Adaptive Assessment Development and Validation

Version 1.2, January 2017 (No IP)

Table of Contents

KITE Overview

KITE is an innovative English language assessment system created by Kaplan's language education and assessment experts. Grounded in evidence-based learning and assessment principles, KITE delivers cloud-based adaptive English language proficiency assessments that address the complex needs of institutions and organizations around the globe.

Educational institutions, businesses, and government agencies worldwide can use KITE as a fair, valid, and reliable assessment tool for:

- Placement
- Student progress tracking
- Workforce evaluation and career development
- Recommendations to focus learning and instruction

KITE accurately measures listening, reading, grammar, writing, and speaking skills that are aligned with the internationally recognized Common European Framework of Reference (CEFR) proficiency levels. KITE assesses communicative proficiency skills ranging from beginning to advanced, and is designed for individuals ages 16 and up.

To pinpoint where an individual's language skills are along a seven-level learning continuum, KITE uses an Item Response Theory (IRT) engine. The engine adapts to each individual's ability level, generating more difficult items for higher-performing test takers and easier items for lower-performing test takers. KITE assessments are flexible in that organizations can choose to use the "Main Flight" (listening, reading, and grammar) on its own or add the "Extras," which are additional productive sections (writing and/or speaking).

IRT and Adaptive Assessment Design

Item Response Theory (IRT), the foundation for KITE, is widely used in education and other fields to: 1) calibrate and evaluate items in tests, questionnaires, and other instruments; and 2) score test takers on their abilities, attitudes, or other traits. Nearly all major educational tests—including the SAT (formerly the Scholastic Aptitude Test), Graduate Record Examinations (GRE), Graduate Management Admission Test (GMAT), Law School Admission Test (LSAT), and many others—rely on IRT because IRT methods offer significant benefits compared with traditional testing models.

In essence, IRT is a probabilistic model that attempts to explain the response of a person to an assessment item. It takes into account the idea that different items require different levels of ability—some items are likely to be answered correctly only by those who have a high ability level, while other items are easier and may be answered by those who have a lower ability level.

One of the most important advantages of IRT is that the performance of different test takers can be compared even when those test takers have answered different items. Also, improvements in individual proficiency over time can be accurately measured even when the individual takes different tests at different points in time. These features mean that IRT is ideally suited for use in online adaptive testing engines like KITE. In KITE, each item is selected to provide the maximum information about the test taker's ability, based on how he or she has answered previous items. Therefore, time is not wasted taking a lot of items that are far too easy or too difficult. KITE matches items to the test taker's ability level and continuously updates that estimate until it is sufficiently precise. The result is a significant reduction in testing time and greater precision and reliability, compared with traditional fixed or static tests in which every student answers the same exact set of items.

KITE Features

KITE also has many other important advantages over traditional fixed tests:

- A computerized adaptive assessment system like KITE provides more efficient testing and more accurate test results because it chooses items that specifically target each test taker's estimated ability level.
- KITE provides instant test results on the listening, reading, and grammar sections, which allow institutions to make informed decisions immediately.
- Organizations can administer assessments in their own setting and on their own schedule because KITE is a cloud-based system.
- KITE recommends skill areas for improvement at each ability level.
- KITE cumulatively tracks the progress of each individual test taker.
- Test results can be used to adapt or personalize instruction to address individual test taker's needs.
- KITE is aligned with the Common European Framework of Reference (CEFR) and is scientifically equated to the IELTS® exam.
- Cheating is reduced because each KITE test taker sees different sets of items.
- KITE is more cost effective because it is paperless and operates on any mobile internet-enabled device and does not require downloaded software.
- Test takers can more accurately show their linguistic knowledge and abilities because the KITE Main Flight assessments are untimed.
- Only highly effective test items are used in the assessments. KITE statistically calibrates and scales items, which effectively identifies items that need to be modified or discarded.

Presentation of evidence in this section follows professional standards established by the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) in their jointly published Standards for Educational and Psychological Testing (American Educational Research Association, et al., 2014), referred to hereafter as "the Standards." Development of content for KITE followed principles and recommendations in the Standards, as well as those in the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (Council of Europe, 2001).

As defined by the Standards, "Validity refers to the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests…" and is "therefore the most fundamental consideration in developing tests and evaluating tests." Establishing validity for a measurement instrument requires accumulating evidence to support the inferences made from the information provided by the instrument. Thus, validity is not considered a feature of a measure, but rather the collection of evidence that supports the intended uses of it (see American Educational Research Association, et al., 2014).

Following the Standards, evidence for the validity of KITE is organized into these major categories:

1. Test Content
2. Internal Structure
3. Relationships to Other Measures
4. Validity and Consequences of Testing

Within each category, we discuss the evidence and theory supporting the interpretations and uses of KITE scores.

1. Test Content

This section addresses the question "What is the relationship between items on the test and the construct the test is supposed to be measuring?" Answering this question requires that we first explicitly define the construct to be assessed. For KITE, the driving construct is English language proficiency as mapped in the descriptive scales of 1) the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (Council of Europe, 2001), the most widely used international standard for profiling language proficiency; and 2) the British Council—EAQUALS Core Inventory for General English (North, Ortega, & Sheehan, 2010).

The Common European Framework of Reference (CEFR) breaks down language learning into six proficiency levels (A1, A2, B1, B2, C1, and C2) and provides a comprehensive description (can-do statements) of what individuals should be expected to do in listening, reading, speaking, and writing at each proficiency level. The Core Inventory for General English outlines specific language skills (e.g., grammar, vocabulary, functions, etc.) expected of English Language Learners (ELLs) relative to the CEFR proficiency levels.

Alignment to the CEFR

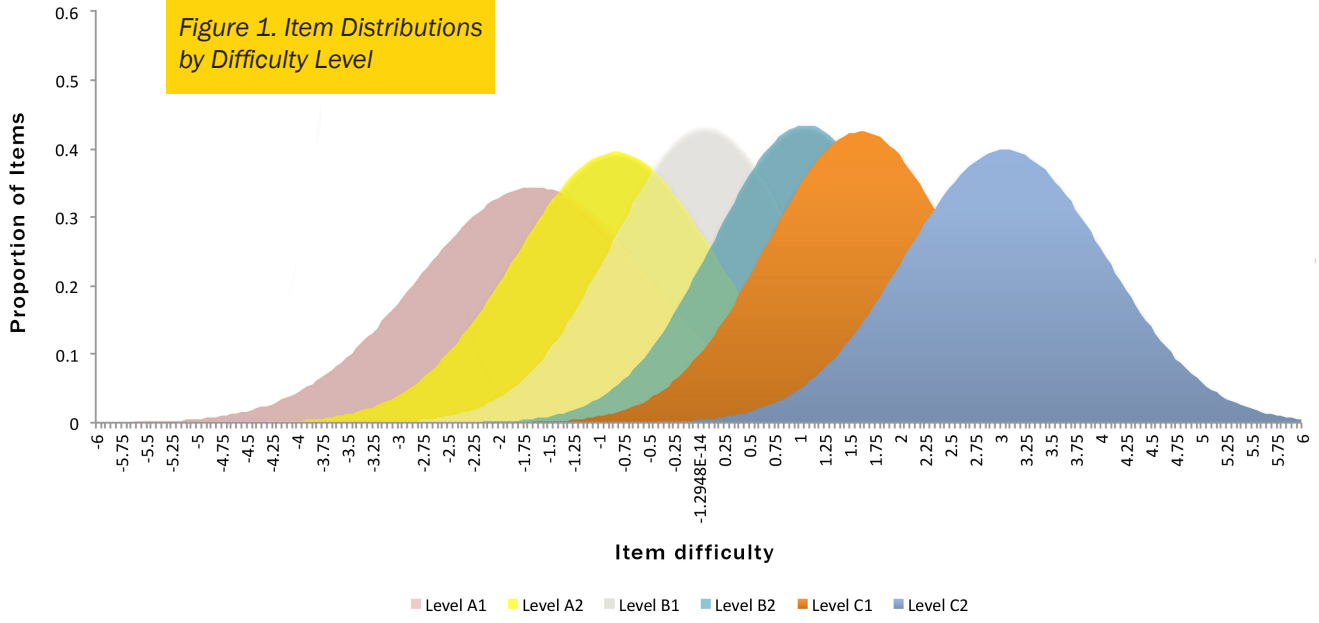
To operationalize the CEFR can-do statements and the Core Inventory into test objectives, language experts at Kaplan International English used the competencies and skills embedded in each descriptor or language point to define how test takers would be expected to demonstrate relevant abilities on an English proficiency test. This entails specifying task purpose, context, content, and constraints that are relevant for each skill in order to clearly distinguish between proficiency levels. These specifications were developed to maximize authenticity of language and to map descriptors from the CEFR and the Core Inventory to each item. In addition, supplementary CEFR assessment manuals, grids, and toolkits developed by the Association of Language Testing in Europe and the Council of Europe informed the development of test blueprints. As a result, KITE content assesses listening, reading, grammar, writing, and speaking skills that are closely aligned with the CEFR framework. To see the KITE content and structure overview, see Appendices (pages 18–19).

2. Internal Structure

The internal structure of the assessment instrument comes from the construct map and the ordering of the skills from different stages on the map. Generally, the skills representing the lower levels on the construct map are associated with items targeted at lower ability levels, and skills representing higher levels are associated with items targeted at higher levels. What should be apparent from the estimated item difficulties is that items measuring skills targeting lower levels of the construct map should be easier, and items measuring skills targeting higher levels of the construct map should be more difficult.

One way of providing evidence that the items support this internal structure is to look at the means and distributions of item difficulties by ability level. A useful plot of this information for each CEFR level is provided below. This plot demonstrates that items targeting progressively higher ability levels are progressively more difficult, which in turn provides evidence that KITE test items effectively operationalize the underlying construct map.

Figure 1. Item Distributions by Difficulty Level



3. Relationships to Other Measures

KITE and IELTS®

When other tests of the same (or similar) construct are administered, evidence of strong relationships between such measures and the instrument being developed can be provided as validity evidence. In the case of the KITE assessment, a linking study was conducted with the IELTS exam. The results of the linking study indicated strong positive relationships between the KITE and each of the existing measures to which it was linked.

A linear regression of IELTS on KITE produces a strong and significant prediction between KITE and IELTS scoring, as the correlations between the KITE overall score and IELTS overall score is .792. In addition, the correlation between the Listening subscales on the KITE and IELTS is .645, and the correlation between the Reading subscales on the KITE and IELTS is .700. Two hundred and two examinees were included in this study, of which 40% were female and 60% were male, and the pool included a diverse population of nationalities and language backgrounds. The data analysis was conducted by an independent psychometrician. A sample of the correlation data from the study with corresponding CEFR levels can be found below.

Table 1. Scale Comparison of KITE, IELTS, and CEFR level

KITE Overall Score Samples and CEFR Level		IELTS Overall Score and CEFR Level	
600	C2	8.5	C2
500	C1	7	C1
425	B2	5.5	B2
350	B1	4.5	B1
300	A2	4	Below B1
250	A1	3.5	Below B1

**KITE and IELTS have been correlated at scores below 4.5, however, IELTS does not report CEFR levels below B1. This table is meant for illustrative purposes only and should only be used for estimation.*

KITE and TOEFL®

Because we have equated KITE and IELTS, and IELTS and the TOEFL iBT exam have been previously equated, we can predict TOEFL scores from KITE scores. Here is a chart indicating the relationships between overall KITE, IELTS, and TOEFL scores:

Table 2. Scale Comparison of KITE, IELTS, and TOEFL iBT

KITE Overall Score Samples	IELTS Overall Score Band	TOEFL iBT Total Score Range
600	8.5	115-117
500	7	94-101
425	5.5	46-59
350	4.5	32-34
300	4	0-31
250	3.5	0-31

**The KITE overall score samples indicate one point that would fall within the corresponding TOEFL iBT score ranges.*

Evidence of Effect of Instruction on Test Scores

Kaplan uses KITE for placement, progress, and exit testing at its international English language schools. In addition, KITE is used for placement/diagnostic and progress testing at other institutions that teach English internationally. Preliminary studies have been conducted to evaluate the effect of instruction on KITE test scores.

One study showed that 92% of test takers who were enrolled in an English language program increased their overall level scores after approximately 10 weeks or 200 hours of instruction. This amount of time is significant because current research reports that it takes students approximately 200 hours to progress to the next CEFR level (Cambridge University Press, 2013). The average increase in score was 56 points, and 67% of the students moved up to the next CEFR level or higher. This study indicates that instruction has a significant effect on KITE scores when studying at an intensive English language school located in a country where English is spoken as the native language.

Another study was conducted in an English as a Foreign Language (EFL) context with native Arabic speakers enrolled in a 15-week academic English course. Results showed that at the end of the program, 92% of the test takers increased their KITE overall level scores. The average increase in score was 60 points, and 58% of the students moved up to the next CEFR level or higher. This study shows that instruction has a significant effect on KITE scores at the 15-week or 300-hour mark when studying at an English language program located in a country where English is not the primary language.

These initial studies show that KITE is an instructionally sensitive English language assessment, as instruction has a measurable effect on KITE scores in both ESL and EFL contexts. KITE researchers intend to continue to measure the effects of instruction in larger, more robust studies spanning multiple countries and contexts.

4. Validity and Consequences of Testing

Early results from the operationalization of KITE suggest that placement decisions made for students based on their KITE assessment scores are consistent with good instructional practice and correspond with the perceptions and judgments of students' classroom teachers and program administrators. In a recent feedback survey, when asked how much they agreed or disagreed with the statement that the KITE Main Flight accurately profiles students' individual language skills, 96% of the Kaplan International administrators and academic staff who were surveyed either agreed or strongly agreed. This indicates that KITE can be used to accurately place students into a program of study that requires the leveling of ability.

Based on the evidence of instruction on test scores mentioned in the previous section, KITE can be used for effectively level testing and progress testing with a language program. However, it is important to recognize that KITE assesses general language proficiency and is therefore not sensitive enough to be used for measurement of progress on specific skills within a short amount of instructional time. It is recommended that KITE be administered at intervals during which test takers can make reasonable progress in their overall language proficiency. Current research suggests this is approximately 200 hours of guided instruction.

Also, it is important to note that research indicates that it generally takes longer to progress as language learners move up the scale, and that length of time needed to progress from one level to the next can vary depending on other factors, including: language learning background, intensity of study, age, and exposure to English outside of the classroom (Desveaux, 2013). This should be taken into consideration when progress testing in the ESL versus EFL context, as well as when assessing progress for higher-level students.

Future Investigations

The KITE team realizes that the process of developing and compiling validity evidence for an assessment product is a continuous one, and we continue to engage in research studies to provide such evidence, such as relating the KITE assessment to other measures of the same or similar constructs. In addition, we continue to enhance and deepen the item pools used in these assessments on a regular basis through our embedding process, thereby strengthening the content coverage and increasing the adaptive efficiency of the assessments. Further, differential item functioning (DIF) analyses are scheduled to occur once sufficiently large subgroup sample sizes have taken the assessment. As with validity, establishing the reliability of an instrument requires providing evidence. In the case of reliability, the evidence pertains to the consistency of the measure, and there are numerous ways to indicate the extent to which the measure consistently operates. In item response models, the degree of measurement error in the score estimates is of the greatest importance.

Standard Errors of Measurement and Separation Indices

KITE uses a Rasch model to analyze item response data. Within the Rasch model, the standard error of measurement is an important aspect of identifying the degree of precision and consistency of estimating student ability. Estimates are affected by many factors, including how well the data fit the underlying model, student response consistency, student location on the ability continuum, matching of items to student ability, and test length. Although there are no specific targets for observed standard errors, lower values of standard errors are preferable to higher values. The mean standard error for subjects in the calibration sample taking the KITE assessment was .38 logits, which is well within the anticipated range for measurement precision.

Other relevant measures provided by WINSTEPS (Lincacre, 2006) in Rasch analysis are separation indices and reliability estimates. Closely related to reliability estimates, separation indices reflect the ratio of person (or item) standard deviation to the standard deviation of error (Wright, 1996). Values above 2.0 indicate that greater than 80% of the variance in scores is not due to error, but rather due to person or item differences. For the KITE assessment, all separation indices are greater than 2.0 for both person and item separation. For the more common reliability measures, equally impressive results were obtained. For person reliability, which is equivalent to the more commonly recognized test reliability in classical test theory settings (such as the KR-20 internal consistency reliability coefficient), a value of .81 was obtained, exceeding the desirable threshold of .80. Rasch analysis also provides item reliability—or the ratio of true item variance to observed item variance—which has no direct counterpart in classical test analysis. For the KITE assessment, these values were quite high, above 1.00, indicating a high level of consistency of item ordering.

Performance Assessment Scoring

Writing and speaking responses are accessed electronically by trained raters who score the responses using analytic scales derived from the CEFR. To ensure maximum reliability, performance assessment raters are selected from ESL/EFL professionals with significant teaching experience and subject matter expertise. Raters attend a standardized webinar training and individually rate a sample set of responses. All raters are reviewed and monitored to ensure accuracy and consistency in scoring. Raters also receive additional training annually to maintain scoring reliability.

A preliminary study was conducted for the KITE writing assessment scoring, which reported very high reliability. The study included three trained raters independently scoring 30 randomly selected sample responses using KITE's seven-level analytic scale. At least two of the three raters agreed on the final overall level of the response on 100% of the responses and all three raters agreed on 57% of the responses. For responses where one rater disagreed, raters never deviated by more than one level. This indicates that after a standardized training, KITE analytic scales can be used to score writing performance reliably.

Speaking performance reliability studies and generalizability analyses are currently underway and will be included in the next version of this report.

Fairness

Providing a fair and unbiased evaluation of an individual's English proficiency is embedded within KITE's mission statement and core values. As a result, test developers strive for fairness in every step of KITE design, development, administration, and use.

ADA Compliance

The system design meets the requirements for Americans with Disabilities Act (ADA) accessible web applications. All efforts have been made to follow Web Content Accessibility Guidelines (WCAG) 2.0 requirements to make the web accessible to people with a wide range of disabilities, including visual, physical, and cognitive disabilities. Note that captions for audio items have not been provided because of the nature of the application (language test).

KITE Content Reviews

In item development, all item writers are trained to avoid a list of topics that would put any test taker at a disadvantage. KITE item writers have extensive experience working with ELLs from diverse backgrounds and from a wide range of nationalities, so they are particularly attuned to possible cultural topics that may have an effect on test takers. In addition to avoiding problematic topics, each item goes through multiple reviews for bias by assessment subject matter experts (SMEs). Items are rewritten and put through the review process again if there is any content that may be considered biased.

Calibration with Diverse Populations

Items are calibrated with thousands of test takers who represent a diverse population that includes a very wide range of demographics and language backgrounds. Over 50 languages are represented in calibration testing, primarily Arabic, Japanese, Spanish, Portuguese, Korean, Chinese, French, Turkish, German, Russian, and Thai, plus several other languages represented in small numbers. The calibration student population is comprised of 47% female and 53% male test takers. Field testing and calibrating items using a representative sample of test takers who reflect the diversity of the larger population helps KITE developers identify items that may be problematic for certain groups of people.

Ongoing Feedback on Use

Future fairness studies will be conducted to analyze the impact of KITE on specific groups. Currently, 97% of clients surveyed agreed that KITE is a fair assessment in that it is not biased toward a particular gender, nationality, or age group.

Scoring and Levels

Scoring—the KITE Main Flight: Listening, Reading, and Grammar

KITE reports an overall score for the assessment, as well as scores for each skill section. The overall score is calculated using a formula that analyzes performance on every item of the assessment. Overall scores are reported in a range of 0–700. The individual skill section scores are calculated based only on performance within each skill section. Each individual skill section is also scored in a range of 0–700. Unlike many traditional assessments, the overall score is not a sum or average of the individual skill section scores. The assessment gathers information and analyzes overall performance and individual skill performance simultaneously.

KITE Overall Level

The KITE overall level score suggests that the test taker is currently performing at the level estimated. The overall level combines the test taker's scores for the listening, reading, and grammar only. This means that the overall demonstrated proficiency corresponds with the minimum requirements of the level proficiency described and interpreted from the CEFR. However, it is very common that a test taker's English proficiency profile in the listening, reading, grammar, writing, and speaking sections is not consistent throughout. For example, a student may have B1 as their overall level; however, he or she may perform higher or lower than B1 in individual sections. See the following overview of KITE overall scores aligned to CEFR levels.

Table 3. KITE Score Ranges and CEFR Levels

KITE Score	Level	Overall Description
535+	C2 (Proficient)	Can use English very fluently, precisely, and sensitively in most social, academic, and professional contexts
500-534	C1 (Advanced)	Can use English fluently and flexibly in a wide range of social, academic, and professional contexts
425-499	B2 (Higher Intermediate)	Can use English effectively, with some fluency, in a wide range of familiar social, academic, and professional contexts
350-424	B1 (Intermediate)	Can communicate important points with some detail in familiar social, academic, and professional contexts
275-349	A2 (Lower Intermediate)	Can communicate simply in English within a limited range of familiar everyday contexts
225-274	A1 (Elementary)	Can communicate in basic English with help and patience from the listener or reader in everyday routine contexts
0-224	Pre-A1 (Beginner)	Can use very familiar everyday expressions and very basic phrases to interact within very limited routine contexts

**Chart modified from Introductory Guide to the Common European Framework of Reference (CEFR) for English Language Teachers (Cambridge University Press, 2013).*

Scoring Productive Sections: Writing and Speaking

Scoring for the writing and speaking sections is done by human raters. These scores are not included in the overall score reported in KITE because they are not scaled using the same IRT psychometrics as the listening, reading, and grammar sections. However, productive section scores are entered into the KITE individual skills report and provide valuable information about the student's English proficiency profile, which may be used to guide placement, progress, or training decisions depending on the organization's emphasis on these skills. The scores are scaled on the same numerical scale as the other sections of the assessment, which allows for easy comparisons and interpretations of test taker strengths and weaknesses.

Scoring for writing and speaking responses may be done by Kaplan raters or organizations may opt to use their own internal raters. These raters receive the same training and rating scales as Kaplan raters.

Test Administration and Score Reports

KITE assessments are administered by institutions or organizations licensing KITE for their own purposes of evaluating English proficiency. Therefore, these organizations and institutions are responsible for ensuring that test taker performance on the test is not influenced by external factors that may cause inconsistent or unreliable results. KITE provides detailed instructions on test delivery, including specifications on testing environment, equipment, test event scripts, and recommended external security measures. The adaptive nature of the assessment provides intrinsic security in that each individual test taker's test is personalized based on how he or she responds to each item, and items are pulled from a large item bank, which increases the likelihood that each test taker receives a unique test.

Score Reports

Immediately after test takers finish a test event, they see a report that profiles their demonstrated English ability. The report includes an overall score and CEFR level, as well as scores and CEFR levels for each section of the test they took. (KITE Main Flight scores are available immediately, but speaking or writing scores are available after human raters have entered the scores). Each individual section also includes a list of recommended skills for improvement or progress. Skill recommendations include specific skills tagged within the KITE item bank and are based on the ability estimate of a test taker during a test event.

Also on the score report is a graph that shows the relationship between the test taker's estimated ability after each item and the items given to the test taker.

It is important to keep in mind that at the end of a test event, test takers will not be able to go back and review individual items. Instead, they are given an individualized profile of their English ability, with recommendations on how to progress based on their performance.

Test Administration and Score Reports (cont.)

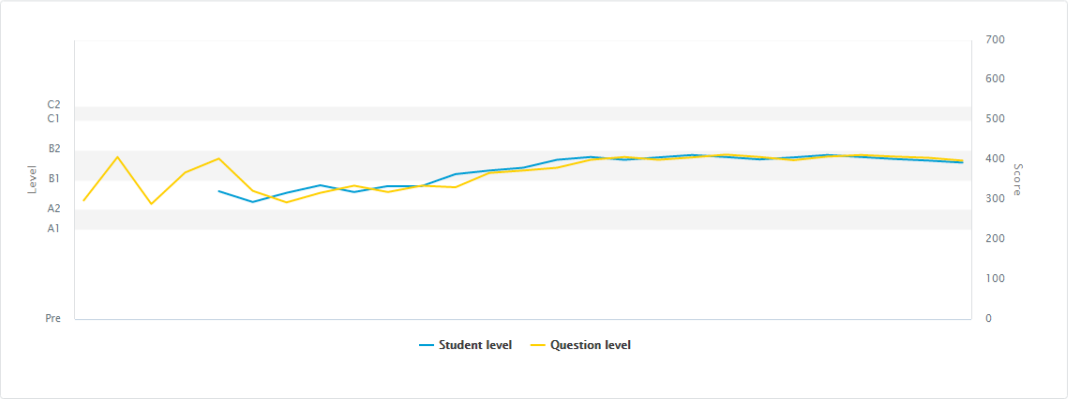
Example Score Report

B1	393	44 mins

Score By Skill

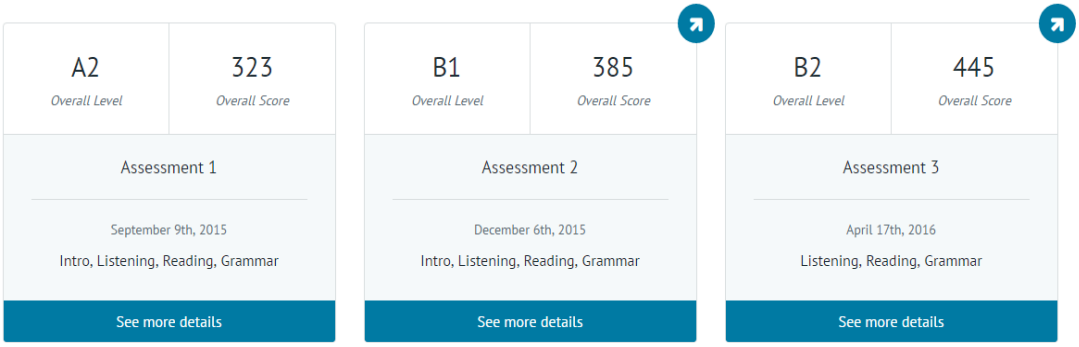
A2	B2	B1
Level	Level	Level
343	477	365
Score	Score	Score
Recommended Skills	*Recommended Skills*	*Recommended Skills*
Overall Listening Comprehension	Identifying Cues and Inferring	Phrasal Verbs
Listening to Audio Media and Recordings	Reading for Information and Argument	Discourse Markers
Understanding Conversation Between Native Speakers	Overall Reading Comprehension	Future
MORE	MORE	MORE

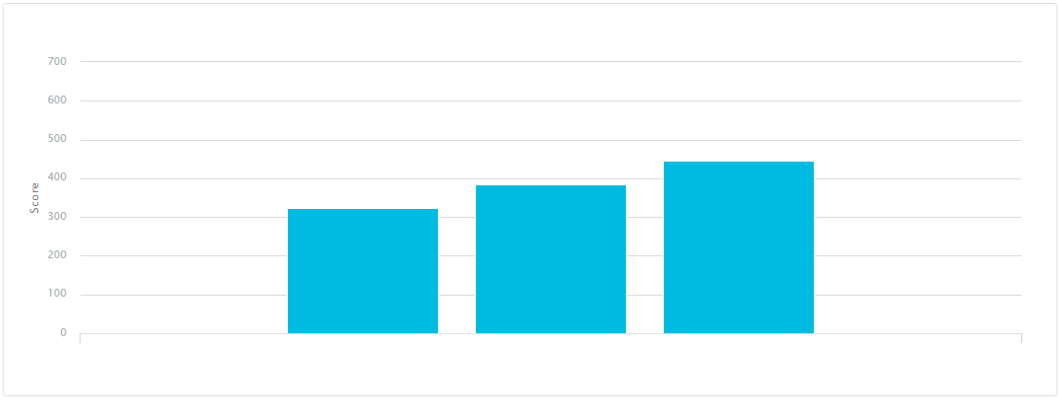
Assessment Graph



Test Administration and Score Reports (cont.)

Example Progress Report (Administrator View)





Progress reporting is designed to give assessment administrators an overview of student performance through multiple testing events.

Interpreting KITE Scores

KITE provides scores and CEFR levels that can be used by organizations and institutions to make decisions (e.g., for recruitment, placement, progress). Because the proficiency requirements for individual clients may vary, it is important that organizations using KITE set their own scoring criteria or cut-off points to suit their context. There are many methods for setting standards that can be used to develop individualized cut scores. KITE offers consultation on best practices with standard setting for individual clients as needed.

Other Information

Tested Population Characteristics

KITE is currently used for placement, progress, and exit testing by a diverse group of students in the ESL and EFL contexts. The following information represents an overview of test takers who took KITE within the ESL context for 2015.

Figure 2 represents the region of national origin reported for tested students. A breakdown of the primary countries represented in each region is as follows: Western Europe—Italy, France, Germany, Turkey, Spain, and Switzerland; Asia—Japan, South Korea, China, Taiwan, and Thailand; Latin America—Brazil, Colombia, Venezuela, Mexico, Argentina, Chile, and Peru; Middle East—Saudi Arabia, Oman, and Kuwait; Eastern Europe—Russia, Czech Republic, Poland, Kazakhstan, Slovakia, Ukraine, and Hungary; Africa—Libya, Angola, Morocco, Tunisia, Jordan, Egypt, Ivory Coast, and Algeria. Figure 3 represents the tested population gender breakdown, and Figure 4 represents the age range of the tested population.

Figure 2: Tested Population National Origin Region

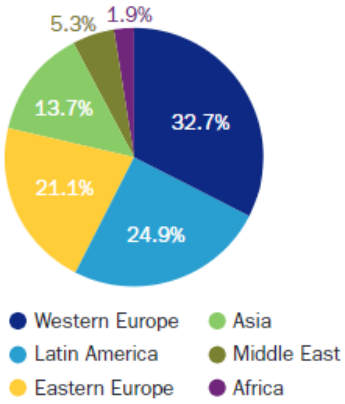


Figure 3: Tested Population Gender

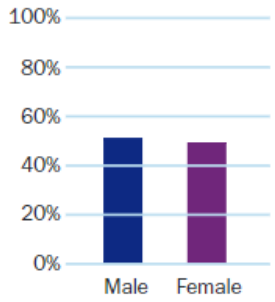
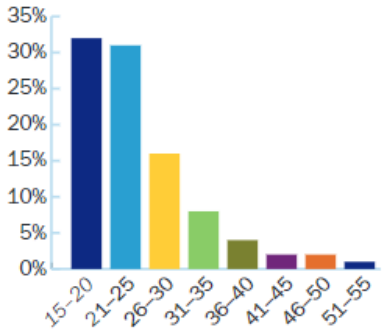


Figure 4: Tested Population Age



Using KITE for Program Accreditation

KITE meets the standards of practice for assessment used for placement, level progression, and program completion within an English language program for major domestic accrediting bodies such as the Accrediting Council for Continuing Education and Training (ACCET) and the Commission on English Language Accreditation (CEA). It also meets the standards for reliable, valid, and fair assessment practices for international organizations such as the European Association for Quality Language Services (EAQUALS); English Language Intensive Courses for Overseas Students (ELICOS, Australia); Languages Canada; and the Private Career Training Institutions Agency (PCTIA, British Columbia). In addition, KITE meets the criterion for quality endorsements as provided by independent organizations such as National ELT Accreditation Scheme Limited (NEAS, Australia).

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association.

Cambridge University Press (2013). Introductory Guide to the Common European Framework of Reference (CEFR) for English Language Teachers. Cambridge: Cambridge University Press. Available online: http://www.englishprofile.org/images/pdf/GuideToCEFR.pdf

Council of Europe (2001). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge: Cambridge University Press. Available online: http://www.coe.int/t/dg4/linguistic/source/framework_en.pdf

Council of Europe (2011). Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR): a manual. Strasbourg: Council of Europe, Language Policy Division. Available online: https://www.coe.int/t/dg4/linguistic/Source/ManualRevision-proofread-FINAL_en.pdf

Desveaux, S. (2013, July 17). Guided learning hours. Posted to https://support.cambridgeenglish.org/hc/en-gb/articles/202838506-Guided-learning-hours

Linacre, J. M. (2006). WINSTEPS Rasch measurement computer program. Chicago: Winsteps.com.

North, B. (2014). English Profile Studies: The CEFR in Practice. Cambridge: Cambridge University Press.

North, B., Ortega, A., & Sheehan, S. (2010). British Council — EAQUALS Core Inventory for General English. British Council and EAQUALS. Available online: http://englishagenda.britishcouncil.org/sites/ec/files/books-british-council-eaquals-core-inventory.pdf

Wright B. D. (1996) Reliability and Separation. Rasch Measurement Transactions 9:4 p. 472.Xi, X. (2010).

How do we go about investigating test fairness? Language Testing, 27(2), 147-170. Available online: http://ltj.sagepub.com/content/27/2/147.abstract

Appendices

KITE Main Flight			
Section	**Format**	**Task Types**	**Focus**
Listening	• Multiple choice • Adaptive • Untimed (Average test time: 73 minutes) • Average number of total items for Listening, Reading & Grammar: 54	• Conversations between native speakers • Announcements & instructions • Lectures & presentations • Radio broadcasts	Identifying & understanding: • main ideas • specific information • opinion & argument • vocabulary in context • contextual cues & inference
Reading		• Personal emails / letters • Business emails / letters • Newspaper / magazine articles • Blog posts • Reviews • Instructions • Announcements & notices • Letters to the editor • Reports	Identifying & understanding: • main ideas • specific information • opinion & argument • vocabulary in context • references & relationships between ideas • contextual cues & inference
Grammar		• Simple sentences • Short dialogues • Short emails & messages • Short descriptions & reports	Identifying & understanding form, meaning & use of: • verb forms & tenses • discourse markers & linkers • question forms • gerunds & infinitives • modals • nouns & noun clauses • prepositions (phrases & clauses) • articles & other determiners • adjectives & adjective clauses • adverbs • intensifiers • possessives • conditionals • phrasal verbs • passives

Appendices (cont.)

KITE Productive Skills (The Extras) *			
Section	**Format**	**Task Types**	**Focus**
Writing	Task 1 Written Interaction • Time: 25 min. Task 2 Written Production • Time: 35-40 min. Scored by human raters	• Personal & business emails • Letters to the editor • Forms & applications • School writing tasks (class surveys, descriptive paragraphs, opinion/ argument essays)	Task 1: • Initiating & responding to invitations & requests • Giving information • Making plans, suggestions & recommendations • Using appropriate correspondence conventions & sociolinguistic skills Task 2: • Filling out a form • Giving factual information • Describing people, places, habits & routines • Developing an argument & justifying opinions • Explaining pros & cons • Speculating causes & effects • Expressing abstract ideas
Speaking	3-5 spoken production tasks Approximate total time: 10 min. Prompts may include charts, graphs, photographs & other visual stimuli. Scored by human raters	• Introductions • Answering questions • Describing & comparing charts, graphs & photographs • Giving a talk to peers	• Answering questions • Giving factual information • Explaining likes & dislikes • Describing daily life • Describing people, places, habits, routines, etc. • Making comparisons • Describing experiences • Giving an extended description • Expressing opinions • Developing an argument & justifying opinions • Explaining pros & cons • Comparing & contrasting • Making suggestions & recommendations • Speculating • Storytelling

*Administered after the KITE Main Flight

Acknowledgments

We would like to acknowledge the following individuals for their seminal and continued roles in the development of KITE:

Rich Brown, PhD, Psychometrician and Founder/CEO of West Coast Analytics; Chief Research Officer at National Math and Science Initiative

David Niemi, PhD, Vice President of Measurement and Evaluation at Kaplan Inc.

Li-Ann Kuan, PhD, Senior Director of Assessment and Test Development at Empowering Education Services