener to determine intended meaning; pronunciation and intonation are markedly non-native
1 ELEMENTARY	• Very little control of the language; no fluency; intended meaning is mostly obscured; significant hesitation and false-starts • Very limited syntactic and grammatical skills • Very limited vocabulary • Content elaboration is neither detailed nor culturally appropriate • Requires extreme effort by reader to determine intended meaning; pronunciation and intonation are significantly non-native
0 BEGINNING	• No response or able to respond only with a few prompt-related words or reading of the prompt • Mostly unintelligible; pronunciation and intonation are extremely non-native; extreme hesitation and false-starts • May be off-topic or canned

Note: Examinee scores can be between two levels, e.g., a score of 4.5 indicates the examinee is between levels 4 and 5.

APPENDIX C: WRITING SCALE RATER SCORING RUBRIC

Rating criteria for each Level are discussed in terms of the following:

- General statement of ability and control of language, fluency
- Syntax and grammar
- Lexicon: sophistication of vocabulary
- Degree of elaboration in content, cultural/stylistic appropriateness
- Intelligibility and required listener/reader effort (includes mechanics such as spelling, punctuation, and capitalization)

Level	Rating Criteria
6 **ADVANCED**	• Highly effective control of the language; high degree of fluency • Fairly strong degree of syntactic variety and sophistication; rare minor errors in grammatical usage • High degree of variety and sophistication in vocabulary for age group, including idiomatic expressions; rare minor errors in grammatical usage • Content elaboration is detailed and relevant to the task; fairly strong degree of cultural and stylistic appropriateness; high degree of organizational markers and coherence • Requires only rare effort by reader to determine intended meaning
5 **LOW ADVANCED**	• Mostly effective control of the language; fairly strong fluency • Some syntactic variety and sophistication; occasional minor errors and awkwardness in grammatical usage, especially in complex structures • Fairly strong degree of variety and sophistication in vocabulary for age group, including occasional idiomatic expressions; occasional minor errors in word usage • Content elaboration is detailed and relevant to the task; fairly high degree of cultural and stylistic appropriateness; fairly high degree of coherence and organizational markers • Requires occasional effort by reader to determine intended meaning
4 **HIGH INTERMEDIATE**	• Fairly effective control of the language; adequate fluency • Some evidence of syntactic variety and sophistication; fairly frequent significant errors and awkwardness in grammatical usage • Fair variety and sophistication in vocabulary; rare use of idiomatic expressions; fairly frequent errors in word usage • Content elaboration is somewhat detailed and mostly relevant to the task; occasional cultural and stylistic appropriateness; some degree of organizational markers and coherence • Requires fair degree of effort by reader to determine intended meaning

Level	Rating Criteria
3 INTERMEDIATE	• Emerging control of the language; some degree of fluency • Some syntactic variety and sophistication; fairly frequent errors and awkwardness in grammatical usage, even in simple structures • Attempts at variety and sophistication in vocabulary; rare use of idiomatic expressions; frequent errors in word usage • Content elaboration is minimally detailed and fairly relevant to the task; attempts at organizational markers and coherence • Requires significant degree of effort by reader to determine intended meaning
2 LOW INTERMEDIATE	• Weak control of the language; little fluency • Little syntactic variety and sophistication; significantly frequent errors and awkwardness in grammatical usage, even in simple structures • Little variety and sophistication in vocabulary; little use of idiomatic expressions; significantly frequent errors in word usage • Content elaboration is very minimally detailed and parts may be irrelevant to the task; little cultural and stylistic appropriateness; few attempts at organizational markers and coherence • Requires sustained effort by reader to determine intended meaning
1 ELEMENTARY	• Very little control of the language; rare fluency; intended meaning is mostly obscured • Very limited syntactic and grammatical skills • Very limited vocabulary • Content elaboration is neither detailed nor culturally appropriate • Requires extreme effort by reader to determine intended meaning
0 BEGINNING	• No response or able to respond only with a few prompt-related words or reading of the prompt • Mostly unintelligible • May be off-topic or canned

Note: Examinee scores can be between two levels, e.g., a score of 4.5 indicates the examinee is between levels 4 and 5.

APPENDIX D: ITEP ABILITY GUIDE

iTEP	CEFR	Listening	Reading	Writing	Speaking
6.0	C1 ADVANCED	• Rarely Comprehends overall meaning and virtually all details of lectures on diverse topics	• Requires little extra reading time and use of dictionary	• Satisfies demands of most general academic tasks with occasional grammar and style mistakes	• Pronunciation demands only slight extra effort from listeners
5.9 5.0	B2 UPPER NTERMEDIATE	• Grasps main ideas and the majority of supporting details from academic lectures	• Utilizes contextual and syntactic clues to interpret meaning of complex sentences and new vocabulary	• Writes reasonably coherent essays on familiar topics, but with some grammatical weakness • Exhibits fairly good organization and development	• Expresses viewpoints in fairly long stretches of discourse • Begins to express abstract concepts, especially on familiar topics • Some errors in grammar, word choice, and cultural appropriateness
4.9. 4.0	B1 INTERMEDIATE	• Occasionally needs to ask for repetition or clarification • Begins to determine the attitudes of speakers • Understands main ideas from academic lectures, but misses significant details	• Gathers most main ideas from textbooks and articles, but has an uneven grasp of details • Limited vocabulary impedes speed	• Communicates basic ideas, but with weak organizational structure and grammatical mistakes that sometimes hinder understanding • Does not have a complete grasp of stylistic features • Vocabulary frequently lacks precision and sophistication	• Generates simple questions, greetings, expressions of needs, and preferences • Pronunciation requires significant effort from listeners

iTEP	CEFR	Listening	Reading	Writing	Speaking
3.9 ↑ 2.5	A2 ELEMENTARY	• Maintains comprehension during conversations on familiar topics • Relies heavily on non-verbal cues and repetition • Unfamiliarity with complex structures and higher-level vocabulary leaves major gaps in understanding	• Begins to determine meaning of words by surrounding familiar context • Understands simple reading materials • Major vocabulary gaps lead to frequently inaccurate or incomplete comprehension, and slow pace	• Expresses him/herself with some circumlocution on topics such as family, hobbies, work, etc. • Considerable effort required by the reader to identify intended meaning • Uses only basic vocabulary and simple grammatical structures	• Generates simple questions, greetings, expressions of needs, and preferences • Pronunciation requires significant effort from listeners
2.4 ↑ 0.1	A1 BEGINER	• Understands very basic exchanges when spoken slowly using simple vocabulary • Understands simple greetings, statements, and questions when spoken with extra clarity • Follows simple familiar instructions • Frequently requires repetition for comprehension • Understands a few isolated words or phrases spoken slowly	• Comprehends only highly simplified phrases or sentences • Identifies the main idea of short passages • Recognizes familiar cohesive devices and basic pronouns • Demonstrates understanding of a few simple grammatical and lexical structures • Recognizes the alphabet and isolated words	• Writes complete sentences on everyday subjects with reasonable phonetic accuracy using short words • Still makes basic mistakes systematically • Writes only short, simple sentences, often characterized by errors that obscure meaning • Provides personal details with correct spelling and can copy familiar words and phrases • Produces isolated words and phrases	• Capable of short, simple presentation on familiar topic • Responds to simple questions • Speech is marked with non-native stress and intonation patterns • Communication is understood for short utterances • Pauses, false starts, and reformulation are common • Communicates with single words and short phrases at "survival level" • Intense listener effort required • Produces a few isolated words and phrases • Pronounciation is mostly unintelligible

APPENDIX E: SUMMARY OF STEPS TO MINIMIZE CONTENT-IRRELEVANT TEST MATERIAL

- Implement best practices in item writing to reduce the likelihood that "test wise" test-takers will be able to select the best answer, through cues in the test, without needing to understand the test item itself (for example, by selecting the lengthiest option, eliminating options that are saying the same thing in different ways)
- Avoid content that may influence test-takers' performance on the test—items respect people's value, beliefs, identity, culture, and diversity.
- Topics on which a set of items may be based are submitted by item writers to BES; BES pre-approves topics prior to item writing
- Assessment content reflects the domain and difficulty of knowledge of someone with the educational level of a seventh or eighth grade student. The content reflects materials that an examinee would be expected to encounter in textbooks, journals, classroom lectures, extra-curricular activities, and social situations involving students and classroom teachers and school administrators. Items do not reflect specialized knowledge.
- Write items at an appropriate reading level (no higher than grade 9; lower reading level for easier items); avoid words that are used with low frequency
- Test items assess comprehension within the item, as opposed to common knowledge. Passages establish adequate context for the topic, but then go on to introduce material that is not generally known. Examinees should be able to gain sufficient new information from the passage to answer the questions.
- Content does not unduly advantage examinees from particular regions of the world.

iTEP®

International Test of English Proficiency