KAPLAN INTERNATIONAL

Tools for English

Adaptive Assessment Development and Validation

Kaplan International Tools for English Technical Report
Version 2.2 March 2022

Table of Contents

Overview

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

Educational institutions, businesses, and government agencies worldwide can use Kaplan International Tools for English as a fair, valid and reliable assessment tool for:

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

Kaplan International Tools for English offers a range of assessment combinations to suit each institution's needs. Kaplan International Tools for English's Main Flight assesses listening and reading ('Skills' option) or listening, reading and grammar ('Extra' option). The Main Flight can be followed by the assessment of writing and/or speaking skills.

Kaplan International Tools for English is closely aligned with the internationally-recognized Common European Framework of Reference (CEFR) proficiency levels. Kaplan International Tools for English accurately assesses communicative proficiency skills ranging from beginning to advanced.

To pinpoint where an individual's language skills are along a seven-level learning continuum, Kaplan International Tools for English 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.

IRT and Adaptive Assessment Design

Item Response Theory (IRT), the foundation for Kaplan International Tools for English, is widely used in education and other fieldsto: 1) calibrate and evaluate items in tests, questionnaires, and other instruments; and 2) scoretest 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 Kaplan International Tools for English. 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. Kaplan International Tools for English 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 sameexact set of items.

Kaplan International Tools for English Features

Kaplan International Tools for English also has many other important advantages over traditional fixed tests:

- A computerized adaptive assessment system like Kaplan International Tools for English provides more efficient testing and more accurate test results because it chooses items that specifically target each test taker'sestimated ability level.
- Kaplan International Tools for English provides instant test results on the Main Flight which allow institutions to make informed decisions immediately.
- Organizations can administer assessments in their own setting and on their own schedule because Kaplan International Tools for English is a cloud-based system.
- Kaplan International Tools for English recommends skill areas for improvement at each ability level.
- Kaplan International Tools for English 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.
- Kaplan International Tools for English is aligned with the Common European Framework of Reference (CEFR) and is scientifically equated to the IELTS® exam.
- Cheating is reduced because each test taker sees different sets of items.
- Kaplan International Tools for English is more cost-effective, because it is paperless and operates on internet-enabled devices and does not require downloaded software.
- Test takers can more accurately show their linguistic knowledge and abilities because the Main Flight assessments are untimed.
- Only highly effective test items are used in the assessments. Kaplan International Tools for English 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 Kaplan International Tools for English 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 Kaplan International Tools for English 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 Kaplan International Tools for English 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 Kaplan International Tools for English, the driving construct is English languageproficiency 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.

As Wilson (2005) and others have argued, documentation of the process used to develop a test is a key component of evidence for the validity of test content: "… documentation of the steps taken…constitute a thorough representation of the content validity evidence for the instrument. It also lays the foundation for the remaining aspects of validity evidence—in a sense, the evidence related to content is the target of the remaining evidence." (Wilson, 2005, p. 156–157). The rigorous and well-documented process used to develop test content aligned to the CEFR is summarized below.

Alignment to the CEFR

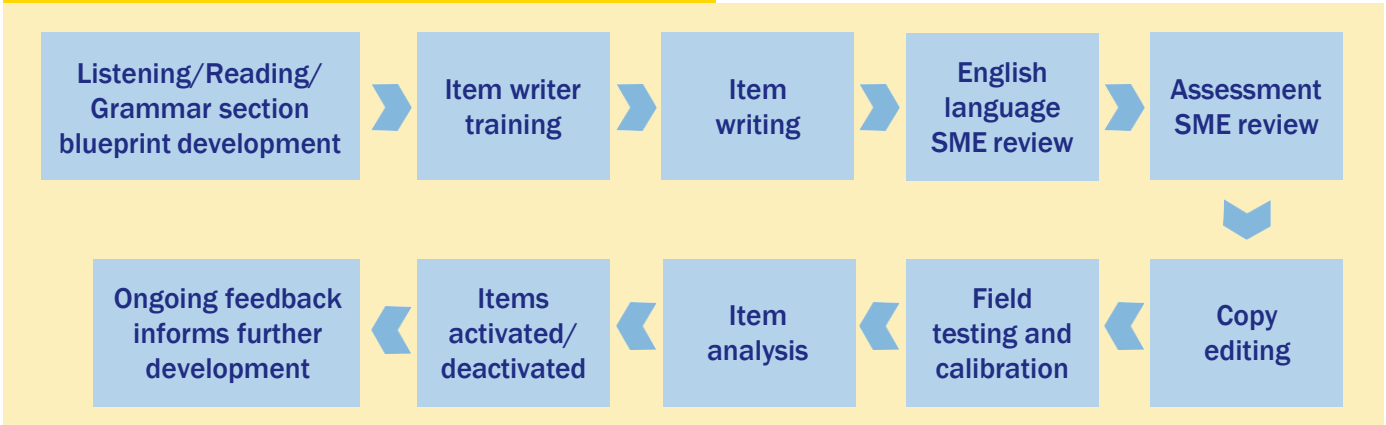
To operationalize the CEFR can-do statements and the Core Inventory into test objectives, language experts at Kaplan International English used the competencies and skills embedded in each descriptor or language point to define how test takers would be expected to demonstrate relevant abilities on an 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, Kaplan International Tools for English content assesses listening, reading, grammar, writing, and speaking skills that are closely aligned with the CEFR framework. To see the Kaplan International Tools for English content and structure overview, see Appendices (pages 20-24).

Item Development Process

The experts at Kaplan International English have developed a rigorous and comprehensive process for creating and reviewing all test items for Kaplan International Tools for English to ensure the highest quality. The processbegins with the careful development of item specifications as mentioned in the CEFR alignment process above. After test objectives are blueprinted, a team of writers is recruited and trained.
Each writer for Kaplan International Tools for English is required to have prior experience with teaching ELLs, which ensures that they are able to fully understand the characteristics of the population, and are able to anticipate common errors made by ELLs and incorporate them into items. This results in items that are moreeffective at discriminating test taker's abilities and providing better diagnostic information. In addition, before writing any test items, each writer must undergo intensive training around (a) the research-based criteria for writing high-quality test items, (b) the test objectives, and (c) the CEFR proficiency levels.

Throughout the item-writing process, Kaplan's team of ELL and assessment experts review every item for quality and accuracy. Collectively, these experts ensure that each item is (a) aligned to the test objectives, (b) clearly worded, (c) accurate, (d) at the appropriate level of difficulty, and (e) free of bias based on culture, gender, and socioeconomic status. Before field testing can begin, each item is copy edited to ensure that it does not contain any typographical or spelling errors.
The following flow chart (figure 1) summarizes the Kaplan International Tools for English Listening, Reading, and Grammar item
development process:

Figure 1. Main Flight Item Development Process



Field Testing and Calibration

Before items are deployed on the testing platform, they are field tested and administered to a diverse group of students—approximately 200 students per item. Typically, the items are organized into testforms of 50–60 items each, including 10 common or anchor items that have been pretestedand pre-calibrated. Items written for a specific proficiency level are administered to Kaplan
International English students who are at the corresponding and adjacent proficiency levels. For example, an item intended to measure a skill at the B1 level is administered to students who have demonstrated language proficiency at the A2, B1, and B2 levels.

The calibration process allows us to determine the difficulty level (and other IRT parameters) for new items and ascertain whether they can be placed on the same scale as other items. Items that scale appropriately are then added to the database of items used by Kaplan International Tools for English; other items are evaluated to see whether they should be modified and retested, or discarded.

Writing and Speaking Section Development

The writing section is administered after the system generates an estimate of the test taker'soverall ability level based on the performance on the listening, reading, and grammar sections.
Based on this estimate, the system presents one or more writing tasks from one of four levels (Pre, A, B, C). The speaking section is administered after the writing section; the system selects one of four levels of test forms (Pre, A, B, C) based on the test taker's Main Flight ability estimate. Each form consists of 3–4 spoken production tasks. The writing and speaking responsesare scored by human raters using analytic scales.

The writing and speaking sections were developed from a blueprint based on the CEFR descriptive scales and Core Inventory for General English. The positive can-do statements were broken down and interpreted into performance outcomes for the tasks. In addition, the CEFR and the Core Inventory for General English were the basis for setting parameters on the input, context, and communicative purpose of the prompts in the productive sections. The rating scales used to assess the communicative proficiency of the test taker's response were also derived from thesubskills illustrated in the CEFR and the assessment grids provided in the Manual for Relating Examinations to the CEFR (Council of Europe, 2011). The rating scale assessment criteria for the writing section include Overall Response, Language Accuracy and Range, and Coherence and Cohesion; the assessment criteria for the speaking section include Overall Response, Linguistic Accuracy and Range, and Fluency. Both of these rating scales include key terms from the CEFR descriptors from levels A1–C2. The Pre-A1 level scales were derived from actual test taker responses that fell below the A1-level criteria. The flow chart that follows (figure 2) illustrates the item writing and rating scale development cycles for the productive sections (e.g., writing). The process is similar for speaking items.

Figure 2. Productive Section Development Process (Writing)

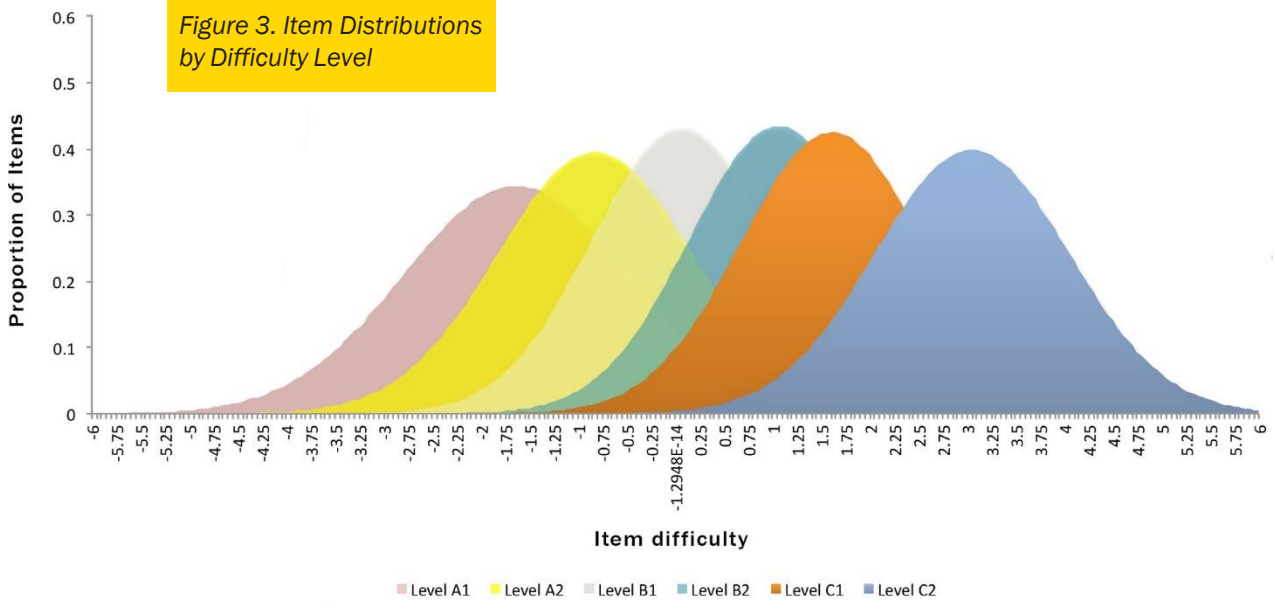
Continuous Development

Test developers at Kaplan International Tools for English are constantly reviewing current items and adding new items to the extensive bank. After inventories are conducted, new items are written for skills and topics that are less represented in the bank. In addition, older items which may be out of date or overusedare retired. Furthermore, the Kaplan International Tools for English team intends to continue developing new item types. As new technologies are developed within Kaplan International Tools for English, new task types will be introduced. Because Kaplan maintains an open and systematic feedback loop with its clients, Kaplan International Tools for English developers are responsiveto client needs. As a result, Kaplan International Tools for English is continuously growing, improving, and innovating.

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. It is widely understood that language learners do not acquire skills in a linear fashion, nor do any two learners acquire the same skills in the same order. Furthermore, in a formal education setting, it is recognized that there is content overlap across levels as learners master skills while they progress. However, what should be apparent from the estimated item difficulties is that, in general, 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 (figure 3). This plot provides support that, by and large, items targeting progressively higher ability levels are progressively more difficult, which in turn provides evidence that Kaplan International Tools for English test items effectively operationalize the underlying construct map. Thus, validity evidence to support the internal structure of the assessments is provided as the item calibrations(e.g., difficulty estimates) support the construct map and item design.

Figure 3. Item Distributions by Difficulty Level



3. Relationships to Other Measures

Kaplan International Tools for English 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 Kaplan International Tools for English assessment, a linking study was conducted with the IELTS exam. The results of the linking study indicated strong positive relationships between each component score and the existing measures to which they were linked.

A linear regression of IELTS on Kaplan International Tools for English produces a strong and significant prediction between Kaplan International Tools for English and IELTS scoring, as the correlations between the Main Flight score and IELTS overall scoreis .792. In addition, the correlation between the Listening subscales on the Kaplan International Tools for English and IELTS is .645, and the correlation between the Reading subscales on the Kaplan International Tools for English and IELTS is .700. Two hundred and two examinees were included in this study, of which 40% were female and 60% weremale, 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 datafrom the study with corresponding CEFR levels can be found below (table 1).

Table 1. Sample Comparison of Kaplan International Tools for English & IELTS

Kaplan International Tools for English	IELTS
560	8.0
527	7.5
495	7.0
459	6.5
426	6.0

Kaplan International Tools for English and TOEFL®

Because we have equated Kaplan International Tools for English and IELTS, and IELTS and the TOEFL iBT exam have been previously equated, we can predict TOEFL scores from Kaplan International Tools for English scores.

Evidence of Effect of Instruction on Test Scores

Kaplan uses Kaplan International Tools for English for placement, progress, and exit testing at its international English languageschools. In addition, Kaplan International Tools for English is used for placement/diagnostic and progress testing at other institutions that teach English internationally. Several studies have been conducted to evaluatethe effect of instruction on Kaplan International Tools for English test scores.

One study showed that 92% of test takers who were enrolled in an English language program increased their Main Flight 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 Kaplan International Tools for English 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 Main Flight 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 Kaplan International Tools for English scores at the 15-week or300-hour mark when studying at an English language program located in a country where Englishis not the primary language.

These initial studies show that Kaplan International Tools for English is an instructionally sensitive English language assessment,as instruction has a measurable effect on scores in both ESL and EFL contexts.

4. Validity and Consequences of Testing

Results from the operationalization of Kaplan International Tools for English in English language programs across the globe suggest that placement decisions made for students based on their Kaplan International Tools for English assessment scores are consistent with good instructional practice and correspond with the perceptions and judgments of students' classroom teachers and program administrators. When asked how much they agreedor disagreed with the statement that the Main Flight accurately profiles students' individual language skills, 96% of administrators and academic staff either agreed or strongly agreed. This indicates that Kaplan International Tools for English 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, Kaplan International Tools for English can be used for effectively level testing and progress testing with a language program. However, it is important to recognize that Kaplan International Tools for English assesses general language proficiency and is therefore not sensitive enough to be used for measurement of progress on specific skills within a short amountof instructional time. It is recommended that Kaplan International Tools for English 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.

It is also 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.

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

Kaplan International Tools for English 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 Kaplan International Tools for English 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 Kaplan International Tools for English 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 Kaplan International Tools for English 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.

To determine the degree of inter-rater reliability, intraclass correlation coefficients (ICCs) were computed for Kaplan International Tools for English Writing and Speaking assessments. For the Kaplan International Tools for English Writing assessment, three trained raters independently scored 30 randomly selected sample responses using Kaplan International Tools for English's seven-level analytic scale, and an ICC estimate was computed based on a single-rater, absolute-agreement, 2-way random effects model. For the Kaplan International Tools for English Speaking assessment, 16 randomly selected responses were graded by three randomly selected trained raters, and an ICC estimatewas computed based on a single-rater, one-way random effects model. Both studies returned an ICC categorized as good to excellent (for Writing, ICC (2,1) = 0.88; for Speaking, ICC (1,1) = 0.87),

indicating a high level of reliability in Writing and Speaking assessments.
Providing a fair and unbiased evaluation of an individual's English proficiency is embedded within Kaplan International Tools for English's mission statement and core values. As a result, test developers strive for fairness in every step of 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).

Kaplan International Tools for English 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. Kaplan International Tools for English 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 problematictopics, 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 Kaplan International Tools for English developers identify items that may be problematicfor certain groups of people.

Main Flight:

Kaplan International Tools for English reports an overall score the Main Flight and scores for each skill section. The overall score is calculated using a formula which analyzes performance on every item of the Main Flight assessment. Overall scores are reported in a range of 0-700. The scores for the individual skills sections (listening and reading ('Skills') or listening, reading and grammar ('Extra')) 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.

Main Flight Overall Level:

The Main Flight overall level score suggests that the test taker is currently performing at the level estimated. The overall demonstrated proficiency corresponds with the minimum requirements of the level proficiency described and interpreted from the CEFR. See the following overview of Main Flight overall scores aligned to CEFR levels (Table 2).

Table 2. Kaplan International Tools for English Score Ranges

Kaplan International Tools for English Score	CEFR Level	Overall Description
535+	C2 (Proficient)	Can use English very fluently, precisely, and sensitively inmost 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 infamiliar social, academic, and professional contexts
275-349	A2 (Lower Intermediate)	Can communicate simply in English within a limited range offamiliar everyday contexts
225-274	A1 (Elementary)	Can communicate in basic English with help and patiencefrom the listener or reader in everyday routine contexts
0-224	Pre-A1 (Beginner)	Can use very familiar everyday expressions and very basicphrases 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).*

Grading Productive Sections: Writing and Speaking

Grading for the writing and speaking sections is done by human raters. Productive section scores are entered into the individual skills report and provide valuable information about the student's English proficiency profile, which may be used to guide score interpretation 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.

Cut Scores

To set cut scores for the Main Flight, a sample of 821 students was drawn from those who were performing successfully in courses at each level, and a set of items was administered to the students at each level. (Altogether, 888 items were administered.) Kaplan International Tools for English researchers then examined: (1) the distributions and mean ability estimates (based on Kaplan International Tools for English) for each level, (2) the distributions and mean item difficulties for each level, (3) the overlap between these distributions,and (4) the probability that a student at the mean ability level for each level would correctly answeran item of mean difficulty for that level. From these analyses, cut scores were derived that were used to place students into course levels. Analysis of successful versus unsuccessful placements then led us to make slight adjustments to the cut scores to improve level scaling accuracy.

Reliability of Cut Scores

The following chart (table 3) shows mean standard errors for the Kaplan International Tools for English ability estimates within 10scale points on either side of each cut-point for the levels. This shows that the standard error is consistent across the ability continuum and is, thus, reliable.

Table 3. Cut Score Reliability

Cut score(+/- 10 points)	Level (lowest end)	Mean Standard Error for Ability Estimates within Range
525-545	A1 (Elementary)	.315
490-510	A2 (Lower Intermediate)	.302
415-435	B1 (Intermediate)	.316
340-360	B2 (Higher Intermediate)	.319
265-285	C1 (Advanced)	.328
215-235	C2 (Proficient)	.327

Test Administration and Score Reports

Client Administration

Kaplan International Tools for English assessments are administered by institutions or organizations licensing Kaplan International Tools for English for their ownpurposes 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.

The assessment can be taken in a test center or remotely. For test center testing, Kaplan International Tools for English provides detailed instructions on test delivery, including specifications on testing environment, equipment, test event scripts, and recommended external security measures. For remote testing, the Kaplan International Tools for English Team offers documentation and consultation on security maximization and support for test takers and administrators.

The Kaplan International Tools for English assessment is accessed by both test takers and administrators via a URL which is unique to their organization. Test takers are required to enter login details to gain access to tests and score reports. Administrators similarly enter login details to gain access to the administration dashboard.

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 thelikelihood that each test taker receives a unique test.

Kaplan Administration

Kaplan International Tools for English provides a referral service to organizations who would like to use the assessment to test English language proficiency but do not wish to administer the assessments themselves. The Kaplan International Tools for English Team manages communication and payment processes, sets and grades assessments, offers test taker support and provides certificates and assessment reports.

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 Main Flight score and CEFR level, as well as scores and CEFR levels for each section of the test. (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 Kaplan International Tools for English item bank and are based on the ability estimate of a test taker during a test event.

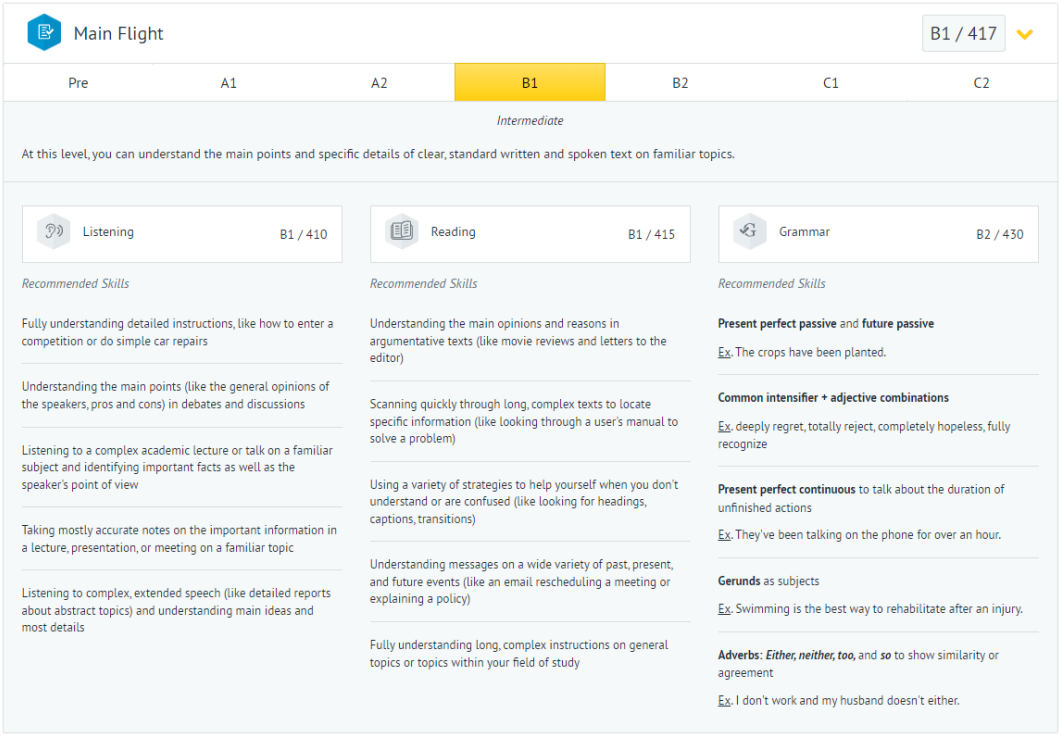
Also on the score report is a graph that shows the relationship between the test taker's estimated ability after each item and the items given to the test taker.

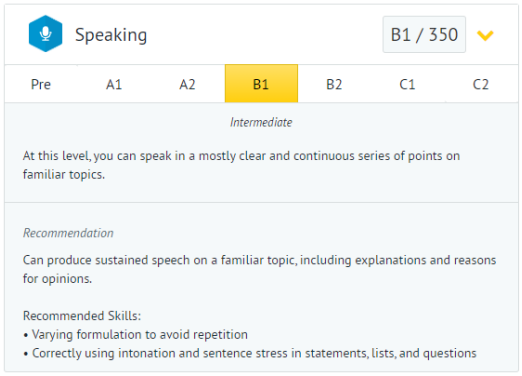
It is important to note that, at the end of a test event, test takers will not be able to go back and review individual items. Instead, they are given an individualized profile of their English ability with recommendations on how to progress based on their performance.

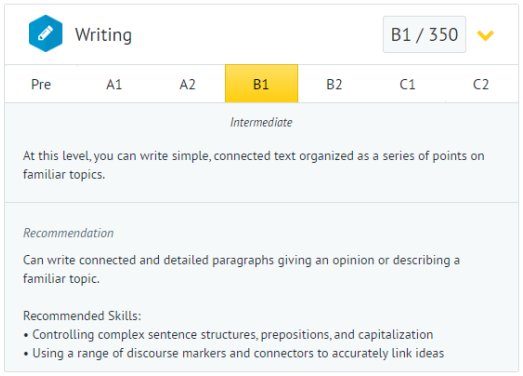
Kaplan International Tools for English Score Reports may be made unavailable to test takers at the

request of test administrators who may wish to communicate the scores and feedback in a different form such as a certificate.

Example Score Report:

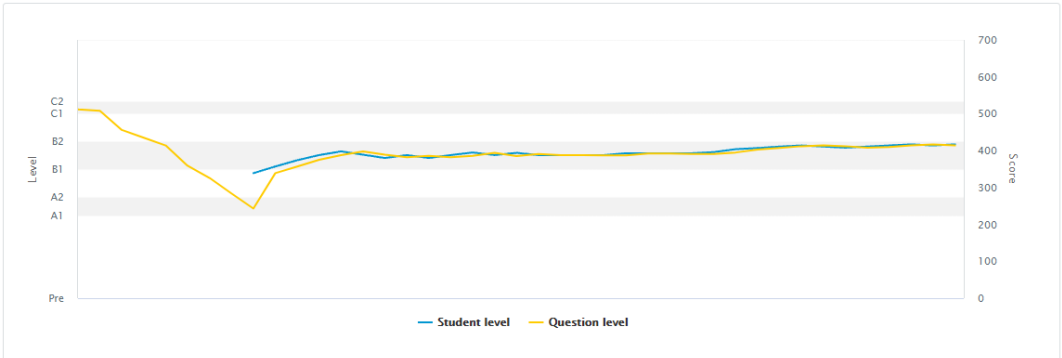




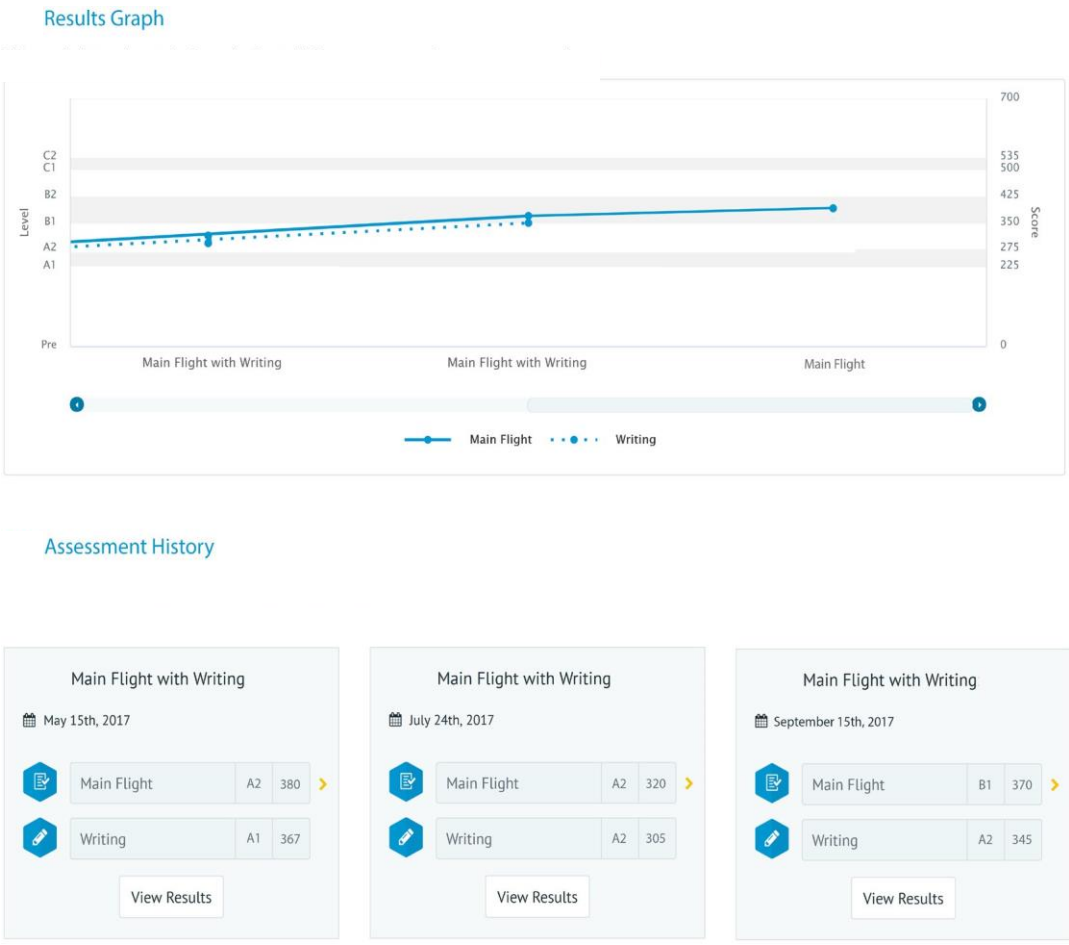


Assessment Graph

The KITE assessment gives you questions that match your level.



Example Progress Report (Administrator View)



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

Interpreting Kaplan International Tools for English Scores

Kaplan International Tools for English 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, organizations are encouraged to set their own scoring criteria or cut-off points to suit their context. The Kaplan International Tools for English Team offers consultative support on these decisions.

Tested Population Characteristics

Kaplan International Tools for English 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 Kaplan International Tools for English within the ESL context for 2015.

Figure 4 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 5 represents the tested population gender breakdown, and Figure 6 represents the age range of the tested population.

Figure 4: Tested Population National Origin Region

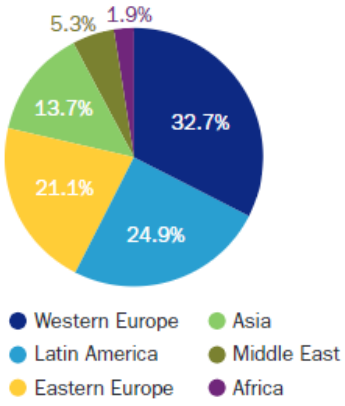


Figure 5: Tested Population Gender

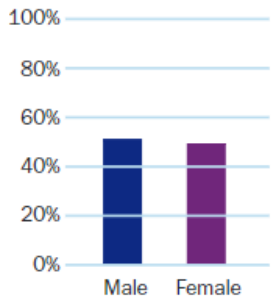
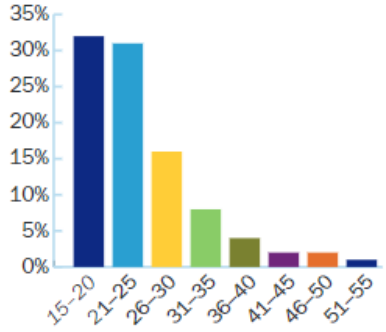


Figure 6: Tested Population Age



Using Kaplan International Tools for English for Program Accreditation

Kaplan International Tools for English meets the standards of practice for assessment used for placement, level progression, and program completion within an English language program for major accrediting bodies suchas 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 QualityLanguage Services (EAQUALS); Languages Canada; and the Private Career Training Institutions Agency (PCTIA, British Columbia). In addition, Kaplan International Tools for English meets the criterion for quality endorsements asprovided by independent organizations such as National ELT Accreditation Scheme Limited (NEAS, Australia), and has received the NEAS Premium Product Endorsement.

Practice Activities

Kaplan International Tools for English offers a range of sample and practice tests to familiarize test takers and administrators with the assessment. These tests use separate item banks from the full assessment to ensure test item security.

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

Main Flight	
Listening Section	
Format	• Computer adaptive • Untimed
Task Type	• Four-option multiple-choice and cloze items • Tasks include 1–4 questions • Test takers can listen to each passage twice
Focus	Depending on ability level, identifying and understanding: • Main ideas • Specific information • Opinion & argument • Vocabulary in context • Contextual cues and inference
Content	Content: • Includes topics, broad language functions, and vocabulary from level descriptors • Ranges from concrete to abstract, depending on level • Length, speed, and articulation of listening passages determined by level Depending on ability level, passages may include: • Informal conversations between native speakers • Announcements and instructions • Lectures and presentations • Radio broadcasts • Debates and discussions • Interviews • Commercial advertisements

Main Flight	
Reading Section	
Format	• Computer adaptive • Untimed
Task Type	• Four-option multiple-choice and cloze items • Tasks include 1–4 questions
Focus	Depending on ability level, identifying and understanding: • Main ideas • Specific information • Opinion and argument • Vocabulary in context • References and relationships between ideas • Contextual cues and inference
Content	Content: • Includes topics, broad language functions, and vocabulary from level descriptors • Ranges from concrete to abstract, depending on level • Length and discourse type determined by level Depending on ability level, passages may include: • Personal emails/ letters • Business emails / letters • Newspaper, magazine, or journal articles • Blog posts • Reviews • Announcements and notices • Instructional manuals and materials • Letters to the editor • Reports • Advertisements • Signs and regulations • Text messages

Main Flight	
Grammar Section	
Format	• Computer adaptive • Untimed
Task Type	• Four-option multiple-choice and cloze items
Focus	Depending on ability level, identifying and understanding form, meaning, and use of: • Verb forms and tenses • Discourse markers and linkers • Question forms • Gerunds and infinitives • Modals • Nouns and noun clauses • Prepositions (phrases and clauses) • Articles • Pronouns • Determiners • Adjectives and adjective clauses • Adverbs • Intensifiers • Possessives • Conditionals • Phrasal verbs • Passives
Content	Content: • Includes topics, broad language functions, and vocabulary from level descriptors • Ranges from concrete to abstract, depending on level • Length and discourse type determined by level Item types include: • Single sentences • Short conversations • Short speeches • Short paragraphs

Productive Skills	
Writing Section: Correspondence Task	
Format	• Typed email or letter to specified audience • Leveled task (Pre, A, B, or C) generated by Main Flight level • Time: 25 minutes • Scored by human raters
Task Type	Written Interaction: • Personal and business emails • Letters to the editor
Focus	Depending on ability level: • Giving information and instructions • Answering questions • Initiating and responding to invitations and requests • Making plans, suggestions, and recommendations • Using appropriate correspondence conventions & sociolinguistic skills
Content	• Includes topics, broad language functions, and vocabulary from level descriptors • Ranges from only concrete to mostly abstract, depending on level
Writing Section: Descriptive/Essay Task	
Format	• Typed academic writing task • Leveled task (Pre, A, B, or C) generated by Main Flight level • Time: 35-40 minutes • Scored by human raters
Task Type	Written Production: • Short descriptive text (A level) • Opinion or argument essay (B–C levels)
Focus	Depending on ability level: • Giving factual information • Describing people, places, habits, and routines • Developing an argument and justifying opinions • Explaining pros and cons • Speculating about causes and effects • Expressing abstract ideas
Content	• Includes topics, broad language functions, and vocabulary from level descriptors • Ranges from only concrete to mostly abstract, depending on level

Productive Skills	
Speaking Section	
Format	• Set of 3–5 tasks related to a theme, specific audience • Leveled task (Pre, A, B, or C) generated by Main Flight level • Total time: Approximately 10 minutes • Scored by human raters
Task Type	Spoken Production: • Making an introduction • Describing visual stimuli (signs, advertisements, photographs, menus, etc.) • Comparing visual stimuli and expressing opinions • Answering text questions • Giving a talk to peers
Focus	Depending on ability level: • Answering questions • Giving factual information • Explaining likes and dislikes • Describing people, places, habits, and routines • Talking about a personal experience • Making comparisons • Giving an extended description • Expressing opinions • Explaining pros and cons • Making suggestions or recommendations • Guessing and speculating • Giving a presentation • Organizing sustained discourse
Content	• Includes topics, broad language functions, and vocabulary from level descriptors • Ranges from only concrete to mostly abstract, depending on level

Acknowledgments

We would like to acknowledge the following individuals for their seminal and continued roles in the development of Kaplan International Tools for English:

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