EF SET

EF STANDARD ENGLISH TEST

EF SET ACADEMIC AND TECHNICAL DEVELOPMENT REPORT

September 2014

TABLE OF CONTENTS

OVERVIEW OF EF SET

- The EF SET is a standardized objectively-scored test of listening and reading skills. It is designed to classify test takers’ reading and listening performances on the test into one of the 6 levels established by the Common European Framework of Reference (CEFR), a set of common guidelines outlining the expected proficiencies of language learners at 6 distinct levels as indicated in the table below. In addition, EF SET PLUS test takers’ combined reading and listening scores are classified by an internal EF scale from 1 to 100. For more information about EF SET's score scale, visit: www.efset.org/score-scale.
- The EF SET is administered as an adaptive test, using a delivery model known as Computer Adaptive Multi-Stage Testing [ca-MST]. This means that as test takers demonstrate their levels of reading and listening comprehension in real time, the test content is adjusted to measure as accurately as possible at the test takers’ empirical level of English comprehension.

Type of Language User	Level	Code	Description
Basic	Beginner	A1	Understands familiar everyday words, expressions and very basic phrases aimed at the satisfaction of needs of a concrete type
	Elementary	A2	Understands sentences and frequently used expressions (e.g. personal and family information, shopping, local geography, employment)
Independent	Intermediate	B1	Understand the main points of clear, standard input on familiar matters regularly encountered in work, school, leisure, etc.
	Upper intermediate	B2	Understands the main ideas of complex text or speech on both concrete and abstract topics, including technical discussions in his/her field of specialisation
Proficient	Advanced	C1	Understands a wide range of demanding, longer texts, and recognises implicit or nuanced meanings
	Proficient	C2	Understands with ease virtually every form of material read, including abstract or linguistically complex text such as manuals, specialised articles and literary works, and any kind of spoken language, including live broadcasts delivered at native speed

Standard Setting Study
Report July 2014

EF SET DEVELOPMENT IN BRIEF

Why build an EF SET?

EF initially set out to develop a free assessment tool for adults (16+ years) that could assess all four language modalities: listening, reading, writing and speaking. The initial purpose for developing the EF SET was to create a placement/advancement/certification test that would be used internally within the EF environment and then be used as a publicly accessible free standardized test comparable in quality to high-stakes and very expensive tests such as IELTS and TOEFL.

The Path to EF SET

EF began the design process by soliciting the assistance of language assessment experts, and engaging in a formal, highly structured design process[1]. Based on the articulation of an Assessment Use Argument (AUA) and Design Statement, an initial Test Blueprint was created.

- From the outset, the test was designed to be fully automated in both delivery and scoring. EF began with a design for separate tests for the receptive skills of reading and listening. Many different task types and test taker response options were explored during the trial period, which began in May 2012 and continued through nine formal trials for approximately 15 months[2].
- Each of the trials took place in EF's International Language Schools; students in these schools are representative of the target population for the test articulated in the initial design documents, the AUA and the Design Statement.
- Test taker responses to the tasks in each trial were subjected to thorough psychometric analyses, using both classical test theory and Rasch model analyses. The table below records the number of test takers in each trial.

Trial Date	Number of test takers
May 2012	500
July 2012	500
August 2012	500
September 2012	1500
November 2012	1500
January 2013	2500
April 2013	2500
May 2013	2500
July/August 2013	2500

[1] Lyle Bachman served as the seminal adviser for the EF SET development, and his book, Language Assessment in Practice (Bachman, L and Palmer, A (2010) Oxford University Press) provided the essential foundations of the EF SET development. Mari Pearlman, of the Pearlman Education Group LLC, facilitated the path from theory to operations throughout the 15-month trial period, with particular attention to the test development and delivery process.

[2] See the Test Blueprint for a partial list of task types evaluated and trialed.

Data-Based Refinement

After the September 2012 trial, the initial Blueprint was further refined, using the principles of assessment engineering[3], into a much more specific and detailed Task Model Grammar for each modality. Task Model Grammars were successively revised as each subsequent trial yielded new information about item performance. The goal of each post hoc analysis of trial data was to improve the quality and replicability of the EF SET reading and listening tasks in the fully automated test delivery environment.

Assessment engineering principles

- Have governed not only test content and task design, but also the final decision about the operational delivery model, ca-MST, or computer adaptive multi-stage adaptive testing.
- At the end of the 2013 trial period, EF SET final ca-MST delivery specifications were defined. In tandem with test content, the EF SET team built a versatile and robust test delivery platform capable of:
  - Delivering a secure test
  - Generating task templates
  - Safeguarding test content
  - Delivering a test built with an ‘adaptive algorithm’
  - Recording and storing large amounts of data from test sessions securely
  - Transferring test session data for psychometric analysis
- The iterative trials and analyses over the 15-month trial period resulted in several design decisions. These included the final decisions on task types to be included in the operational tests (May 2013), the decision to use Rasch model analyses for trial task analysis (January 2013), calibration of the task pool for the adaptive test panels and operational score reporting and decision making (finished by April 2013), and the decision to make the publicly available EF SET reading and listening assessments multi-stage adaptive tests (ca-MST) (April 2013).
- In June of 2013, EF conducted an initial formal Standard Setting Study to establish the operational definitions for each of the CEFR levels on the beta version of the EF SET. The decisions reached by the first Standard Setting Panel participants were implemented in the multi-stage adaptive EF SET beta test. In July of 2014, EF did a second formal Standard Setting Study to establish the final operational definitions for each of the CEFR levels on the official version of the EF SET and EF SET PLUS, launched in September of 2014. The decisions reached by the 2014 standard setting panel participants have been implemented in the final version of the multi-stage adaptive test, and they govern the reported EF scale score bands for the listening and reading EF SET PLUS test. A scoring matrix, validated by language teaching experts, determines the combined total score for each EF SET Certification test taker.

From its initial launch in October 2013, EF SET has continued to examine the performance of the test items, the operation of the ca-MST, and the functionality of the internet delivery platform. Ongoing validation of the test content, statistical model, and performance standards is an integral part of the EF SET’s management at EF.

1. ASSESSMENT USE ARGUMENT

The Assessment Use Argument (Bachman and Palmer, 2010) is a formal statement of the rationales and justifications for the design and development decisions the test developer makes during the process of creating the test, from conceptualization to operational use. The EF SET AUA, below, was created at the beginning of the design and development process as an explicit and disciplined way to record the reasoning that led to specific design decisions. Because the EF SET, like most standardized tests, produces a score that is meant to support inferences about the language proficiency of the test taker, the “argument” for the trustworthiness of those inferences must be supported. The AUA is intended to demonstrate the foundation for this support in an explicit and incremental fashion

[3] Richard M. Luecht, of University of North Carolina Greensboro, and his colleague John Willse, advised EF on all aspects of the EF SET in its use of assessment engineering, Rasch model analysis, and ca-MST.

Prototype Phase Process Documentation

Stages	Name of Document
Plan	I. Initial Planning Questions
Design	II. Assessment Use Argument III. Design Statement
Develop	IV. Blue Print V. User Specifications VI. IT Specifications
Trial	VII. Back up Evidence during Trial
Use	VIII. Back up Evidence during Use

Background

EF's goal is to assure, maintain, and improve the quality of instruction and learning. To this end, EF is developing a system of tests that will provide information that is useful for learners, teachers, course developers, program administrators and test developers.

- Learners: motivate their learning of English, demonstrate progress and receive confirmation of their achievement in English language ability,
- Teachers: improve their teaching
- Course developers, curriculum designers, and materials writers: improve the quality of the learning activities/materials with which students and teachers interact,
- Program administrators: make decisions about placing students into courses/levels that are appropriate for their English language learning needs.

Assessment Use Argument

The Assessment Use Argument (AUA) is the conceptual framework which explicitly states the rationale and justification for the decisions we make in designing and developing the test.

Consequences

Claim 1: The use of placement/advancement/certification tests and of the decisions that are made based on the test results are beneficial to the stakeholders

Descriptions of the stakeholders

1. Test Takers (EF Students)
2. Sponsors of EF Students
3. EF Instructors
4. EF Course Administrators
5. EF Content Developers
6. EF Test Product Managers
7. Instructors at local schools or corporations
8. Course Administrators at local schools or corporations
9. Managers at local schools or corporations, including HR managers
10. Admission Staff at educational institutions to which EF graduates might apply
11. Potential Employers
12. General Public, including academics, ministries of education, corporations

Warrants: Consequences of using the placement/advancement/certification tests are beneficial

A.1. Consequences of using the placement/advancement/certification tests that are specific to the students (test takers), families, instructors, course administrators, content developers, test product managers, admission staff and employers will be beneficial.

A.2. Scores from the placement/advancement/certification tests of individual students are treated confidentially.

A.3. Scores from the placement/advancement/certification tests are reported in ways that are clear and understandable to students, families, instructors, the course administrator, test product manager, admission staff and employers.

A.4 Scores from the placement/advancement/certification tests are reported in a timely manner.

A.5 The placement/advancement/certification tests help promote good instructional practice and effective learning. The use of this is thus beneficial to students, instructors, course administrators, test product managers, admission staff and employers.

Warrants: Consequences of the decisions made are beneficial

B.1. The consequences of the placement/advancement/certification decisions that are made will be beneficial for each of the following groups of stakeholders:

1. The students studying at EF
2. The sponsor of the student studying at EF
3. The instructors at EF
4. The course administrators at EF
5. The content developer at EF
6. The test product managers at EF
7. The instructors at local schools or corporations
8. The course administrators at local schools or corporations
9. The managers at local schools or corporations, including HR managers
10. The admissions staff at educational institutions to which graduates of EF courses might apply
11. Potential Employers of EF graduates
12. The general public, including academics, ministries of education, corporations

Rebuttal: The consequences of false positive and false negative classification errors are as follows:

1. False positive classification errors:

- Placement and advancement decisions: Placing and advancing students into courses that are more advanced than their actual level will have negative consequences for the students, instructors and course administrators. The students may feel overwhelmed, frustrated and lose confidence. Furthermore, even though these students may be struggling in class, they may not report this. Instructors may also feel frustrated because their students are not able to keep up with assignments and classroom tasks. Both students and instructors may complain to the course administrators who will need to make adjustments to classroom size and timetables.
- Certification decisions: Certifying students at a given CEFR level when they are actually at a lower level of proficiency will have detrimental consequences for the students, admission staff, and potential employers. A certified level of proficiency warrants that the recipient can demonstrate certain language skills. If student’s English ability is lower than indicated, other stakeholders may question the legitimacy of the individual or in some cases suspect cheating. Meanwhile, the admission staff or employer could have rejected other eligible candidates based on the stated English language ability, only to find that the records do not reflect reality after employment decisions have been made or progressed.

2. False negative classification errors:

- Placement and advancement decisions: Placing students into lower level courses will have detrimental consequences for students. Students may feel bored with the coursework and resent the whole experience. They may also suffer because taking the recommended course prevented them from taking different or additional courses.
- Certification decisions: Not certifying students at a given CEFR level when they are actually at a higher level will have detrimental consequences for students and employers. The student may decide to spend and time to take another test or study materials again, even when it is unnecessary. Employers or admission staff may not chose an otherwise perfect candidate because they make the assumption that the candidate's level of English is lower than it actually is, based on the information.

Decisions

Claim 2: The decisions to place students into or allow them to advance from a given level in an EF course or to certify them at a given CEFR level are made with careful consideration of all laws, rules and regulations that may govern the use of the EF test, reflect relevant educational and societal values in each country of operation and are equitable for those students and other stakeholders affected by the decisions made.

The decisions, stakeholders affected by decisions, and individuals responsible for making the decisions are provided in the table below.

Decision	Stakeholders affected by the decision	Individual(s) responsible for making the decision
Place students into appropriate level in an EF course.	Student, Instructor , Sponsors	Course Administrator
Allow students to advance to next level in an EF course.	Student, Instructor , Sponsors	Course Administrator
Adjust curriculum, course materials, and instruction in an EF course.	Student, Instructor, Content Developers	Instructor
Certify at CEFR level	Student, Admission Staff, Employer, Sponsors	Test Product Manager

Warrants: Values Sensitivity

A.1. Relevant educational values, regulations, and legal requirements of the EF corporation and EF programs, as well as relevant educational and societal values in each country of operation are carefully considered when placement, advancement, and certification decisions are made.

A.2. Relevant educational values, regulations, and legal requirements of the EF corporation and EF programs, as well as relevant educational and societal values in each country of operation are carefully considered in determining the relative seriousness of false positive and false negative classification errors.

A.3. Cut scores are set to minimize the most serious classification errors.

- Relative seriousness of classification decision errors: For placement, advancement, and certification decisions, false positive classification errors are more serious than false negative ones.

Warrants: Equitability

B.1. Students taking the EF placement/advancement/certification tests are classified only according to the cut scores and decision rules, and not according to any other considerations.

B.2. Test takers, EF instructors and other individuals within EF community, as well as other relevant stakeholders (e.g., families, admissions staff, potential employers) are fully informed about how the decisions will be made and whether decisions are actually made in the way described to them.

B.3. Test takers will have equal opportunity to learn or acquire the areas of English ability to be assessed.

Interpretations

Claim 3: The interpretations about the students' "reading, writing, listening, and speaking abilities in English" are meaningful in terms of the content of instruction and instructional materials used in EF courses, and of the language use demands of tasks in test takers' TLU domains, and for certification decisions, in terms of the level descriptors in the CEFR, impartial to all groups of test takers, generalizable to instructional content and learning activities in EF courses, and to tasks in students' TLU domains, and relevant to and sufficient for the decisions that are to be made.

Warrants: Meaningfulness

A1. The interpretations about the test taker's integrated skills with emphasis on communicative and functional language use are meaningful with respect to EF course materials, other relevant documents and students' Target Language Use (TLU) domains, which are aligned to the bands and levels defined by the CEFR.

- Definition of the construct is "speaking, listening, reading and writing skills with emphasis on fluent and accurate use of English as a communication tool rather than a field of study", reflecting proper grammar, vocabulary, lexis, pronunciation and spelling.
- The interpretation of the test taker's integrated skills is required to be:
    - A measure of students' levels which is transparent and understandable to the administration, teachers and students.
    - A clear indication of the student's progress and what they should aim to achieve in future.
    - Aligned with regard to the terminology used to describe expectations of the indicated level and range of levels.
    - Correlated with international standards, descriptions and other external examinations
    - Worthy of the certificates which states a student's English language level

A2. The assessment task specifications clearly specify the conditions under which we will observe or elicit performance from which we can make inference about the construct we intend to assess.

A3. The procedures for administering EF placement/advancement/certification tests enable the test takers to perform at their highest levels of "speaking, listening, reading and writing skills with emphasis on fluent and accurate use of English as a communication tool rather than a field of study".

A4a. The algorithm for producing machine rated test scores focus on elements of the construct "speaking, listening, reading and writing skills with emphasis on fluent and accurate use of English as a communication tool rather than a field of study" and are transparent to stakeholders.

A4b. The procedures for producing human rated test scores focus on elements of the construct "speaking, listening, reading and writing skills with emphasis on fluent and accurate use of English as a communication tool rather than a field of study" and provide clear instructions for raters and are transparent to stakeholders.

A5. The EF test tasks engage the "speaking, listening, reading and writing skills with emphasis on fluent and accurate use of English as a communication tool rather than a field of study".

A6. Scores on the EF test can be interpreted as indicators of "speaking, listening, reading and writing skills with emphasis on fluent and accurate use of English as a communication tool rather than a field of study".

A7. The definition of the construct is explained in non-technical language via the instructions for tasks and sample questions and answers. The construct definition is also included in non-technical language in the assessment report for test takers and other stakeholders.

Warrants: Impartiality

B1. The EF test tasks do not include response formats or content that may either favor or disfavor some test takers.

B2. The EF test tasks do not include content that may be offensive to some test takers.

B3a. The procedures for producing assessment reports are clearly described in terms that are understandable to all test takers.

B3b. The procedures for dealing with unanswered or incomplete tasks are clearly described in terms that are understandable to all test takers.

B4. Test takers are treated impartially during all aspects of the administration of the assessment.

- Test takers have equal access to information about the assessment content and assessment procedures.
- Test takers have equal access to the assessment, in terms of cost, location, and familiarity with conditions and equipment.
- Test takers have equal opportunity to demonstrate their knowledge of communication skills in English.

B5. Interpretations of the test takers' "speaking, listening, reading and writing skills with emphasis on fluent and accurate use of English as a communication tool rather than a field of study" are equally meaningful across students from different first language backgrounds and academic disciplines.

Warrants: Generalizability

C1. The characteristics of the EF test tasks correspond closely to a) those of tasks in the test taker's TLU domains, which are closely tied to EF courses and b) language use descriptions outlined in the CEFR.

C2. The criteria and procedures for evaluating the responses to the EF test tasks correspond closely to those that are typically used by other language users in test takers' TLU domains and are aligned to the CEFR.

Warrant: Relevance

D. The interpretation "speaking, listening, reading and writing skills with emphasis on fluent and accurate use of English as a communication tool rather than a field of study" provides the information that is relevant to the EF Course Administrators' decisions about placement and exemption, and to the Test Product Manager's decisions about certification.

Warrant: Sufficiency

E. The assessment-based interpretation of "speaking, listening, reading and writing skills with emphasis on fluent and accurate use of English as a communication tool rather than a field of study" provides sufficient information to make the placement, advancement, and certification decisions.

Assessment Records

Claim 4: The scores from the EF test are consistent across different forms and administrations of the test and across students from different cultural, linguistic, or socio-economic backgrounds.

Warrants: Consistency

1. The EF test is administered in a standard way every time it is offered.
2. The scoring criteria and procedures for human ratings and the computer scoring algorithm are well specified.
3. The computer scoring algorithm was developed through extensive trialing and comparison with multiple human ratings.
4. The computer scoring algorithm was developed through trialing with several different groups of test takers.

5. Raters undergo training and must be certified.

6. Raters are trained to avoid bias for or against different groups of test takers.

7. Scores on different items are internally consistent.

8. Ratings of different raters are consistent.

9. Different ratings by the same rater are consistent.

10. Scores from different forms of the EF test are consistent..

11. Scores from different administrations of the EF test are consistent.

12. Scores on the EF test are of comparable consistency across different groups of students.

2. DESIGN STATEMENT

The Design Statement is closely related to the Assessment Use Argument. In the Design Statement, the rationales and justifications for the claims made in the AUA and the warrants for those claims, are viewed from the perspective of the production of test tasks, the trialing of those tasks, the analysis of the resulting data, and the interpretation of results in light of the AUA claims and warrants. The Design Statement formally poses the questions that require the test developer to consider available resources, operational constraints, consistent rules and procedures for both task development and task delivery and scoring. The Design Statement is the initial practical engagement with the development of the test.

Phase 0 (Prototype)
Process 2 (Design Statement)

Stages	**Name of Document**
Plan	IX. Initial Planning Questions
Design	X. Assessment Use Argument XI. Design Statement
Develop	XII. Blue Print XIII. User Specifications XIV. IT Specifications
Trial	XV. Back up Evidence during Trial
Use	XVI. Back up Evidence during Use

Background

EF's goal is to assure, maintain, and improve the quality of instruction and learning. To this end, EF is developing a system of tests that will provide information that is useful for learners, teachers, course developers, program administrators and test developers.

- Learners: motivate their learning of English, demonstrate progress and receive confirmation of their achievement in English language ability,
- Teachers: improve their teaching
- Course developers, curriculum designers, and materials writers: improve the quality of the learning activities/materials with which students and teachers interact,
- Program administrators: make decisions about placing students into courses/levels that are appropriate for their English language learning needs.

Design Statement

The Design Statement (DS) is a guide for test developers information that backs the warrants in the AUA.

1. Description of Test Taker and Stakeholders

Stakeholders	Attributes
1. Test Takers (EF Students)	Kids and Teens aged 7-15 Adults aged 16-25+ with different native languages and English levels ranging from A1-C1. Enrolled in varying levels of courses, and come from a wide range of educational, linguistics, cultural, and social backgrounds. Enrolled in courses that vary from 2 to 52 weeks depending on individual needs and interests. Can receive varying degrees and intensity of lessons per week.
2. Sponsors	Sponsors provide the student support to study at any one of EF's programs. Often provides financial resources and / or emotional support for students to attend course at EF. Expecting a return on investment in the form of improved language ability by the student.
3. EF Instructors	EF teachers in classroom and online teach students, assign homework and grades student's work. Interacts with each student anywhere from 2 to 32 times a week, depending on type of course taught (general vs special interest). Teachers qualifications range from CELTA (bachelors) at a minimum to DELTA (diploma) or Masters degrees, meeting all accreditation standards.
4. EF Course Administrators	EF school directors and director of studies. These persons typically have DELTA, Masters or Doctoral degree qualifications and substantial years of teaching experience. In charge of school operations, staff management, classroom assignments and academic counseling for students when necessary.
5. EF Content Developers	EF employees who design and write content suitable for each age group and medium of delivery. Typically a combination of project manager and, full time staff and freelancers. Content developers typically have at least masters or doctoral degree and suitable teaching experience.

2. Intended Beneficial Consequences (Claim 1)

Stakeholders	Intended Beneficial Consequences	
	Of Using the assessment	Of the Decisions made
1. Test Taker (EF Student)	Will realize that the test tasks are similar to instructional tasks, and thus relevant to their target language use (TLU) needs. Will have been tested in a way that is consistent with ways in which their performance during and after the course is being evaluated.	Will benefit from being placed in a course they need. Will have documented evidence of their current proficiency.
2. Sponsors	Will be able to track progress and see return on investment. (If test results are shared by the test taker with the student's families)	Can be involved more actively in helping the EF student learn and make progress (If test results are shared by the test taker with the student's families)
3. EF Instructors	Will benefit from using a test in which the criteria for making placement decisions are similar to those used in making decisions about the effectiveness of their instruction. Will better understand the areas of strength and weakness in their instruction. Will be able to improve their instruction to more effectively meet the needs of their students.	Will benefit from being able to focus their instruction on a group of students who are relatively homogeneous in their English language ability. Can pay closer attention to each individual student in class.
4. EF Course Administrators	Will benefit from using a test whose scoring criteria are consistent with the performance objectives for the course they supervise.	Will have to deal with fewer complaints from bored or frustrated students and frustrated teachers.

Stakeholders	Intended Beneficial Consequences	
	Of Using the assessment	Of the Decisions made
5. EF Content Developers	Will better understand the areas of strength and weakness in the course materials they develop and making pacing decisions. Will be able to improve the course materials they develop to more effectively meet the needs of the students.	Will be able to monitor effectiveness of content more closely and make improvements based on feedback and data.
6. EF Test Product Managers	Will benefit by getting feedback on the test's quality of measurement.	EF test product managers can make improvements to the EF test engine based on feedback and data.
7. Local Instructors	Similar to EF Instructors	Similar to EF Instructors
8. Local Course Administrators	Similar to EF Course Administrators	Similar to EF Course Administrators
9. Local Managers	Will be able to track usage snap shot across multiple users, skills and criteria. Will be able to track progress over time across multiple users, skills and criteria.	Base system wide decisions on purchasing educational service based on data Base system wide decisions on purchasing educational service based on data
10. Admission Staff	Will receive clear and informative score report and gain confidence in what that scores mean	Can make judgment and admissions decisions with confidence on candidates' English language abilities
11. Potential Employers	Will receive clear and informative score report and gain confidence in what that scores mean	Can make judgment and employment decisions with confidence on candidates' English language abilities
12. General Public	Will have clear and thorough information about how the assessment was developed and for what purpose.	Can reference our research findings Can contribute to further research topics

3. Description of Decisions to be made (Claim 2)

Decision	Stakeholders affected by the decision	Individual(s) responsible for making the decision
Place students into appropriate level in an EF course.	Student, Instructor	Course Administrator
Allow students to advance to next level in an EF course.	Student, Instructor	Course Administrator
Provide diagnosis of language skills and show progress over time	Student, Instructor	Course Administrator
Adjust instruction in an EF course.	Student, Instructor	Instructor
Certify at CEFR level	Student, Admission Staff, Employer	Test Product Manager

4. The Relative Seriousness of Classification Errors, Policy-Level Decisions about Standards, The Standards themselves (Claim 2, Warrants A2 and A3)

- Relative seriousness of classification errors: False positive classification decisions are relatively more serious than false negative classification decisions.
- Policy-level procedures for setting standards: The standard scoring mechanism and cut-off scores were set by the EF Test Product Managers in consultation with industry expert, content developers, instructors and course administrators.
- Standard for placing/advancing/certifying students: Test taker must demonstrate that they are able to successfully communicate in English as described by the standards set by EF and in relation to the CEFR.

5. Definition of The Construct (Claim 3, Warrant A1)

- “Speaking, listening, reading and writing skills with emphasis on fluent and accurate use of English to interpret and express personal experiences, ideas, themes and conflicts embodied in all oral, written, and visual texts. Focusing on English as a communication tool rather than a field of study.”
- This definition of the construct is based on the following:
  - Test taker’s TLU domains based on extensive survey of English language users, instructors and policy makers
  - EF course materials, and other relevant documents, which correspond to the test taker’s TLU domains and are aligned to the content descriptions of the bands and levels defined by the CEFR

6. Description of the TLU Domain (Claim 3, Warrant C1)

Test takers’ Target Language Use Domains are broadly based as the learners go on to a wide variety of situations, ranging from further education in English to working in an international company or applying to academic programs.

- The table below shows the TLU domains that correspond to the different language programs at EF. The EF syllabus and proficiency levels correspond to the TLU domains and the CEFR framework.

Category	Target Language Use Domain	Lang	B2C	B2B	B2S	E1 Kids	E1 Smt
General English Adults	• Informal interactions with other users of English in social and professional settings	•	•	•	•		•
Primary and Secondary School (K-12)	• Informal social interactions with other users of English in preparation for long-term personal, academic and career goals.				•	•	
Higher Education	• Undergraduate and graduate academic research, presentations, seminars and other academic endeavors.	•			•		
Professional	• Business negotiations, meetings, international conferences and other professional interactions in English			•			•

7. TLU Tasks Selected as a Basis for Developing Assessment Tasks

For illustration purposes, below is one task from each domain which is relevant to the test taker’s TLU.

- Listening: Understanding conversation between native speakers
- Reading: Reading for information and argument
- Speaking: Describing experience in a monologue
- Writing: Note Taking

8. Description of The Characteristics of TLU Tasks that have been selected as a Basis for Assessment Tasks

	Question	Options
SETTINGS	**The circumstance under which each test takes place**	
Physical Location	Where will the test be taken?	[Classroom, Any location
Medium	How will the test be delivered?	[Stationary Computer, Mobile Device, Paper]
Participants	Other than the test taker, who else will be present during the test?	[None, Invigilator, Other test takers]
	Question	Options
RUBRIC	The context in which the tasks are performed	
Instructions	Specification of structure, procedures to be followed by test takers, and procedures for producing assessment records	
Language	What language will the instruction be in?	[Native, English, Both]
Channel	How will the instruction be presented?	[Aural, Visual, Text]
Explicitness	How much information will the test taker get about recording their response?	[Text, Example Question, Video]
Structure	**Test structure**	
Number of parts	How many parts will the exam have?	[Reading, Listening, Grammar, Writing, Speaking)
Number of tasks per part	How many tasks should there be per part?	[Custom]
Salience of parts/tasks	How distinguishable should the parts be for the test taker?	[Separated, Continuous]
Sequence of parts/tasks	In what order should the parts (tasks) be presented?	[Fixed, Random, Custom]
Relative importance of parts/ tasks	Should the parts (tasks) be weighted differently?	[Equal, Zero, Custom]
Time Allotment	Amount of time for the test, parts, tasks.	
	How much time should there be per Test / Part / Task?	[Minutes]
Recording Method	Outcome of assessment process	
Type of Assessment Record	How will the test taker be informed of test outcome?	[Number Score, Description]
Criteria for Correctness	What components of language ability will be scored and how will the scores be assigned?	[ ]
Procedure for producing an assessment record	Who (what) will do the recording?	
Where will it take place?		
What sequence will be followed?	[Human, Computer]	

	Question	Options
INPUT	**Materials the test takers are expected to process and respond to.**	
Format		
Channel	How will the test takers receive information related to the question?	[Aural, Visual, Both]
Form	Is the information received language based or non-language based or both?	[Language, Non-language, Both]
Language	Is the information related to the test question in their native language, target language or both?	[Native, Target, Both]
Length of Time	How much time is given to process the input?	[Minutes]
Vehicle	Is the information related to the test question live or reproduced?	[Live, Reproduced, Both]
Rate of Processing	What is the rate at which the information is presented?	[Speed]
Type	What type of input is it?	[Input for Interpretation, Item, Prompt]
Language	What is the nature of the language presented to test takers?	
Organizational Characteristics	How are utterances or sentences organized?	
	Grammatical Vocabulary Syntax Phonology / graphology	
	Textual Cohesion Rhetorical or Conversational	
Pragmatic Characteristics	How are utterances or sentences related to communicative goals of the language user and the setting?	
	Functional Ideational Manipulative Heuristic Imaginative	
	Sociolinguistic Genre Dialect/Variety Register Naturalness Cultural References Figure of Speech	
Topical Characteristics	What topic is referred to?	[Personal, Cultural, Academic, Business, Literature, Technical etc.]

	Question	Options
EXPECTED RESPONSE		
Format		
Channel	How will the test takers receive information related to the question?	[Aural, Visual, Both]
Form	Is the expected response language based or non-language based or both?	[Language, Non-language, Both]
Type	What type of expected response is it?	[Selected, Limited Production, Extended Production]
Language	Is the expected response to the test question in their native language, target language or both?	[Native, Target, Both]
Length of Time	How much time is given per expected response?	[Minutes]
Rate of Processing	What is the rate at which the expected response should be executed?	[Speed]
Language	What is the nature of the language of the expected response?	
Organizational Characteristics	How are utterances or sentences organized?	
	Grammatical Vocabulary Syntax Phonology / graphology	
	Textual Cohesion Rhetorical or Conversational	
Pragmatic Characteristics	How are utterances or sentences related to communicative goals of the language user and the setting?	
	Functional Ideational Manipulative Heuristic Imaginative	
	Sociolinguistic Genre Dialect/Variety Register Naturalness Cultural References Figure of Speech	
Topical Characteristics	What topic is referred to?	[Personal, Cultural, Academic, Business, Literature, Technical etc.]
	Question	Options
RELATIONSHIP BETWEEN INPUT AND RESPONSE		
Type of external interactiveness	What is the relationship between input and expected response? (feedback provided on relevance or correctness of response, response affects subsequent input)	[Reciprocal, Non-Reciprocal, Adaptive]
Scope	What is the amount or range of input that must be processed in order for the user to respond as expected?	[Reciprocal, Non-Reciprocal, Adaptive]

	Question	Options
Directness	How much of the answer is in the input versus found in context or through background topical knowledge?	[Direct (loads), Indirect(small)]

3. TEST BLUEPRINT

The Test Blueprint, the third in the series of foundational documents for developing a new test, takes all of the procedural, operational, and practical questions articulated in the Design Statement, and formally organizes them into a specification document for test developers. In the sample of the initial EF SET Test Blueprint below, for test tasks intended for CEFR levels B1 or B2, a very broad range of characteristics of the test taker, the test delivery methodology, and the test content are articulated. At this initial stage of development, many different task types are articulated in the section labeled "item format." Most of these task types were developed and trialed during the 15-month period leading up to the final summer 2013 trials, and the final decisions about the operational EF SET.

EF SET – Reading Test Blueprint, April 2012
Reading: B1-B2

Color Key:
Predictable or specified as fixed characteristics
Characteristics to be specified by Task/Test Developer
Row Heading (not to be filled in)

Test Developer:
Date:
Form:

Attributes of Test takers	
Level of English (EF Level)	• B1 or • B2
Age	16+ (adult) The majority of this target audience is between 16-25.
Gender	Both female and male
L1	Many different L1s
Level of Education	At a minimum, high school level education in their countries of origin. Often students are in undergraduate studies preparing for study abroad or graduate studies in an English speaking country
Purpose for learning English	TTs are interested in furthering schooling or careers. TTs are a language school, learning on-line, or in a blended-learning setting are learning English to communicate in everyday situations and aiming for fluency.
Prior familiarity with equipment	TTs are familiar with the computer and related equipment.

Characteristics of the setting				
Characteristics of the setting				
Place	Computer lab or laptop at home or office			
Equipment	Computer, keyboard, mouse, head phones			
Participants	Test takers			
Time of task	24/7 on internet			
	Task 1	**Task 2**	**Task 3**	**Task 4**
Level of Ability tested (EF level)	• B1 or • B2			
Constructs tested	• Gist • Major ideas • Specific details • Inference • Sens-rhet-org • Sens-cohes	• Gist • Major ideas • Specific details • Inference • Sens-rhet-org • Sens-cohes	• Gist • Major ideas • Specific details • Inference • Sens-rhet-org • Sens-cohes	• Gist • Major ideas • Specific details • Inference • Sens-rhet-org • Sens-cohes
Characteristics of the input				
PASSAGE	**PASSAGE ID #:**	**PASSAGE ID #:**	**PASSAGE ID #:**	**PASSAGE ID #:**
Channel	Visual	Visual	Visual	Visual
Form	• Language • Non-language	• Language • Non-language	• Language • Non-language	• Language • Non-language
Language	English	English	English	English
Text type				
Length	Passage: # words:	Passage: # words:	Passage: # words:	Passage: # words:
Speededness	Unspeeded	Unspeeded	Unspeeded	Unspeeded
Type	Input for interpretation	Input for interpretation	Input for interpretation	Input for interpretation
Vocabulary	*primarily concrete and familiar, non-technical*	*primarily concrete and familiar, non-technical*	*primarily concrete and familiar, non-technical*	*primarily concrete and familiar, non-technical*
Syntax	*primarily declarative, with some compound and occasional complex sentences*	*primarily declarative, with some compound and occasional complex sentences*	*primarily declarative, with some compound and occasional complex sentences*	*primarily declarative, with some compound and occasional complex sentences*
Spelling & punctuation	*Standard American or British spelling and punctuation*	*Standard American or British spelling and punctuation*	*Standard American or British spelling and punctuation*	*Standard American or British spelling and punctuation*
Cohesion Rhetorical organization				
Functions				
Genre (Select from COMPASS or EFFEKTA)				

Variety	*Standard American or British variety, without regional variants*	*Standard American or British variety, without regional variants*	*Standard American or British variety, without regional variants*	*Standard American or British variety, without regional variants*
Register	Formal and informal	Formal and informal	Formal and informal	Formal and informal
Naturalness	Natural	Natural	Natural	Natural
Cultural references and figures of speech				
Topic/theme (Select from COMPASS or EFFEKTA)				
TTs' Prior familiarity with topical content ("Prior familiarity" will be specific to each task, and will include any topical knowledge the task presupposes about the test takers.				
Other (specify) ("Other" will be specific to each task, and will be any other TT attributes that might affect TTs' performance on the task				

ITEM 1	**ITEM ID#:**	**ITEM ID#:**	**ITEM ID#:**	**ITEM ID#:**
Channel	Visual	Visual	Visual	Visual
Form	• Language • Non-language	• Language • Non-language	• Language • Non-language	• Language • Non-language
Language	English	English	English	English
Length	Passage: # words:	Passage: # words:	Passage: # words:	Passage: # words:
Speededness	Unspeeded	Unspeeded	Unspeeded	Unspeeded
Type	Input for interpretation	Input for interpretation	Input for interpretation	Input for interpretation
Item Format	• Matching • Multiple select • Categorization • Completion • Linked short texts • Speaker Match • Gap fill completion • Incomplete outline • (We can add new task types as we develop them)	• Matching • Multiple select • Categorization • Completion • Linked short texts • Speaker Match • Gap fill completion • Incomplete outline • (We can add new task types as we develop them)	• Matching • Multiple select • Categorization • Completion • Linked short texts • Speaker Match • Gap fill completion • Incomplete outline • (We can add new task types as we develop them)	• Matching • Multiple select • Categorization • Completion • Linked short texts • Speaker Match • Gap fill completion • Incomplete outline • (We can add new task types as we develop them)
Vocabulary	primarily concrete and familiar, non-technical	primarily concrete and familiar, non-technical	primarily concrete and familiar, non-technical	primarily concrete and familiar, non-technical

Syntax	primarily declarative, with some compound and occasional complex sentences	primarily declarative, with some compound and occasional complex sentences	primarily declarative, with some compound and occasional complex sentences	primarily declarative, with some compound and occasional complex sentences
Spelling & punctuation	Standard American or British spelling and punctuation	Standard American or British spelling and punctuation	Standard American or British spelling and punctuation	Standard American or British spelling and punctuation
Cohesion				
Rhetorical organization				
Functions				
Variety	Standard American or British variety, without regional variants	Standard American or British variety, without regional variants	Standard American or British variety, without regional variants	Standard American or British variety, without regional variants
Register	Formal and informal	Formal and informal	Formal and informal	Formal and informal
Naturalness	Natural	Natural	Natural	Natural
Cultural references and figures of speech				
Topic/theme	Same as topic of passage.	Same as topic of passage.	Same as topic of passage.	Same as topic of passage.
Characteristics of the expected response				
Channel	Visual	Visual	Visual	Visual
Form (language, non-language, both)	Completion: language All others: non-language	Completion: language All others: non-language	Completion: language All others: non-language	Completion: language All others: non-language
Language (native, target, both)	Completion: English All others: N/A	Completion: English All others: N/A	Completion: English All others: N/A	Completion: English All others: N/A
Length (# of words)	# of words:	# of words:	# of words:	# of words:
Degree of speededness	Unspeeded	Unspeeded	Unspeeded	Unspeeded
Syntax	Completion: primarily concrete and familiar All others: N/A	Completion: primarily concrete and familiar All others: N/A	Completion: primarily concrete and familiar All others: N/A	Completion: primarily concrete and familiar All others: N/A
Vocabulary	Completion: primarily declarative, with some compound and occasional complex sentences All others: N/A	Completion: primarily declarative, with some compound and occasional complex sentences All others: N/A	Completion: primarily declarative, with some compound and occasional complex sentences All others: N/A	Completion: primarily declarative, with some compound and occasional complex sentences All others: N/A

Spelling & punctuation	Completion: Standard American or British spelling and punctuation All others: N/A	Completion: Standard American or British spelling and punctuation All others: N/A	Completion: Standard American or British spelling and punctuation All others: N/A	Completion: Standard American or British spelling and punctuation All others: N/A
Topic	Same or similar to topic of input	Same or similar to topic of input	Same or similar to topic of input	Same or similar to topic of input
Scope of relationship: **Gist: broad** **Specific details: Narrow**				
Directness of relationship **Specific details: direct** **Major ideas: direct** **Gist: relatively indirect** **Inference, prediction: indirect**				
Scope of relationship: **Gist: broad** **Specific details: Narrow**				
Instructions for answering task				
Instructions				
Scoring Method				
Criteria for correctness				
Number of points to be assigned				
How points are assigned				
Maximum points				

ITEM 1	**ITEM ID#:**	**ITEM ID#:**	**ITEM ID#:**	**ITEM ID#:**
Channel	Visual	Visual	Visual	Visual
Form	• Language • Non-language	• Language • Non-language	• Language • Non-language	• Language • Non-language
Language	English	English	English	English
Length	Passage: # words:	Passage: # words:	Passage: # words:	Passage: # words:
Speededness	Unspeeded	Unspeeded	Unspeeded	Unspeeded
Type	Input for interpretation	Input for interpretation	Input for interpretation	Input for interpretation

Item Format	• Matching • Multiple select • Categorization • Completion • Linked short texts • Speaker Match • Gap fill completion • Incomplete outline (We can add new task types as we develop them,)	• Matching • Multiple select • Categorization • Completion • Linked short texts • Speaker Match • Gap fill completion • Incomplete outline (We can add new task types as we develop them,)	• Matching • Multiple select • Categorization • Completion • Linked short texts • Speaker Match • Gap fill completion • Incomplete outline (We can add new task types as we develop them,)	• Matching • Multiple select • Categorization • Completion • Linked short texts • Speaker Match • Gap fill completion • Incomplete outline (We can add new task types as we develop them,)
Vocabulary	primarily concrete and familiar, non-technical	primarily concrete and familiar, non-technical	primarily concrete and familiar, non-technical	primarily concrete and familiar, non-technical
Syntax	primarily declarative, with some compound and occasional complex sentences	primarily declarative, with some compound and occasional complex sentences	primarily declarative, with some compound and occasional complex sentences	primarily declarative, with some compound and occasional complex sentences
Spelling & punctuation	Standard American or British spelling and punctuation	Standard American or British spelling and punctuation	Standard American or British spelling and punctuation	Standard American or British spelling and punctuation
Cohesion				
Rhetorical organization				
Functions				
Variety	Standard American or British variety, without regional variants	Standard American or British variety, without regional variants	Standard American or British variety, without regional variants	Standard American or British variety, without regional variants
Register	Formal and informal	Formal and informal	Formal and informal	Formal and informal
Naturalness	Natural	Natural	Natural	Natural
Cultural references and figures of speech				
Topic/theme	Same as topic of passage.	Same as topic of passage.	Same as topic of passage.	Same as topic of passage.
Characteristics of the expected response				
Channel	Visual	Visual	Visual	Visual
Form (language, non-language, both)	Completion: language All others: non-language	Completion: language All others: non-language	Completion: language All others: non-language	Completion: language All others: non-language
Language (native, target, both)	Completion: English All others: N/A	Completion: English All others: N/A	Completion: English All others: N/A	Completion: English All others: N/A
Length (# of words)	# of words:	# of words:	# of words:	# of words:
Degree of speededness	Unspeeded	Unspeeded	Unspeeded	Unspeeded

Syntax	Completion: primarily concrete and familiar All others: N/A	Completion: primarily concrete and familiar All others: N/A	Completion: primarily concrete and familiar All others: N/A	Completion: primarily concrete and familiar All others: N/A
Vocabulary	Completion: primarily declarative, with some compound and occasional complex sentences All others: N/A	Completion: primarily declarative, with some compound and occasional complex sentences All others: N/A	Completion: primarily declarative, with some compound and occasional complex sentences All others: N/A	Completion: primarily declarative, with some compound and occasional complex sentences All others: N/A
Spelling & punctuation	Completion: Standard American or British spelling and punctuation All others: N/A	Completion: Standard American or British spelling and punctuation All others: N/A	Completion: Standard American or British spelling and punctuation All others: N/A	Completion: Standard American or British spelling and punctuation All others: N/A
Topic	Same or similar to topic of input	Same or similar to topic of input	Same or similar to topic of input	Same or similar to topic of input
Scope of relationship:				
Gist: broad **Specific details: Narrow**				
Directness of relationship **Specific details: direct** **Major ideas: direct** **Gist: relatively indirect** **Inference, prediction: indirect**				
Scope of relationship: **Gist: broad** **Specific details: Narrow**				
Instructions for answering task				
Instructions				
Scoring Method				
Criteria for correctness				
Number of points to be assigned				
How points are assigned				
Maximum points				
ITEM 3 **(same as above)**	ITEM ID#:	ITEM ID#:	ITEM ID#:	ITEM ID#:
ITEM 4 **(same as above)**	ITEM ID#:	ITEM ID#:	ITEM ID#:	ITEM ID#:

4. ASSESSMENT ENGINEERING

Assessment engineering[4] (AE) is an approach to test design, development and operational delivery that imposes principles and practices of manufacturing engineering to the work of creating and maintaining assessments. Building on the kind of evidence-centered design[5] principles that are embodied in documents like the Assessment Use Argument, the Design Statement, and the Blueprint, assessment engineering practices move the design and development process toward much greater specificity and precision about the characteristics of test tasks, their intended performance across the range of test taker abilities, and the interrelationships between these two categories of information and specification.

Three major assertions underlie the assessment engineering approach to test development, analysis, delivery and scoring. The first is that the content of test tasks should vary systematically across the difficulty scale of a test. That means that test developers should quite deliberately specify the differing characteristics of test content in the domain to be tested as those characteristics affect the complexity of the content, and, presumably, the difficulty of the test tasks. The accuracy of these specifications is empirically verified by trialing and analysis.

A second AE assertion is that once both the precise characteristics of the skills needed to perform a task (procedural knowledge) along with the precise characteristics of the task itself—specifications for the declarative knowledge elicited by the task, the auxiliary aids or tools provided to the test taker and the contextual complexity of the task setting—are articulated in a model, and the model is empirically verified as working as intended, an entire family of test items or tasks can be generated from the model, and the performance of those tasks will be consistent. Finally, AE asserts that such model-based task specification and development can lead to item and task templates of sufficient precision to support automated item and task generation.

For EF SET, efforts to implement AE began with an articulation of the difficulty drivers for reading and listening comprehension tasks for English language learners across the CEFR levels from A1 through C1. Below is the initial articulation of the differences in content complexity as they were expected to vary across proficiency levels. Included in this initial schema is the hierarchical data model that governs the management of test taker responses in the database, and subsequent item analysis.

This initial specification of content variability across the target levels of proficiency guided the next phase of task development and trial.

Difficulty levers in assessment tasks for reading and listening

READING	LISTENING
Stimulus materials	Stimulus materials
Amount of text in fixed time	Length of aural stimulus
Increasing complexity of syntax	Increasing complexity of syntax
Increasing breadth and complexity of vocab	Increasing breadth and complexity of vocab
Concrete to abstract topical continuum	Concrete to abstract topical continuum
Familiar to unfamiliar topical continuum	Familiar to unfamiliar topical continuum
	Number and accents of speakers
Questions:	Questions:
Distance from direct reference in stimulus to key	Distance from direct reference in stimulus to key
Closeness of the distractors	Closeness of the distractors

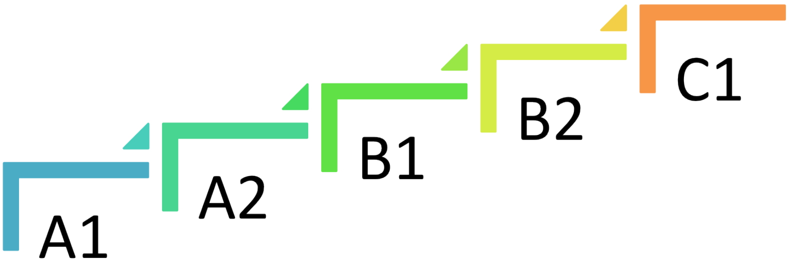
[4] *Luecht R. M. (2012) Assessment Engineering Task Model Maps, Task Models and Templates as a New Way to Develop and Implement Test Specifications, paper presented at NCME, April 2012; Luecht, R.M. (2012). In M. Gierl and T. Haladyna eds, Automatic Item Generation (pp. 57-76) New York: Routledge.*

[5] *A Brief Introduction to Evidence-Centered Design, Mislevy, R., Almond, R.G, Lukas, J.F. (2004), CSE Report 632, National Center for Research on Evaluation, Standards, and Student Testing (CRESST) Center for the Study of Evaluation (CSE) Graduate School of Education & Information Studies University of California, Los Angeles;Mislevy, R. J. and Haertel, G. D. (2006), Implications of Evidence-Centered Design for Educational Testing. Educational Measurement: Issues and Practice, 25: 6–20.*

READING	LISTENING
Questions requiring inference, with inference on rhetorical features (tone, attitude) and structures (conclusion? Preface?) hardest; this is another concrete to abstract continuum	Questions requiring inference, with inference on rhetorical features (tone, attitude) and structures (conclusion? Preface?) hardest; this is another concrete to abstract continuum
Number of questions and complexity of question format (multiple right answers; matching; categorization)	Number of questions and complexity of question format (multiple right answers; matching; categorization)

QUERY: Can we design tasks for reading and listening, using the same task model, that systematically exploit these difficulty features from A1 through C1?

- Task model possibilities:
  - RDNG
    - Single passage with m/c or m/o questions + completions
    - Fixed format two-passage
  - LIST
    - Single speaker with m/c or m/o questions
    - Dialogue with m/c or m/o questions
    - Speaker match



READING

A1	A2	B1	B2	C1
SINGLE PASSAGE/MULTIPLE CHOICE OR MULTIPLE OPTION QUESTIONS WITH COMPLETION B2/C1				
100 words	150-200 words	350 words	450-500 words	450-600 words
4 questions	5 questions	6 questions	6 questions + COMP	8 questions + 2 COMP
TWO PAIRED SHORT PASSAGES + 6 FIXED FORMAT QUESTIONS				
50 word each pssg	100 word each pssg	150 words each pssg	150-200 words each pssg	150-200 words each pssg
4 statements/fixed choices	6 statements/4 fixed choices	6 statements/4 fixed choices	6 statements/4 fixed choices	6 statements/4 fixed choices

LISTENING

A1	A2	B1	B2	C1
MONOLOGUE + M/C OR M/O QUESTIONS + COMP AT B2/C1				
100 words	60 “	60’	90”	240”
	4 M/C questions	4 M/O questions	4 M/O questions + COMP	5 questions + 1 COMP
DIALOGUE WITH MULTIPLE CHOICE				
60 “	60 “	120”	180”	180”
4 M/C questions	7 M/C questions	4 M/O questions	4 M/O questions + 1 COMP	6 FF questions
SPEAKER MATCH seconds by speakers by matches by choices				
90” seconds/4 of 5	120 seconds/4 of 5	180 seconds/5 of 6	220 seconds/5 of 6	400 seconds/5 of 6 x 2

A three-tiered hierarchical data object model will support the three levels of IA reporting functionality and is displayed in Figure 3. Here, we use the generic term data object to refer any level of data within the hierarchy. Further, this hierarchical structure can be mapped to a fairly wide variety of assessment task response-capturing controls including selected-response item types using radio-button group, list-select or check box controls, short-answer input boxes, and even human-scored essays or constructed response items. This hierarchical structure is therefore highly generalizable across item types. Problem sets are the most general level objects and may comprise individual (discrete) items or multiple items (e.g., items associated with a reading passage). In turn, items have one or more scoring objects. These scorable objects can be discrete selections (e.g., multiple-choice selections), text entries, or any compatible response-capturing format that resolves to satisfying Boolean rule (true/false) when the response is compared by some prescribed logic pattern to a scoring evaluator or rubric.

For multiple-choice (MC) items, the scoring objects are the usual distractor options. For example, a one-best answer, four-option multiple-choice item has a correct answer key and three incorrect distractors—a total of four scoring objects. In contrast, a matching/categorization type of item with multiple free form text entries would let the item set represent the collection of text entry boxes. For generality, we can refer to the response-capturing control as a container. Examples of containers include text entry boxes/controls, drag-and-drop types of items with multiple selections dragged by the examinee to one of several boxes or slots, and multiple-select list or checkbox controls. Each container functions as an item. In turn, there are one or more scoring objects associated with each container (item).

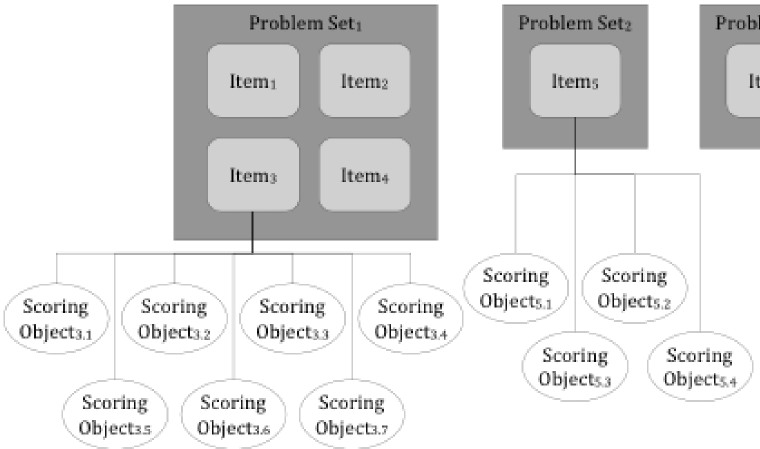


Figure 3. Representation of Data Object Hierarchies for IA

The data types needed to support the data object model shown in Figure 3 is listed in Table 1 for four item types: (a) two or more MC items linked to an item set (passage or other stimulus) and using a correct-answer key (CAK) scoring evaluator; (b) (a) singular (discrete) MC items that use a CAK scoring evaluator; (c) categorization items that require the examinee to enter or drag one or more choices to a box or container ; and (d) free-form, short-answer entries in a text box or container.

TABLE 1. GENERALIZING ITEM SETS, ITEMS AND SCORING OBJECTS ACROSS ITEM TYPES
Identifier and Hierarchical Linking Mechanisms

Item Type	Problem Set Level Data Object	Item Level Data Object	Scoring Object Level
Two or more MC items linked to a common stimulus (passage or exhibit)	Problem Set ID and linked item list	Discrete MC items, scored using CAK evaluator	MC distractor options
Discrete multiple-choice items	Problem Set ID (fixed item set length = 1)	Discrete MC items, CAK evaluator	MC distractor options
Multi-select categorization items	Item ID	Input box ID within item	Selection rule within input box
Multiple short-answer text entry	Item ID	Input box ID within item	Selection rule within input box

5. TASK MODEL GRAMMARS

Assessment Engineering practices replace a traditional test blueprint of the kind created in the initial phases of EF SET development with analytical processes that begin, in the case of a test like EF SET, by articulating the proficiency claims we want to make about test takers at each level of the scale—in the case of EF SET, each CEFR level. These claims articulate with as much precision as possible what test takers know and are able to do in particular language contexts. EF pursued this part of the development work by completing an exhaustive review of all of the EF teaching materials, and then creating a model of the proficiency expectations at each level. The resulting document, a work in progress called the COMPASS (COMmon Principles of Academic Scope and Sequence), maps macro and micro “can do” statements to each CEFR level and connects the associated language functions, grammar, vocabulary, topics, themes, and settings appropriate for each level.

The application of this construct map to assessment design is the second phase of the AE process, and it results in two concrete tools for test construction. The first is a set of task models and task model maps, called Task Model Grammars. These, in turn, lead to task and item templates and a detailed task development manual to guide item writing.

In the excerpt from one of the EF SET Task Model Grammars below, you will see the specification of the declarative knowledge expected from the test taker, the contextual complexity of the stimulus materials and the cognitive demands of the items. The excerpt covers one task type, reading comprehension with multiple-choice questions, across three CEFR levels, A1, A2, and B1.

Note that the task design alters as the difficulty level changes, sometimes quite dramatically. This reflects the basic AE principle, that content is not invariant across the scale, but varies deliberately with the intended measurement target. In this way, difficulty is “designed in” from the outset, as our concept map indicates that what test takers at different points in the language proficiency scale know and are able to do varies systematically, both in terms of the complexity of vocabulary and grammar as well as sheer volume of reading and type of topic and situation.

TASK MODEL GRAMMAR: NO REPETITION ACROSS TABLES—
EACH NEW TABLE SUBSUMES THE TABLE BEFORE OR TASK TYPES CHANGE
Reading at A1

Action/Skills: Simple	**Definitions**
Comprehend	Decode task directions and task/item texts sufficiently to choose correct response at lowest level 30% of the time; mid-level 60% of the time; A1/A2 level 80% of the time
Compare	Brief signs/posters to gist/appropriate context; one brief passage to a second on same topic
Evaluate	Location of information in passage[s]; correctness of each of 4 options as answer to question; connections between posters/signs and statements,
Classify	Statement as "found in passage 1," "found in passage 2," "Found in both passages", statement as not a match to any poster/sign
Choose	Correct option

In social and travel contexts, users at this level can understand simple notices and information, for example in airports, on store guides and on menus. They can understand simple instructions on medicines and simple directions to places.

In the workplace they can understand short reports or product descriptions on familiar matters, if these are expressed in simple language and the contents are predictable.

If studying, they can read basic notices and instructions.

CONTEXTUAL COMPLEXITY OF STIMULUS MATERIALS
Reading at A1

Context: Simple	**Descriptions by task**
Task 3: M/C RCMP	• 100 to 150 word passage • Description/narrative prose • Concrete and familiar topic [constrained by EF Compass] • Vocabulary 85 to 90% in top 1000 • Simple syntax; primarily declarative sentences

Variable elements within tasks

COMPLEXITY OF ITEMS
Reading at A1
Task by Task Item Verbal Components/Variables/Distractor Characteristics
[21 points/1task each type]

RCMP pssg + 8 MC questions	**Item Description[s]**	**Component variables**	**Distractor Characteristics**
	• 4 option M/C questions • 8 questions • Single right answer	• Gist (1) • Specific details (7)	• All distractors are plausible in the overall context of the passage stimulus topic • For 3 of 5 questions, all 4 options come from the passage • For GIST the options are a plausible and connected set, connected to the overall topic of the stimulus • No question can be correctly answered without reading the passage • In general, options must be very short, and answers a single word or phrase • Stems can be phrased cloze style, with a blank inserted
Item/stimulus interaction[s] design	• Elicit RECOGNITION of details appropriate to particular question • Elicit COMPARISON of passage details in order to choose best fit with question • Elicit reasoning beyond the specific sentences in the passage by gist question • Elicit COMPREHENSION of BOTH stimulus and items [note that in RC/MC the total set is regarded and used as the focus of the comprehension assessment]		

Variable elements within tasks

TASK MODEL GRAMMAR:
No repetition across tables—Each new table subsumes the table before OR task types change
Reading at A2

Action/Skills: Simple	**Definitions**
Comprehend	Decode task directions and task/item texts sufficiently to choose correct response at lowest level 30% of the time; mid-level 60% of the time; A2/B1 level 80% of the time
Compare	8 statements to each other and to 6 posters to find a connection of poster to statement; information in linked passages to classify statements
Evaluate	Correctness of reasoning beyond the stated information (inference questions)
Classify	
Choose	Correct option

CONTEXTUAL COMPLEXITY OF STIMULUS MATERIALS
Reading at A2

Context: Simple	**Descriptions by task**
Task 3: M/C RCMP	• 200 – 250 word passage • Description/narrative prose • Concrete and familiar topic [constrained by EF Compass] • Vocabulary 85 to 90% in top 1000 • Simple syntax; primarily declarative sentences

Variable elements within tasks

COMPLEXITY OF ITEMS

Reading at A2

Task by Task Item Verbal Components/Variables/Distractor Characteristics

[18 points/1 task each type]

RCMP pssg + 8 MC questions	Item Description[s]	Component variables	Distractor Characteristics
	• 4 option M/C questions • 8 questions • Single right answer	• Gist (1) • Specific details (3) • Inference (1)	• All distractors are plausible in the overall context of the passage stimulus topic • For 6 of the 8 questions, all 4 options come from the passage • For GIST and INFERENCE the options are a plausible and connected set, connected to the overall topic of the stimulus • No question can be correctly answered without reading the passage
Item/stimulus interaction[s] design	• Elicit RECOGNITION of details appropriate to particular question • Elicit COMPARISON of passage details in order to choose best fit with question • Elicit reasoning beyond the specific sentences in the passage by inference & gist questions • Elicit COMPREHENSION of BOTH stimulus and items [note that in RC/MC the total set is regarded and used as the focus of the comprehension assessment]		

Variable elements within tasks

READING AT B1

[nb: No repetition across tables—Each new table subsumes the table before OR task types change]

Action/Skills: Simple to Moderate	Definitions
Comprehend	Decode task directions and task/item texts sufficiently to choose correct response at lowest level 30% of the time; mid-level 60% of the time; B1/B2 level 80% of the time
Compare/Contrast	Compare gist and details of each of five passages in order to match passage to statement; compare 10 statements to each other and to the 5 passages in order to match passage to statement
Evaluate	Correctness of each of 5 to 7 options as answers to question requiring two correct options
Classify	Statement as found in “passage 1, passage 2, passages 1 and 2, neither”
Organize	Details in a passage according to categories specified in question
Choose	Correct option/options

CONTEXTUAL COMPLEXITY OF STIMULUS MATERIALS

Reading at B1

Context: Simple to Moderate	Descriptions by task
Task 2: M/C RCMP	• 400 to 450 word passage • Descriptive/expository prose • Concrete topic [constrained by EF Compass] • Vocabulary 80 to 85% in top 1000 • Syntax includes occasional complex sentences

COMPLEXITY OF ITEMS
Reading at B1
Task by Task Item Verbal Components/Variables/Distractor Characteristics
[29 pts/1 task each type]

RCMP pssg + 8 M/O MC questions	**Item Description[s]**	**Component variables**	**Distractor Characteristics**
	• 8 questions TOTAL • One 5-option M/C gist/purpose/main idea question • 7 five-option M/C questions • Specific details and inference questions have 2 correct answers	• Gist (1) of whole passage OR • Main idea (1) of individual paragraph OR • Main purpose of passage • Specific details (4-5) • Inference (2-3)	• All distractors are plausible in the overall context of the passage stimulus topic • For 4 of 6 questions, all • Options come from the passage • For GIST and INFERENCE the options are a plausible and connected set, connected to the overall topic of the stimulus • No question can be correctly answered without reading the passage
Item/stimulus interaction[s] design	• Elicit COMPARISON of passage details in order to support RECOGNITION of multiple correct fits with question		

Variable elements within tasks

6. COMPASS

COMPASS stands for COMmon Principles for Academic Scope and Sequence. It is a metadata framework which is intended to help EF organize its own universe of educational content. The long-term goal is that all of EF's teaching and learning materials – or at least a critical mass – is described using a consistent system, whether it is online, in a book, part of a group activity,a test, or a teacher-led session., a book, a group activity, a test, or a teacher led session. The initial draft maps macro and micro "can do" statements to each CEFR level and connects the associated language functions, grammar, vocabulary, topics, themes, and settings appropriate for each level.

7. COMPUTER-ADAPTIVE MULTI-STAGE TESTING (CA-MST)

The measurement challenges the EF SET seeks to meet are complex. First, the desired outcome is a score that is reasonably reliable across a very broad range of language proficiency, from beginners to advanced English language users. Second, the goal was to make the test as efficient as possible in terms of test takers' time, so every minute of testing time needed to contribute to the accuracy of the measurement. Third, the EF SET was designed to be completely automated, so that human intervention in any test taker's test session would be completely unnecessary: from start of the test to the final score report, everything was intended to work reliably every time without human intervention. The ca-MST approach to test assembly, delivery and scoring promised the most trustworthy support for meeting these challenges.

Computer-adaptive multistage tests combine the advantages of more traditional computer-adaptive tests—more accurate and precise measurement—with enhanced quality control of the item banks, the test forms, the exposure rate (or security) of the items, operational test delivery and management of test taker data. In addition, the fixed module and panel test assembly and delivery mechanism simplifies the actual "next-step" routing and scoring of each test taker's responses. This is particularly important in a web-based delivery environment, like that of EF SET, because it can help to avoid problems caused by data transmission or connectivity problems.

Like all computer-adaptive tests, the ca-MST uses real-time information about the test taker's ability, based on the test taker's responses to items, to ever more accurately match the difficulty of subsequent items with the ability reflected in item responses thus far. In the ca-MST, however, the path to the final score is determined by a particular "route" the test taker takes through items grouped in modules, rather than by an item-by-item calculation.

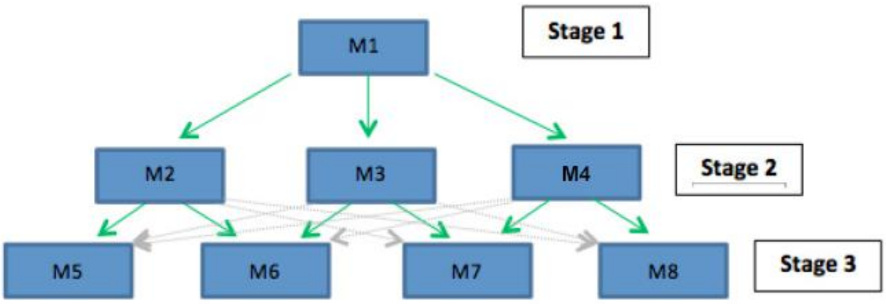
Items in a ca-MST are organized into "panels," which consist of a pre-determined number of individual "modules" organized into a pre-determined number of "stages." Each module is constructed to precise content and statistical specifications and test takers' performance on the first module or Stage 1 (usually constructed to measure best at middle difficulty) determines which module they will see in Stage 2. Performance on the Stage 2 module determines the module presented at Stage 3. At each stage, increasingly precise estimates of the test taker's ability are available.

The application of Assessment Engineering (AE) principles to the test development process, in combination with the decision to use Rasch model item analysis made it possible for EF to create a calibrated pool of items that could be constantly renewed and replenished. Specifications for the assembly of each module include both content and statistical specifications, but the module information function is the controlling target for the assembly of all modules.

8. FINAL EF SET CA-MST DELIVERY SPECIFICATIONS

In the EF SET ca-MST, a test taker's ability is estimated initially by performance—number of questions answered correctly—on Module 1, which is constructed to measure best at the mid-point of the B1 CEFR level. Based on that initial estimate, the test taker is automatically routed to one of three modules in Stage 2, modules constructed to measure best at the mid-point of the A2, B1, and B2 CEFR levels. Based on the Stage 2 estimate of ability, the test taker is automatically routed to one of four modules in Stage 3. These modules are constructed to measure best at the cut scores that separate A1 from A2, A2 from B1, B1 from B2, B2 from C1, and C1 from C2 CEFR levels. These cut scores were determined by the Standard Setting Study conducted in July 2014.

Panel Overview:



While there are 12 possible routes through these stages, the EF SET reading and listening tests enable only six. The disabled routes are regarded as less likely than the enabled routes, but they can be enabled at any time, should the data indicate the need to activate them. Here are the EF SET routes through the ca-MST:

- M1-M2-M5
- M1-M2-M6
- M1-M3-M6
- M1-M3-M7
- M1-M4-M7
- M1-M4-M8

At each routing point, intermediate thresholds, or score targets, are embedded in the panel's data management software. These thresholds determine each test taker's next module: based on performance on the Stage 1 module, the test taker is routed to one of the three Stage 2 modules, and based on the test taker's cumulative performance on the Stage 1 plus Stage 2 modules, the test taker is routed to one of the four Stage 3 modules. The total EF SET reading or listening score is the result of performance on all three stages.

EF SET CONTENT SPECIFICATION

In general, EF SET panels consist of the following configuration of tasks in each module:

Stage/Module	Target level of difficulty	Maximum information at CEFR*	No. of tasks in EF SET (Reading - Listening)	No. of tasks in EF SET PLUS (Reading - Listening)
Stage 1/Module 1	Medium	Mid-B1	1 – 1	2 – 2
Stage 2/Module 2	Easy	Mid-A2	1 – 1	2 – 2
Stage 2/Module 3	Medium	Mid-B1	1 – 1	2 – 2
Stage 2/Module 4	Difficult	Mid-B2	1 – 1	2 – 2
Stage 3/Module 5	Very Easy	A1/A2 cut	1 – 1	3 – 3
Stage 3/Module 6	Easy-Medium	A2/B1 cut	1 – 1	3 – 3
Stage 3/Module 7	Medium-Difficult	B1/B2 cut	1 – 1	3 – 3
Stage 3/Module 8	Very difficult	B2/C1 cut	1 – 1	3 – 3

*All mid-points and cut points are determined by the results of the 2014 Standard Setting Study

As well as keeping task specifications per panel uniform, caution is taken to ensure that the standard error of measurement of any route taken on the 120-minute EF SET PLUS does not exceed 0.29, and on the 50-minute EF SET, no more than 0.45.

9. RASCH MODEL ANALYSES

Throughout the iterative trials that preceded the operational launch of EF SET, items were subjected to both Classical Test Theory[6] (CTT) and Rasch model[7] analyses. The CTT analyses provided successive views of the item performance with each new trial sample, thus yielding empirical evidence of the differences in these samples. CTT analyses also were used for in-depth reviews of the difficulty and discrimination of each response option in each multiple-choice item (distractor analysis). The limitation of CTT analysis for EF SET was the dependence of the item statistics on a particular sample of test takers, and the need for very elaborate controls and statistical analyses to generalize results and item characteristics over multiple populations and multiple forms of tests.

Because the EF SET is designed to be an online, on demand test of a very broad range of English language proficiency levels in a testing population that may vary dramatically in its composition from one testing occasion to another, a psychometric model that could facilitate automation of stable and reliable scoring and reporting was desirable.

Rasch model analyses were initially performed to establish the appropriateness of this kind of analysis for the EF SET data. The Rasch model was particularly attractive as the psychometric approach to EF SET because of its most basic assertion: the difficulty of test items must be invariant across test takers. That is, an item intended to measure at the B1 CEFR level must always measure at that level, regardless of the language proficiency level of the test taker. B1 does not get reinterpreted as a result of a more or less able test taking population.

In the actual implementation of the model for EF SET, initial Rasch model analyses of trial items were confirmed and reconfirmed in a series of iterative trials. Once item statistics, including the aforementioned difficulty, are established, the items are said to be “calibrated.” Calibrated items can be embedded with new items. These new items are then on the same scale as the original items (i.e., item results can be interpreted in the same way). This process of calibrating items to a common scale was used over several waves of item-writing, test administration, and item analysis. Once sufficient data had been collected and analysed, the formal reporting scale for the item pool was established. All subsequent new items were calibrated to that scale using the process of embedding items that were already calibrated (i.e., using anchor tests for linking the scales).

[6] Kline, T. J.B(2005). Classical Test Theory: Assumptions, Equations, Limitations, and Item Analysis. In T.J.B. Kline. Psychological Testing: A Practical Approach to Design and Evaluation .Sage Publications.

[7] Bond, T. G. ,Cook, J., & Fox, C. M. (2007) Applying the Rasch Model. Fundamental Measurement in the Human Sciences. 2nd Edition..University of Toledo; http://en.wikipedia.org/wiki/Rasch_model;. Hambleton, R. K. and Jones, R. W.(1993) Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development. Educational Measurement Issues and Practice, 12, 39-47.

These calibrated items comprised the test forms taken by EF students in January 2013. The students' language proficiency scores were estimated on the same scale as that used for item calibrations. In Rasch model analyses, the scales for item difficulty and person ability are reported in the same units. Item difficulty is usually indicated by b, while person achievement is usually indicated by the Greek letter theta, θ. By having items and people on the same scale, many other tasks related to test creation are made easier. Those uses of the scale are discussed in other sections of this document. The calibrated Rasch scale for these test forms was used as the basis for the initial standard setting study in June 2013, which established the ability levels, on the θ scale, for each of the CEFR levels reported on the EF SET. Concomitant with the second standard setting study in July 2014, which established the official and final definitions of the CEFR levels on the θ scale for the operational EF SET launched in September 2014, the entire EF SET item pool was recalibrated to ensure the accuracy of the item statistics.

10. STANDARD SETTING STUDY

Overview

In the pages that follow you will find the Standard Setting Study report from July 2014. EF did an initial Standard Setting Study in June 2013 to establish cut scores and CEFR level ability definitions on the θ scale for the beta version of the EF SET (published in October 2013). In the year that followed this initial Standard Setting Study, EF performed intensive analysis of item performance, new trial data for new items, and the operational performance of the initial cut scores. In order to be certain that the cut scores that would govern reported scores on the operational forms of the EF SET from 2014 forward represented a consensus of expert judgment that was a fair and thoughtful as possible, EF decided to perform a second Standard Setting Study in July of 2014. While the technical procedures that were used in the two studies are very similar, they are not identical and we present the full report of the final study below for those interested in the technical processes and procedures.

The initial paragraphs are part of the orientation to the process involved in of any standard setting study sent to the language experts who participated in each study.

Orientation for Panelists

Most people are all-too familiar with tests and the notion of a score scale. Test scores can be reported as the number of total points earned or as a percentage of correct answers—for example, "I received a sore of 90% correct on my math test." Or, score scales can be derived using rather sophisticated statistical methods to ensure that even examinees taking different forms of a test can be fairly and accurately compared to one another or evaluated relative to fixed set of standards that define what various levels of proficient mean. Test scores almost always have a purpose—to assign course grades, to qualify candidates for a certificate or license, for college or university admissions, to assign scholarships or awards, or to place students in appropriate courses of study. The test scores therefore need to be elaborated by a hopefully useful set of evaluative interpretations—the meaning of the scores within the context of the test's purpose.

Standard setting is a judgmental process that adds that type of meaning to a test score scale. For example, when a student gets a score of 80, is that good, is it only marginal, or is it excellent? The simple answer is, "It depends." What was the average score? Was the test easy or hard? Were all of the students amply prepared to take the test? Although there are many myths about the absolute interpretation numbers—for example, the myth that 70 or 75 percent is always "passing"-- the reality is that some level of human judgment is needed to determine appropriate and defensible cut scores on a test score scale based on the content of the test, proficiency claims that we associated with the scores, the quality of the scores, and the purpose for which the scores are to be used. Standard setting is one of the best ways of arriving at those cut scores.

So what are cut scores? Cut scores are specific values on the score scale that classify individuals into two or more categories, based on their test performance. For the standard-setting activity in which you have agreed to take part, there will be four cut scores that are used to classify English language students into the five categories shown in the table, below[8]. That is, we will be setting cut scores that distinguish levels A1 and A2, levels A2 and B1, levels B1 and B2, and levels B2 and C1, and levels C1 and C2.

[8]These levels are based on the first five levels of the Common European Framework of Reference for Languages: Learning, Teaching and Assessment (Council of Europe, a 47-member international organization). The proficient level (C2) is not included here

Type of Language User	Level	Code	Description
Basic	Beginner	A1	Understands familiar everyday words, expressions and very basic phrases aimed at the satisfaction of needs of a concrete type
	Elementary	A2	Understands sentences and frequently used expressions (e.g. personal and family information, shopping, local geography, employment)
Independent	Intermediate	B1	Understand the main points of clear, standard input on familiar matters regularly encountered in work, school, leisure, etc.
	Upper intermediate	B2	Understands the main ideas of complex text or speech on both concrete and abstract topics, including technical discussions in his/her field of specialisation
Proficient	Advanced	C1	Understands a wide range of demanding, longer texts, and recognises implicit or nuanced meanings
	Proficient	C2	Understands with ease virtually every form of material read, including abstract or linguistically complex text such as manuals, specialised articles and literary works, and any kind of spoken language, including live broadcasts delivered at native speed

Standard Setting Study Report July 2014

It is important to understand that standard setting is a group process—no one person gets to make the final cut-score decisions. Rather, the standard-setting process allows a designated group of qualified panelists—the standard-setting panel, in this case qualified English language educators—to make an informed and reasoned set of judgments that ultimately set the four cut scores on two different tests: an English reading test and an English listening test.

There are several purposes for these classifications. The first is to place prospective students in the appropriate English language course of study. The second is to provide English language reading and/or listening certifications for the students at the completion of one or more courses of EF study. The third is to provide language learners outside EF some trustworthy information about their level of language proficiency on an EF standard English test.

SETTING STANDARDS FOR THE EF READING AND LISTENING ASSESSMENTS

JULY 2014

Richard M. Luecht, Ph.D.
The University of North Carolina at Greensboro

INTRODUCTION

Many tests use cut scores to denote “passing performance” or to otherwise classify the test takers into two or more groups. However, contrary to some popular mythology “70% correct” is NOT always “passing.” For example, if two individuals each take a different test form—test forms that differ substantially in difficulty—the person taking the more difficult test form would be at a distinct disadvantage. Unfortunately, there are many such misconceptions floating around about the source and nature of cut scores on a test. The simple fact is that tests developed, scored, and used in line with the comprehensive Standards for Educational and Psychology Tests (American Psychological Association, American Educational Research Association, and National Council on Measurement in Education, 2014) and/or the International Testing Commission’s Guidelines for Test Use (ITC, 2007) are very likely to engage in a systematical process called standard setting to determine not only the cut scores but also meaningful performance-based descriptions about the proficiency categories demarcated by the cuts scores (Cizek. 2013).

Most standard setting studies attempt to set a single cut score—as noted above, a passing score. This report describes a very ambitious and comprehensive standard-setting process that resulted in a total of ten cut scores for the new EF tests of listening and reading. EF developed their standardized, objectively scored reading and listening assessments to be aligned to the Council of Europe’s Common European Framework of Reference (CEFR). For more information, visit: www.efset.org/english-score/cefr. The CEFR provides a set of conceptual guidelines that describe the expected proficiency of language learners from beginning to advanced language proficiency. However, ad hoc or logical alignment of test materials to the CEFR is not entirely sufficient to ensure proper interpretation of EF test scores. Standard setting must be used to go beyond conceptual alignment and more directly map the conceptual CEFR performance expectations and interpretations to statistically determined cut scores on the corresponding EF reading and listening test score scales.

Figure 1 presents the six conceptual CEFR proficiency levels: A1, A2, B1, B2, C1, and C2. A1 is the lowest level and C2 is the highest level. Therefore, five cut scores are required to distinguish between adjacent levels: Cut #1 for A1 and A2; Cut #2 for A2 and B1; Cut #3 for B1 and B2; Cut #4 for B2 and C1; and Cut #5 for C1 and C2.

Type of Language User	Level	Code	Description
Basic	Beginner	A1	Understands familiar everyday words, expressions and very basic phrases aimed at the satisfaction of needs of a concrete type
	Elementary	A2	Understands sentences and frequently used expressions (e.g. personal and family information, shopping, local geography, employment)
Independent	Intermediate	B1	Understand the main points of clear, standard input on familiar matters regularly encountered in work, school, leisure, etc.
	Upper intermediate	B2	Understands the main ideas of complex text or speech on both concrete and abstract topics, including technical discussions in his/her field of specialisation
Proficient	Advanced	C1	Understands a wide range of demanding, longer texts, and recognises implicit or nuanced meanings
	Proficient	C2	Understands with ease virtually every form of material read, including abstract or linguistically complex text such as manuals, specialised articles and literary works, and any kind of spoken language, including live broadcasts delivered at native speed

Figure 1. Six CEFR Language Proficiency Levels

It is important to understand that all EF reading and listening test forms are scored using statistically calibrated (and equated) score scales[1]. Therefore, from a fairness perspective, it does not matter whether a particular test taker gets an easier or more difficult test form. Any and all differences in test-form difficulty are automatically managed by the statistical calibration and equating process. That is, every examinee can be scored and their performance judged in terms of the SAME underlying score scale, regardless of the difficulty of his or her test form. The standard setting process, in turn, determines the cut scores on the underlying reading and listen scales. As long as those scales continue to be statistically maintained over time, the same five cut scores per scale can be applied to classify all examinees into the corresponding six CEFR categories.

This report summarizes standard-setting activities and rating results obtained from a panel of fifteen seasoned English language teachers and learning experts that convened in London, UK in early July 2014[2]. The panel completed online versions of the EF reading and listening test forms, reviewed the associated assessments tasks in depth, and rated actual test-taker performance to come up with ten cut scores: five for the reading scale and five for the listening scale. The standard setting process was considered to be highly successful and defensible on technical grounds and generated results that further made practical sense. In general, the panelists were highly engaged and seemed adequately informed and competent to complete the standard-setting tasks.

Standard-Setting Procedures

Standard setting is a reasoned process for linking proficiency-anchored expectations and interpretations to cut scores on a score scale—in this context, deciding how much reading or listening skill and knowledge are required to support the proficiency claims we might make about examinees at various CEFR levels. For example, suppose that we determine a particular cut score on the reading scale that separates CEFR levels A2 and B1 (elementary versus intermediate reading proficiency). That cut score implies a reading proficiency transition (per Figure 1) from reading and understanding sentences and frequently used expressions to being able to understand the main points of passages and more extended blocks of text. In essence, the standard setting certifies that any examinee scoring at or above the designated cut score has made that type of transition and can demonstrate that B1 level of reading proficiency most of the time.

This section of the report discusses selection of the panelists for the standard setting, the selection of test materials, and the actual standard setting process used. Statistical outcomes are summarized in the next section of the report (see Analysis and Results).

Selection and Orientation of Panelists

All standard setting procedures are fallible insofar as being based on human judgments about what “proficient” means in a particular assessment context. Employing different methods of setting the standards, using different standard-setting panelists, or merely altering the choice of different test materials and information presented to the standard setting panel can result in different standards and associated cut scores. If the standard-setting study process of collecting ratings from qualified panelists is reasonably objective and carried out in good faith, most standards have been successfully defended against technical, legal and ethical challenges.

A defensible standard-setting process usually begins with the careful selection of panelists (Raymond & Reid, 2001). This is one of the most critical aspects of the entire process. The panel needs to be representative of the constituent population, have the requisite knowledge and skills to judge test items and/or examinee performances, and be willing to determine who is and who is not in at the designated levels of proficiency terms of demonstrated test performance. Fifteen language subject-matter experts (SMEs) were selected to participate in the EF standard-setting panels. All of the panelists were familiar with EF, had English as their native language, had extensive experience with English language teaching and learning, and holders of either a Master’s degree in TESOL or a Cambridge Diploma in English Language Teaching to Adults (DELTA). The panelists were also independently interviewed by a linguistic expert on the EF staff as part of the vetting process to confirm their credentials. Finally, all of the panelists were expected to independently complete an online version of the EF listening and reading tests (without access to any answer keys) prior to coming to the standard-setting exercise as a means of familiarizing themselves with the assessment tasks and items, as well as experiencing the test in a manner similar to that of a typical test taker. Finally, the panelists were provided with various

[1] EF test forms are calibrated and equated to a common metric or scale using a statistical measurement model known as the item response theory (IRT) Rasch model (Rasch, 1960, Masters, 1982; Wright & Masters, 1980; Lord, 1980; Hambleton & Swaminathan, 1985). The advantage of this calibration and equating process is that examinees receive comparable scores on the SAME underlying score scale, regardless of possible changes in test form difficulty.

[2] An initial formal standard setting study was conducted in June 2013 to establish the operational definitions of each of the CEFR levels on the beta version of the EF listening and reading tests. The decisions reached in the second standard setting study done in July 2014, govern the final version of the EF reported score scale for the reading and listening EF SET.

orientation materials about EF and the standard setting activities in advance of the meeting.

The panel of fifteen individuals was convened for a one-day meeting on 05-July 2014 at the EF offices in London. EF staff members were instrumental in selecting and communicating with the panel, providing logistical support, preparing materials for the standard-setting study, and ensuring the successful completion of the activities. Each panelist was randomly pre-assigned to one of five groups, A to E, so that there were three panelists within each group. Everyone within each group reviewed exactly the same reading and listening test tasks and response data.

Selection of Test Materials

Operational listening and reading test materials were used for the standard setting. The selection of test materials was somewhat complicated because of the adaptive test design used for the operational EF examinations. That is, every EF test form is comprised of adaptively administered listening or reading tasks using a test design known as computerized adaptive multistage testing or caMST (Luecht & Nungester, 1998; Luecht, 2014). A caMST specifically targets each selected task to each examinee's apparent listening or reading proficiency and results in a statistically optimal test from a measurement perspective. As noted earlier, item response theory (IRT) is then used for scoring and automatically adjusts for the differential task difficulty of each examinee's ultimate "test form". However, while an adaptive test design highly useful from a measurement perspective, the caMST design also slightly complicated the selection of the test materials for standard setting because very few examinees took exactly the same listening and reading tasks making it infeasible to create intact "test forms" for the panelists to evaluate. Instead, the test materials and examinee response data were selected at the task level and included ten listening tasks that spanned the A2 to C1 CEFR levels and ten reading tasks that bridged the A2 to B2 levels[3].

Each EF task chosen for the standard setting comprised a set of selected-response (SR) items linked to a particular mode-specific stimulus. The reading tasks were composed of a text-based reading passage and six to eight SR items associated with that particular passage. Each listening task included an aural passage, presented via headphones, and the associated items in the set (five to six items). A variety of matching and choice-based SR item formats are included on the EF examinations (i.e., not all of the items were one-best-answer multiple-choice).

All of the tasks used in the standard setting were previously administered to large samples of examinees. That large-sample response data was then calibrated using an item response theory (IRT) model known as the partial-credit model (Wright & Masters, 1982; Masters, 2010). The partial-credit model can be written as follows:

$$P\left(x=X\middle|\theta;b_i,\mathbf{d}_i\right)\equiv P_{ix}\left(\theta\right)=\frac{\exp\left[\sum_{k=0}^{x}\theta-\left(b_i+d_{ik}\right)\right]}{\sum_{j=0}^{m}\exp\left[\sum_{k=0}^{j}\theta-\left(b_i+d_{ik}\right)\right]}$$

(Equation 1)

where θ is the examinee's proficiency score, bi denotes an item difficulty or location for item i, and dik denotes two or more threshold parameters associated with separations of the category points for items that use three or more score points (k=0,…,xi). All reading items and tasks for the EF standard setting were calibrated to one IRT scale, θR; All listening items and tasks were calibrated to another IRT scale, θL.

There were two criteria used in selecting tasks for the standard-setting: (1) the items had to have good statistical/psychometric characteristics, and (2) in aggregate, the task sets of items were expected to represent a reasonable spread of item difficulty. As noted above, both the listening and reading tasks spanned the A1 to C1 levels. All included tasks and items were jointly vetted by the psychometric advisors on this project and EF content experts. Table 1 provides a summary of the listening tasks assigned to each of the five groups, A to E. The stimulus topic, approximate CEFR level assigned to the stimulus and items, number of items and average (mean) IRT item difficulty are listed left to right. The coding used in the 'CEFR Level" column corresponds to the level associated with each task as a whole. Table 2 shows a corresponding summary of the reading tasks by group.

[3] The average difficulty of the items associated with each task was used to classify the tasks by approximate CEFR level. The cut scores set in the initial standard setting study of June 2013 were used as the basis for those approximate classifications.

There were two criteria used in selecting tasks for the standard-setting: (1) the items had to have good statistical/psychometric characteristics, and (2) in aggregate, the task sets of items were expected to represent a reasonable spread of item difficulty. As noted above, both the listening and reading tasks spanned the A1 to C1 levels. All included tasks and items were jointly vetted by the psychometric advisors on this project and EF content experts. Table 1 provides a summary of the listening tasks assigned to each of the five groups, A to E. The stimulus topic, approximate CEFR level assigned to the stimulus and items, number of items and average (mean) IRT item difficulty[4] are listed left to right. The coding used in the 'CEFR Level" column corresponds to the level associated with each task as a whole. Table 2 shows a corresponding summary of the reading tasks by group.

TABLE 1. LISTENING TASK DESCRIPTIONS BY STANDARD SETTING PANEL GROUPS (A TO E)

Listening Tasks	CEFR Level	No. of Items	Mean Difficulty
	Group A		
Carnival	A1(i)	5	-2.45
Parting Ways	A1(ii)	6	-1.03
Tickets	A2(i)	6	-0.69
United Nations	B1(i)	6	0.53
	Group B		
Tickets	A2(i)	6	-0.69
Dubai Girl	A2(ii)	6	-0.58
United Nations	B1(i)	6	0.53
New Species	B2(i)	6	1.06
	Group C		
Being a Pilot	B1(ii)	6	0.24
United Nations	B1(i)	6	0.53
New Species	B2(i)	6	1.06
Entrepreneurs	C1(i)	6	1.87
	Group D		
United Nations	B1(i)	6	0.53
New Species	B2(i)	6	1.06
Coffee	B2(ii)	6	0.99
Entrepreneurs	C1(i)	6	1.87
	Group E		
United Nations	B1(i)	6	0.53
New Species	B2(i)	6	1.06
Genetic Engineering	C1(ii)	6	1.19
Entrepreneurs	C1(i)	6	1.88

[4] *Easier items, on average, are denoted by negative values on the IRT scale. Larger, positive numbers denote difficult items.*

TABLE 2. READING TASK DESCRIPTIONS BY STANDARD SETTING PANEL GROUPS (A TO E)

Listening Tasks	CEFR Level	No. of Items	Mean Difficulty
Group A			
Eating Out	A1(i)	8	-1.98
Holiday Accommodations	A1(ii)	8	-1.61
Boat Race	A2(i)	8	-0.79
Steinway Prize	B1(i)	8	0.39
Group B			
Boat Race	A2(i)	8	-0.79
Food Trucks	A2(ii)	8	-0.39
Steinway Prize	B1(i)	8	0.39
Future of Technology	B2(i)	6	1.23
Group C			
Steinway Prize	B1(ii)	8	0.39
Dream Control	B1(i)	8	0.81
Future of Technology	B2(i)	6	1.23
Malaria Treatment	C1(i)	8	1.27
Group D			
Steinway Prize	B1(i)	8	0.39
Future of Technology	B2(i)	6	1.23
Celestial African Art	B2(ii)	8	1.84
Malaria Treatment	C1(i)	8	1.27
Group E			
Steinway Prize	B1(i)	8	0.39
Future of Technology	B2(i)	6	1.23
Working from Home	C1(ii)	8	2.18
Malaria Treatment	C1(i)	8	1.27

Rather than have each of the five groups contend with the full span of proficiency and difficulty (A1 to C2), the four listening and four reading tasks assigned to each group were selected to cover only three adjacent CEFR levels (e.g., Group B's tasks covered A2, B1 and B2). Table 3 shows how tasks were further spiraled across the five CEFR levels. Some of the same tasks, A1(i), B1(i), B2(i) and C1(i) were reused across groups to provide additional "connectivity" in the data. For example, "B1(i)" is the same task assigned to all five groups. Note: there were no specific listening or reading tasks at the C2 levels. The C2 level was nonetheless considered by the panelists as a viable proficiency rating category[5].

[5] *Since the panelists were rating examinee responses, not the items or tasks per se, there was no inherent reason why C2 tasks HAD to be included.*

Overview of 'The Integrated Judgment Method' (IJM)

The standard-setting method of choice for the EF reading and listening examinations is called the integrated judgment method or IJM (Jaeger & Mills, 2001). The IJM requires the standard setting panelists to simultaneously focus on the test content, on the statistical operating characteristics of actual test items, and on patterns of actual examinee performance on the test items. Using the information provided, the panelists assign holistic ratings to effectively classify each examinee into the designated proficiency-anchored categories. The statistical relationship between their classifications and the examinees' scores is then exploited to derive the cut scores on the score scale. The IJM was specifically selected for the EF standard setting for three reasons. First, it grounds the standard setting panelists in both the test content and in actual examinee performance on the test. Second, it is well-suited to for standard setting involving multiple cut scores. Third, it is readily adaptable for different item types ranging from selected-response items to complex performance-based tasks and simulations.

Under the IJM, qualified panelists already familiar with the test content and items review raw response strings from a selected sample of examinees and rate each response string according to the relevant proficiency categories[6]. The panelists must consider the relative difficulty of each item, total score performance and the patterns of choices or answers provided by a select number of examinees. The panelist, working independently, evaluates each examinee's response string and assigns a rating that reflects the panelist's judgment about the examinee's proficiency level. In the present context, panelists were instructed to use ratings of 1 (corresponding to CEFR A1-level performance) to 6 (C2-level performance). The C2 level was allowed but the panelists were also instructed that there was no requirement to actually use it unless definitely warranted.

Figure 2 provides a visual depiction of the type of data provided to the panelists for carrying out their IJM ratings. Responses strings for three examinees are shown in the figure (rows in the rectangle labeled RESPONSE STRINGS). There are 8 item responses corresponding to a single task completed by the three examinees in this example. Ones ("1") denote correct answers and letters denote a particular incorrect letter choice. The stimulus—described as a reading passage in this example—and actual item text and response-selection options are available to the panelists at all 8 items. The statistical item difficulties were provided to the panelists (P1=.80 indicates that 80 percent of the all examinees got item #1 correct) . The total scores and self-reported proficiency levels are ancillary information for the panelists to consider. In short, the panelists get to see exactly what each examinee did on every item and the examinees' overall performance. The panelists also had full statistical performance information about every item. A unique and important aspect of the IJM is that the panelists must integrate all of this information in real-time and use it to make a holistic, reasoned judgment about each examinees' apparent level of proficiency.

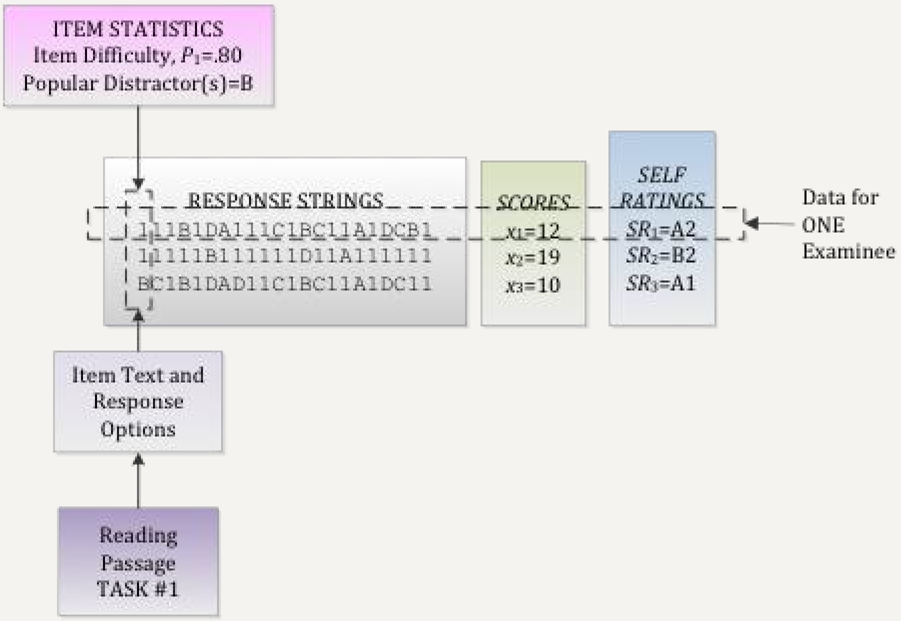


Figure 2. Information Presented to the Panelists for the IJM

[6] Jaeger & Mills (2001) presented additional rating categories such as "far below", "clearly below", "just below", "just barely", and "clearly above", around each conceptual cut score. A decision was made here to eliminate those multiple intermediate levels and focus exclusively on the six CEFR levels to essentially force the panelists to make distinct choices. The panelists appeared to work well with only the primary CEFR levels in the present standard setting study and did not indicate any need to use additional levels of separation.

An important and interesting aspect of the IJM is that each panelist can consider the pattern of responses for each sampled examinee in making their judgments. For example, in the context of multiple-choice items, two examinees having the same total score may have rather different response strings for their incorrect responses. That is, each response string contained the raw responses—including incorrect choices—for a single listening or reading task. Panelists were also provided with auxiliary information about the difficulty of each item as well as the popularity of incorrect response options. Each response string therefore provided the panelists with a multi-item profile of listening or reading performance.

Selection of Examinee Response Strings

A stratified random sampling strategy was employed to choose the response strings. As noted in the prior section, there were ten unique listening tasks and ten unique reading test tasks (see Tables 1, 2 and 3). EF had relatively large data sets comprised of operational examinees who had taken the tests over the past 12 months. In addition, each examinee had already been assigned to a CEFR-level English course. The random sampling targeted selecting 40 to 60 examinees for each task such that their smaller distributions of scores on each task were proportional to the full sample of examinees completing that same task. This process was repeated for each of the ten listening and each of the ten reading tasks. The sampling generated 36 to 63 unique response strings per task as shown in Table 4 (listening on the left; reading on the right).

TABLE 4. NUMBERS OF RESPONSE STRINGS CHOSEN PER LISTENING AND READING TASK

Listening Tasks	CEFR Level	No. of Response Strings	Reading Tasks	CEFR Level	No. of Response Strings
Group A					
Carnival	A1(i)	63	Eating Out	A1(i)	57
Parting Ways	A1(ii)	42	Holiday Accommodations	A1(ii)	59
Tickets	A2(i)	59	Boat Race	A2(i)	58
United Nations	B1(i)	58	Steinway Prize	B1(i)	60
Group B					
Tickets	A2(i)	59	Boat Race	A2(i)	58
Dubai Girl	A2(ii)	56	Food Trucks	A2(ii)	56
United Nations	B1(i)	58	Steinway Prize	B1(i)	60
New Species	B2(i)	58	Future of Technology	B2(i)	36
Group C					
Being a Pilot	B1(ii)	61	Steinway Prize	B1(ii)	60
United Nations	B1(i)	58	Dream Control	B1(i)	51
New Species	B2(i)	58	Future of Technology	B2(i)	36
Entrepreneurs	C1(i)	56	Malaria Treatment	C1(i)	60

TABLE 4. NUMBERS OF RESPONSE STRINGS CHOSEN PER LISTENING AND READING TASK (CONTINUED)

Listening Tasks	CEFR Level	No. of Response Strings	Reading Tasks	CEFR Level	No. of Response Strings
Group D					
United Nations	B1(i)	58	Steinway Prize	B1(i)	60
New Species	B2(i)	58	Future of Technology	B2(i)	36
Coffee	B2(ii)	55	Celestial African Art	B2(ii)	52
Entrepreneurs	C1(i)	56	Malaria Treatment	C1(i)	60
Group E					
United Nations	B1(i)	58	Steinway Prize	B1(i)	60
New Species	B2(i)	58	Future of Technology	B2(i)	36
Genetic Engineering	C1(ii)	57	Working from Home	C1(ii)	56
Entrepreneurs	C1(i)	56	Malaria Treatment	C1(i)	60

The July 2014 Standard-Setting Process

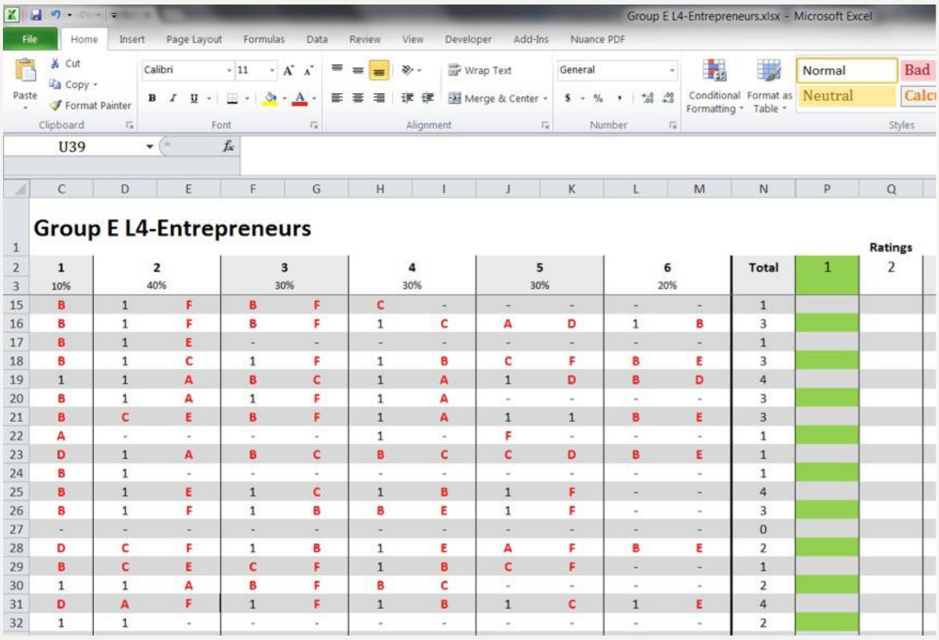
The standard setting was carried out the EF offices in London on 05 July 2014. The agenda and some relevant advance materials provided to the panelists are shown in Appendix A. No actual test materials were released prior to standard setting because of the secure nature of the tasks and items. EF staff included five facilitators for the corresponding groups (A to E), several administrative and technical support staff, and two psychometric consultants. Fifteen external panelists arrived at the EF offices at approximately 8AM for check in and to begin the activities.

The EF standard setting exercise began with a presentation by one of the psychometric consultants that outlined the purpose of the standard setting and a description of the day's activities. The overview introduced the panelists to the test materials they would be reviewing as well as providing an introduction to the IJM standard-setting procedure that would be used. A copy of the presentation is provided in Appendix B. Following this introduction and procedural overview, the panelists broke into their five assigned groups: A, B, C, D and E. One facilitator accompanied each group to a separate room. Three panelists were assigned to each group (A1, A2, and A3 assigned to group A, B1, B2 and B3 to group B, etc.). A relatively brief amount of time was allowed for introductions. Working with the group facilitators[7], the panelists discussed their perceptions about reading and listening proficiency in the context of the six CEFR levels (see Figure 1). A decision was made to introduce the panelists to the actual IJM rating task as quickly as possible. Therefore, the initial discussion of proficiency was also operationally combined with "practice" on the rating task for a small number of examinees. The response strings and other examinee-level data, as well as the panelists' ratings were all displayed/collected in Microsoft Excel®, using spreadsheets customized for the listening and reading forms seen by each of the four groups.

Figure 3 shows a sample rating spreadsheet presented to the panelists. The second and third rows in the spreadsheet, respectively, list the test items associated with each task. This example from Group E (listening) shows the responses corresponding to the six multiple-response items (two responses per item) associated with a listening task entitled "Entrepreneurs". In addition to the printed materials, the panelists could also listen via headphones to the same stimulus that the examinees heard. The raw responses for the selected examinees to the six items are shown in the body of the worksheet. If the examinee got the item correct, a "1" is shown; otherwise, the examinee's actual [incorrect] response selection is shown as a red letter. The column labeled "Total" contains the number-correct scores for each examinee on the task. All panelists assigned to the same group saw exactly the same tasks, test items, and responses to those items.

[7] The ratio of three panelists to one facilitator was extremely important to the success of the standard-setting exercise. This small panelist:facilitator ratio ensured timely assistance and response by the facilitators to every issue that arose. In addition, the facilitators periodically touched based with one another to share insights.

The panelists also entered their ratings directly into the spreadsheet for all of the examinees presented.



	C	D	E	F	G	H	I	J	K	L	M	N	P	Q
1	Group E L4-Entrepreneurs													Ratings
2	1	2		3		4		5		6		Total	1	2
3	10%	40%		30%		30%		30%		20%				
15	B	1	F	B	F	C	-	-	-	-	-	1		
16	B	1	F	B	F	1	C	A	D	1	B	3		
17	B	1	E	-	-	-	-	-	-	-	-	1		
18	B	1	C	1	F	1	B	C	F	B	E	3		
19	1	1	A	B	C	1	A	1	D	B	D	4		
20	B	1	A	1	F	1	A	-	-	-	-	3		
21	B	C	E	B	F	1	A	1	1	B	E	3		
22	A	-	-	-	-	1	-	F	-	-	-	1		
23	D	1	A	B	C	B	C	C	D	B	E	1		
24	B	1	-	-	-	-	-	-	-	-	-	1		
25	B	1	E	1	C	1	B	1	F	-	-	4		
26	B	1	F	1	B	B	E	1	F	-	-	3		
27	-	-	-	-	-	-	-	-	-	-	-	0		
28	D	C	F	1	B	1	E	A	F	B	E	2		
29	B	C	E	C	F	1	B	C	F	-	-	1		
30	1	1	A	B	F	B	C	-	-	-	-	2		
31	D	A	F	1	F	1	B	1	C	1	E	4		
32	1	1	-	-	-	-	-	-	-	-	-	2		

Figure 3. Sample Spreadsheet Used for Collecting the Panelist's IJM Ratings

In addition to the spreadsheet, a bound, hard-copy booklet was provided to each panelist containing the test items in the same sequence as the items presented (column-wise) in their spreadsheet. The booklets also contained the text passages for reading, item text, and answer key(s). As noted above, the listening prompts were made available via headphones. In general, comments from the panelists and facilitators suggested that, following the initial training and "hand-holding" by the facilitators, all of the panelists mastered locating needed information and carried out their assigned ratings task without any difficulty.

The panelists reviewed the pattern of correct and incorrect responses within their spreadsheet, considered the examinee's self-reported CEFR level, and assigned a rating of 1 to 6 in the column labeled "1" under "Ratings" (see Figure 3). Panelists were allowed to re-evaluate and change their ratings for each examinee in Round 2 (see the column labeled "2"). Cell protection within spreadsheet was activated to prevent the panelists from inadvertently overwriting any of the fixed-presentation data (response strings, etc.). In additional, each panelist's ratings were locked prior to starting the next round of ratings. As noted earlier, working with their facilitators, the panelists in each group jointly reviewed [approximately] the first five examinees in lieu of a formal rating practice exercise[8]. Each panelist was instructed on how interpret the response string, scores, and all other relevant item-level information. The panelists were able to review the test materials and statistical item data using the hard-copy test booklets. Although the panelists worked through the initial "practice" cases as a group, each panelist was allowed to enter their own rating choices for all of the task-specific response strings. In aggregate, the panelists provided 3,426 classification ratings for Listening and 2,934 ratings for Reading.

There were two rounds of ratings, with provision for a potential third round had the panelists wanted it for a final review of their ratings. The panelists unanimously decided that the third round was not needed. Round #1 was preliminary and meant to both orient the panelists to the review and rating tasks and to develop confidence in their perceptions and judgments about reading and listening proficiency. In standard setting, Round #2 is often viewed as the "movement" round where panelists, provided with feedback specific to their personal ratings from their first round, may elect to correct minor inconsistencies in their individual ratings after reflecting on their results and/or more globally increase or decrease their ratings to affect a change in the cut scores. At no time are the facilitators or anyone else observing the standard setting allowed to intervene to directly change or otherwise coerce any panelists to change their ratings. Instead, any and all changes to the cut scores are expected to occur solely because of reasoned decisions by one or more panelists to alter their earlier ratings, given group discussions and personal reflection.

[8] Given the small facilitator-to-panelist ratios and multiple rounds of ratings, it made little practical sense to provide separate practice exercises. Instead, a decision was made to get the panelists working on their real rating task as soon as possible.

Feedback was provided after Round #1 (see Appendix C). A plenary session was held with all of the 15 panelists to jointly explain the feedback. Figure 4 shows an example of the feedback shared with the panelists. There are six plots in Figure 4, where each numbered plot heading shows the CEFR level ratings that the panelists provided (1=A1, 2=A2, 3=B1, etc.). For example, the upper left-most plot #1 shows the results only for those examinees rated as "A1". The vertical scale within each plot shows the average test scores for examinees classified by the panelist at that CEFR level. Higher elevation on the vertical scale indicates higher test performance. The letters plotted correspond to the panelists' assigned groups A to E. Each cluster of three letters (e.g., all scores plotted with the letter "A") corresponds to the three panelists within that group. Therefore, each plot shows the examinees' average scores rated at that CEFR level by each of the 15 panelists. Note that some of the panelists did not assign any examinees to the C2 level (plot #6).

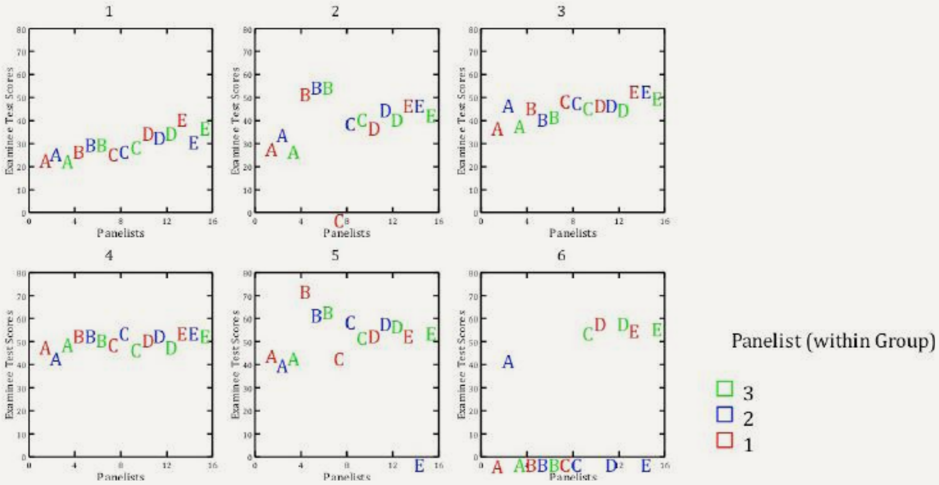


Figure 4. Average Test Scores by Rated Category (Three Panelist per Group A to E)

One caution in reviewing these results seems needed. Each of the five groups (A to E) was reviewing the performance of a different sample of examinees. Therefore, across-group comparisons of the raw ratings (e.g., Group A vs. E) are not necessarily relevant. The plots confirm that there was certainly some variation within and between the panelists. In general, the patterns of average scores shown in Figure 4 make sense. That is, as we move up the panelists classifications of examinees into the CEFR levels (1=A1 to 6=C2), the same examinees' average scores also go up. All of the panelists were asked to reflect on the results and their ratings in particular. Panelists who varied significantly from the rest of their group were further requested to closely scrutinize their ratings and either confirm or modify their ratings in Round #2.

By the end of Round #2, all of the panelists stated that they were completely satisfied with the ratings. None of the panelists felt the need for a third round. Table 5 shows the counts of examinees (response string records) classified by the 15 panelists into each of the six CEFR categories. Average test scores for those same examinees are also show. Listening and reading results are shown side-by-side for each panelist. In general, the average scores increase by category suggesting reasonable consistency between the ordering of the examinees based on their ratings and via their corresponding test scores.

TABLE 5. DETAILED LISTING OF EXAMINEE COUNTS AND AVERAGE RATINGS BY PANELIST ASSIGNED RATING CATEGORIES (1=A1,...,6=C2).

			Listening Tasks		Reading Tasks	
Group & Panelist	**CEFR**	**Rating Category**	**Count of Examinees**	**Average Test Scores**	**Count of Examinees**	**Average Test Scores**
Group A						
A1	A1	1	35	27	4	29
A1	A2	2	61	35	78	50
A1	B1	3	78	47	60	41
A1	B2	4	44	55	26	50
A1	C1	5	4	59	7	59
A1	C2	6	0	--	0	--
A2	A1	1	60	29	16	30
A2	A2	2	52	40	112	51
A2	B1	3	54	45	76	39
A2	B2	4	51	56	22	50
A2	C1	5	3	56	8	58
A2	C2	6	2	59	0	--
A3	A1	1	36	30	20	30
A3	A2	2	52	33	69	42
A3	B1	3	69	43	126	49
A3	B2	4	40	58	13	50
A3	C1	5	24	50	6	61
A3	C2	6	1	58	0	--
Group B						
B1	A1	1	21	25	9	30
B1	A2	2	21	36	66	54
B1	B1	3	14	50	55	49
B1	B2	4	3	62	23	56
B1	C1	5	0	--	1	75
B1	C2	6	0	--	0	--
B2	A1	1	29	27	16	5
B2	A2	2	56	40	58	57
B2	B1	3	85	47	39	45
B2	B2	4	58	55	36	56

TABLE 5. DETAILED LISTING OF EXAMINEE COUNTS AND AVERAGE RATINGS BY PANELIST ASSIGNED RATING CATEGORIES (1=A1,...,6=C2). (CONTINUED)

			Listening Tasks		Reading Tasks	
Group & Panelist	**CEFR**	**Rating Category**	**Count of Examinees**	**Average Test Scores**	**Count of Examinees**	**Average Test Scores**
Group B						
B2	C1	5	3	62	5	64
B2	C2	6	0		0	--
B3	A1	1	22	25	12	33
B3	A2	2	49	38	56	58
B3	B1	3	109	47	46	46
B3	B2	4	48	55	35	55
B3	C1	5	3	61	5	64
B3	C2	6	0	--	0	--
Group C						
C1	A1	1	11	38	3	45
C1	A2	2	15	46	10	39
C1	B1	3	113	46	23	45
C1	B2	4	94	53	16	49
C1	C1	5	0	--	6	55
C1	C2	6	0	--	2	66
C2	A1	1	9	34	7	29
C2	A2	2	38	41	29	41
C2	B1	3	139	49	100	50
C2	B2	4	44	56	48	57
C2	C1	5	3	63	22	62
C2	C2	6	0	--	1	61
C3	A1	1	3	26	7	40
C3	A2	2	24	41	22	42
C3	B1	3	119	47	58	48
C3	B2	4	82	54	79	53
C3	C1	5	5	53	37	58
C3	C2	6	0	--	4	61

TABLE 5. DETAILED LISTING OF EXAMINEE COUNTS AND AVERAGE RATINGS BY PANELIST ASSIGNED RATING CATEGORIES (1=A1,...,6=C2). (CONTINUED)

			Listening Tasks		Reading Tasks	
Group & Panelist	CEFR	Rating Category	Count of Examinees	Average Test Scores	Count of Examinees	Average Test Scores
Group D						
D1	A1	1	33	44	6	38
D1	A2	2	29	45	11	42
D1	B1	3	62	49	47	50
D1	B2	4	41	51	65	54
D1	C1	5	55	55	52	57
D1	C2	6	7	56	27	61
D2	A1	1	8	40	5	31
D2	A2	2	45	45	17	42
D2	B1	3	65	48	57	49
D2	B2	4	78	53	82	56
D2	C1	5	29	54	47	61
D2	C2	6	2	50	0	--
D3	A1	1	26	43	5	34
D3	A2	2	27	47	9	43
D3	B1	3	43	46	13	43
D3	B2	4	81	52	70	50
D3	C1	5	50	54	59	58
D3	C2	6	0	--	0	--
Group E						
E1	A1	1	60	44	19	43
E1	A2	2	48	50	37	50
E1	B1	3	84	52	75	56
E1	B2	4	31	54	48	57
E1	C1	5	6	61	19	56
E1	C2	6	0	--	14	58
E2	A1	1	12	38	4	33
E2	A2	2	83	48	45	47
E2	B1	3	107	52	115	56
E2	B2	4	23	57	41	57

TABLE 5. DETAILED LISTING OF EXAMINEE COUNTS AND AVERAGE RATINGS BY PANELIST ASSIGNED RATING CATEGORIES (1=A1,...,6=C2). (CONTINUED)

			Listening Tasks		Reading Tasks	
Group & Panelist	**CEFR**	**Rating Category**	**Count of Examinees**	**Average Test Scores**	**Count of Examinees**	**Average Test Scores**
Group E						
E2	C1	5	4	61	7	59
E2	C2	6	0	--	0	--
E3	A1	1	1	34	3	29
E3	A2	2	2	42	24	44
E3	B1	3	28	45	58	51
E3	B2	4	27	49	67	58
E3	C1	5	0	--	41	56
E3	C2	6	0	--	16	60

Analysis and Results

One of the more serious technical challenges in any standard setting exercise involves converting the panelists ratings to one or more cut scores on the corresponding official score scale. Multiple statistical mapping procedures can be used; only rarely do these procedures produce exactly the same cut scores.

The official score scales for the EF tests are based on an item response theory (IRT) model—in this case, the partial-credit model (see Equation 1). IRT software is used to calibrate all of the items in a particular item bank so that all scores from any associated test form will be on a common score scale. The score scale is maintained over time by linking new test forms and examination results to those same scales. A statistical mechanism was therefore needed by which the IJM classification ratings by the panelists could be statistically mapped to five cut scores on the IRT scale for listening (A1 vs. A2, A2 vs. B1, B1 vs. B2, B2 vs. C1, and C1 vs. C2) and five corresponding cut scores on the IRT scale for reading.

Every examinee response string used in the IJM standard setting was required to have a corresponding IRT-based test score. Proficiency scores of θ_R (reading) and θ_L (listening) had been previously estimated under the IRT partial-credit model (see Equation 1) and served as final test scores for the purpose of computing the cut scores. That is, the examinees actual test scores were considered along with the panelists' ratings of those same examinees to determine appropriate cut scores on the test score scale of interest.

Once the cut scores are determined, any examinee scored using that scale can be appropriately classified into one of the corresponding proficiency category demarcated by the cut scores. Based on preliminary analyses, it was decided that the most stable statistical outcomes would be obtained by merging all of the ratings and examinee IRT scores across panelists. This issue was carefully considered since it implied that the same examinees would be in every analysis at least three times (once for each panelist in each group and more often for tasks shared across some of the five groups). However, since the panelists ratings were independently determined, there was no reason to suspect any hidden statistical dependencies in the data nor inappropriate weighting of the statistical estimates.

Jaeger and Mills (2001) originally investigated using various statistical linear and polynomial regression functions to obtain the cut scores. Here, four different approaches were investigated with these data: (1) multinomial logistic regression, (2) discriminant function analysis (DFA), (3) weighted means, and (4) equipercentile equating (setting the cuts so that the marginal distribution of scores match the percentages within classification categories).

Multinomial regression employs a computational method known as maximum likelihood estimation to derive regression coefficients that can then be used to estimate the cut scores. Unfortunately, the multinomial regression results for this application were highly erratic and, in one case, completely unreasonable from a pragmatic perspective. This was true whether the analyses were carried out "within panelist" or "across panelists". This logistic regression technique was therefore dropped from consideration as a viable method of obtaining stable cut scores.

DFA is another well-known technique for minimizing classification errors when two or more groups or categories are involved (e.g., Klecka, 1980). In this case, there are six categories. Although DFA classifications are actually based on a component score—the discriminant function—there is a direct functional association between the discriminant function component and the observed scores when only one predictor is used.

The third method involves locating the cut score as a point between two distributions (e.g., A1 and A2) where each distribution is represented by a mean (centroid in the multivariate case). The means for the adjacent distributions can be weighted to determine the cut scores. When the weights are chosen to statistically minimize error variances of estimate, the procedure can be called an optimally weighted mean (OWM) method. Two weighting mechanisms were evaluated for the weighted means approach employed here: (1) weighting proportional to frequencies within classes (assigned proficiency levels, A1 to C2) and (2) optimally weighting inversely proportional to the error variance of the mean (Graybill & Deal, 1959). The weights were computed to be proportional to the sample variances of the sample as a means for finding optimal midpoints between the distributions (see Graybill & Deal , 1959, for the optimization rationale). That is, a particular cut score for adjacent categories c and c+1 can be computed as

$$\hat{\mu}\left(\hat{\theta}_{c,c+1}\right)=\frac{\hat{\mu}\left(\hat{\theta}_{c}\right)\hat{\sigma}^{-2}\left(\hat{\theta}_{c}\right)+\hat{\mu}\left(\hat{\theta}_{c+1}\right)\hat{\sigma}^{-2}\left(\hat{\theta}_{c+1}\right)}{\hat{\sigma}^{-2}\left(\hat{\theta}_{c}\right)+\hat{\sigma}^{-2}\left(\hat{\theta}_{c+1}\right)}$$

(Equation 2)

where "^"denotes the estimates of population parameters (conditional means and variances of the distributions of MLE(θ) scores).The choice of weighting method did not produce any substantial difference in the results. Nonetheless, because the OWM approach combines the frequency weighting and the within-class variance in the cut score computation, it was considered to be somewhat more statistically defensible than simply weighting by within-class frequencies.

The final method of computing cut scores is called equipercentile equating (Kolen & Brennan, 2010) and was originally suggested by Michael Kane (personal communication) for use with IJM. Equipercentile equating ensures the proportion of examinees classified into each of the six categories matches the proportions of examinees based on the cut scores—that is, finding the cut scores so that the aggregate proportions match for all six categories. Although equipercentile equating ensures equivalence of the marginal frequency distributions, there is no guarantee that the same examinee classified by the panelists as B1", for example, would classified as "B1" based on his or her test score. This equating approach is conceptually fairly straightforward. We have two distributions of "scores": (1) the distribution of test scores, MLE(θ), and (2) the distribution of ratings, y. Under equipercentile equating, we find the equating transformation function that makes the cumulative frequency distributions of MLE(θ) and y as equal as possible[9]. The equating function can be written as

$$eq_{y}\left(\hat{\theta}\right)=Q^{-1}\left[P\left(\hat{\theta}\right)\right]$$

(Equation 3)

where "^" denotes that we are using estimates of the θ scores, *P*(.) is a cumulative frequency [counting] function and Q^{-1}(.) is the inverse cumulative frequency function. A nice property of this equating approach is that it is symmetric and maintains the SAME percentages in the sampling distributions of test scores—as grouped by the four cut scores—as resulted from the panelists' ratings.

[9] *Frequency estimation was used without pre- or post-smoothing of the cumulative distributions.*

Table 6 contains the combined OWM, DFA and equipercentile cut scores for listening. Table 7 contains the combined cut scores for Reading. The rightmost column in each table contains the average cut score across the three computational methods. It should be apparent that there is more variability in the cuts nearer the extremes. For example, the A1-A2 cuts are more varied across methods than B1-B2. That variability needs to be considered from a robustness perspective (i.e., more variability for the same cut score across computational methods implies less stability in the results and greater susceptibility to assumptions or weights used in the computations). That does not imply that there are inherent problems. Instead, using an average across methods provides a reasonable estimate of each cut, somewhat independent of computational method. The OWM and DFA methods produce similar cut scores. The equipercentile results are somewhat more varied.

TABLE 6.OWM, DFA AND EQUIPERCENTILE CUT SCORES FOR 2014 FOR LISTENING

Cuts	OWM	DFA[a]	Equi-Percentile	Average Cut Score
(A1,A2)	-0.91408	-1.04326	-1.46969	-1.14234
(A2,B1)	-0.36015	-0.39849	-0.61255	-0.45706
(B1,B2)	0.05055	0.04767	0.23019	0.10947
(B2,C1)	0.50117	0.49512	1.25942	0.75190
(C1,C2)	0.67678	0.66299	2.01339	1.11772

TABLE 7.OWM, DFA AND EQUIPERCENTILE CUT SCORES FOR 2014 FOR READING

Cuts	OWM	DFA[a]	Equi-Percentile	Average Cut Score
(A1,A2)	-0.53339	-0.35771	-1.77863	-0.88991
(A2,B1)	-0.10796	-0.10223	-0.45942	-0.22321
(B1,B2)	0.10269	0.08136	0.21111	0.13172
(B2,C1)	0.44724	0.44832	1.10489	0.66682
(C1,C2)	0.7719	0.75957	1.97883	1.17010

The impact of applying these cuts to samples of EF examinees (i.e., examinees previously administered forms and scored on the new EF score scales) are respectively shown in Table 8 (listening) and Table 9 (reading). The cut-score based (assigned) CEFR level is shown in the leftmost column. The frequency of examinees classified into that category is shown in column #2, along with the percentage (relative frequency). The means and standard deviations on the IRT-based scales are shown in the next two columns for the corresponding examinees in each CEFR category. Finally, the "normal equivalents" are shown in the rightmost columns and reflect the likely impact results if the distributions of examinees' scores in the more complete populations of listening and reading examinees are assumed to be "normally distributed" (i.e., bell-shaped). The fact that the percentages correspond to the empirical impact suggests that a normality assumption is reasonable.

TABLE 8. IMPACT OF NEW CUTS FOR LISTENING (N=37,058 EXAMINEES)

			IRT Theta Scores		
Assigned CEFR	**Count**	**Percent**	**Mean**	**Std. Dev.**	**Normal Equivalent Freq. %**
A1	5828	15.73%	-1.9933	0.8983	19.5%
A2	6961	18.78%	-0.7684	0.1948	20.3%
B1	9751	26.31%	-0.1430	0.1611	19.6%
B2	7909	21.34%	0.4101	0.1838	19.4%
C1	2856	7.71%	0.9245	0.1051	8.1%
C2	3753	10.13%	1.7274	0.5871	13.1%
ALL	37058		-0.1617	1.14137	

TABLE 9. IMPACT OF NEW CUTS FOR READING (N=37,159 EXAMINEES)

			IRT Theta Scores		
Assigned CEFR	**Count**	**Percent**	**Mean**	**Std. Dev.**	**Normal Equivalent Freq. %**
A1	6796	18.29%	-1.7133	0.9286	23.4%
A2	8290	22.31%	-0.5263	0.1878	22.0%
B1	6976	18.77%	-0.0363	0.0919	12.9%
B2	7347	19.77%	0.3797	0.1534	17.5%
C1	4281	11.52%	0.8861	0.1413	11.9%
C2	3469	9.34%	1.7290	0.5530	12.2%
ALL	37159		-0.0990	1.0908	

An implication of using the average cut scores from Tables 8 and 9 is that approximately 20 percent of the ER Listening and Reading examinees would likely fall into CE categories A1 to B2. The remaining approximately 20 percent of the examinees would be distributed among the higher C1 and C2 categories. Ultimately, a decision was made to use unweighted averages of the cut scores estimated under all three computational methods. The rationale for averaging makes good statistical sense given some of the disparities between the results under the various cut-score computational methods. A more complete discussion of these computational methods and the derivation of the final, recommended cut scores is provided further on.

The cut scores shown in Tables 6 and 7 serve two purposes. First, they represent plausible ways to classify EF examinees into the CEFR-anchored proficiency categories (A1, A2, B1, B2, C1, and C2) in the future, based solely on their MLE(θ_L) or MLE(θ_L) scores. Second, they can serve to target item writing and test assembly so that measurement precision can be maximized in the region of the cuts to ensure optimal classification decisions. As long as the IRT scales underlying the listening and reading tests are maintained over time through appropriate linking and equating procedures, these cuts should suffice.

Commentary on the IJM Process and the Final Cut Scores

Operationally, there were a few unanticipated data issues during the standard setting, but all were resolved without serious incident or damage to the process. Ultimately, five cut scores for listening and five cut scores for reading were determined (see Tables 6 and 7) and are recommended to EF management. These cut scores are plausible and defensible because the standard-setting "science" behind the IJM, the statistical and psychometric methods used, and the overall standard setting process used were all sound. Given the rather data-intensive nature of the IJM, technical advisors to the EF standard-setting study felt that successful implementation of the IJM would depend on three factors. Those factors were all present for the EF standard setting.

The first success factor was having a sufficient number of well-qualified facilitators. A ratio of one facilitator per three panelists was chosen for this study. The facilitators were NOT envisioned to be passive observers. Each facilitator was expected to be intimately familiar with the proficiency level descriptors, the test materials, the rating process, and the software, as well as knowing when to bring in additional resources (e.g., summoning technical assistance).

The second success factor is to provide full access of the panelists to all information germane to their task. That is, rather than parceling out information in small doses to supposedly make it easier to digest the information, all of the information was available about the examinee response and test materials from the onset. The facilitators were there to teach the panelists how and where to find the information. The organization and presentation of the materials was also optimized from a data visualization standpoint. For example, the assessment task sets were ordered in difficulty, easiest to hardest, and the panelists were able to directly see the responses on the same line in the spreadsheet where they entered each rating. In addition, shading, color-coding border lines, cell over-write protection and other formatting features were included within the rating spreadsheets to help panelists readily distinguish important information and avoid confusion. A limited amount of usability testing was performed with the materials and formats prior to the live standard setting exercise.

The final success factor was providing adequate feedback to the panelists, accompanied by useful information for each panelist on how to use that feedback to change their ratings (if they chose to do so) in order to accomplish a particular goal such as increasing or decreasing the cut scores separating the five CEFR levels. Graphical displays (e.g., Figure 4) of the panelist's aggregate results were used for this purpose, with additional feedback provided to the panelists after Round 1. A brief plenary session was held (all panelists) to explain the graphics and how to interpret and use the information. The graphics appeared were well-received and understood by a majority of the panelists.

In addition to providing credible cut scores for EF, one of the more important contributions of this EF project to the broader area of standard-setting research is that it appears to confirm the practical utility of the IJM. That is, some researchers have avoided the IJM because of the amount of detailed information provided to the panelists and potential cognitive complexity of the rating task. This research study suggests that the IJM can be effectively implemented.

References

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

Cizek, G. (2013). Setting performance standards: Foundations, methods, and innovations, 2nd edition. New York: Routledge.

Graybill, F. A. & Deal, R. B. (1959). Combining unbiased estimators. Biometrics, 15, 543-550.

Hambleton, R. K.; Swaminathan, H. & Rogers, J. (1991). Fundamentals of Item Response Theory. Thousand Oaks, CA: SAGE Publications.

International Test Commission. (2007). ITC guidelines on test use. (www.intestcom.org/itc_projects.htm#ITC Guidelines on Test Use).

Jaeger, R. M. & Mills, C. N. (2001). An integrated judgment method for setting performance standards on complex, large-scale assessments. In G. J. Cizek (Ed.), Setting Performance Standards: Concepts, Methods and Practices, pp. 313-338. Mahwah, NJ: Lawrence Erlbaum Associates.

Klecka, W. R. (1980). Discriminant Analysis Quantitative Applications in the Social Sciences. Thousand Oaks, CA: Sage Publications

Kolen, M. J. & Brennan, R. L. (2010). Test equating, scaling and linking: Methods and practices. New York: Springer.

Luecht, R. M. (2014). Computerized adaptive multistage design considerations and operational issues (pp. 69-83). In D. Yan, A. A. von Davier & C. Lewis (Eds.) Computerized Multistage Testing: Theory and Applications. Taylor-Francis.

Luecht, R. M. & Nungester, R. J. (1998). Some practical applications of computerized adaptive sequential testing. Journal of Educational Measurement, 35, 229-249.

Masters, G. (2010). The Partial Credit Model. In M. Nering and R. Ostini (Eds.), Handbook of Polytomous Item Response Theory Models, pp. 109-122. New York: Routledge Academic.

Raymond, M. R. & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants in standard setting. In G. J. Cizek (Ed.), Setting Performance Standards: Concepts, Methods and Practices, pp. 119-157. Mahwah, NJ: Lawrence Erlbaum Associates.

Wright, B. D. & Masters, G. (1982). Rating Scale Analysis. Rasch Measurement. Chicago, IL: MESA Press.

11. REPORTED SCORE BANDS FOR EF SET AND EF SET PLUS

The final result of the 2014 Standard Setting Study was the establishment of expert-judgment based “cut scores” that define the boundaries of each of the CEFR levels the EF SET measures for reading and for listening. These cut scores are expressed in terms of the theta (θ) scale, which is the psycho- metric scale used in the analysis and scoring of the EF SET tests (see Rasch model analyses). The θ scale has a mean of zero and a range of -3 to +3 in most practical applications (the theoretical range is -6 to +6).

	Cut Scores (Theta or θ)	
CEFR Categories	**Listening**	**Reading**
A1/A2	-1.142	-0.890
A2/B1	-0.457	-0.223
B1/B2	0.109	0.132
B2/C1	0.752	0.667
C1/C2	1.118	1.170

The scores of EF’s two tests, EF SET which is a 50-minute reading and listening test, and EF SET PLUS which is a 120-minute reading and listening test, are reported differently. For the EF SET, the final reading and listening scores are reported as individual modality-specific CEFR classifications based on where the respective final θ scores fall on the scale. A combined score or combined CEFR classification is not reported for EF SET. However, for EF SET PLUS each reading and listening final θ score is reported as a CEFR classification from A1 through C2. In addition, a combined EF SET PLUS total score that averages the 0-100 EF scale scores on the listening and reading tests is also reported for test takers who complete both sections of the EF SET PLUS, and whose results fall within the acceptable limits of the Scoring Matrix explained in Section 12 of this report. A detailed explanation of how the EF 0-100 scale was established and is used can be found on the next page of this report, in the section entitled “The ‘EF Scale’ Reported Score Scale.”

Reporting a reading or listening numerical score to test-takers on the θ scale presents a number of challenges. Firstly, a θ score is difficult to interpret given the negative to positive range on the θ scale (from -3 to +3). Secondly, the cut scores for reading and listening are not uniform, thus introducing a further issue of interpretability. This is due to the fact that the same score achieved for either reading or listening may result in different CEFR classifications, which may confuse the test taker. Therefore, after careful consideration of a number of reporting score scale transformation options, a consensus was reached to report a more comprehensible EF SET PLUS score scale using the linear transformation (LT) approach. The final linear transformation scale
(LTS) is what represents the current operational ‘EF scale’. The EF scale ranges are as follows:

EF scale* ranges (Listening)	EF scale* ranges (Reading)	CEFR classification
1 – 30	1 – 30	A1
31 – 40	31 – 40	A2
41 – 50	41 – 50	B1
51 – 60	51 – 60	B2
61 – 70	61 – 70	C1
71 – 100	71 – 100	C2

*Only the results of EF SET PLUS are reported on the EF scale

The EF Scale Reported Score Scale report examines the different approaches considered by EF in an effort to find a suitable reported score scale, and details the psychometric principles that underlie the linear transformation scale (LTS) approach that was finally adopted by EF.

THE ‘EF SCALE’ REPORTED SCORE SCALE

SEPTEMBER 2014

Richard M. Luecht, Ph.D.
The University of North Carolina at Greensboro

The official scales for the Listening and Reading EF SET are based on item response theory (IRT). Items in the corresponding Listening and Reading item banks are *calibrated* to a metric that psychometricians refer to as "theta" (the Greek "θ"). The calibration process ensures that all estimated scores are on a common scale—the θ scale—regardless of whether an examinee takes an easier or more difficult test form. Examinees can therefore be fairly compared to one another without EF needing to rely on exposing the same test form over time. IRT also makes it possible to implement adaptive testing like the multistage testing framework employed for the EF tests.

Cut scores are set with respect to the θ scales using a well-established process known as *standard setting*. The cut scores are used to classify each examinee taking the Listening and/or Reading tests into one of six Council of Europe Framework of Reference (CEFR) language proficiency categories: A1, A2, B1, B2, C1 or C2. Therefore, we have five new cut scores for the Listening scales and five new cut scores for the Reading scales. The standard setting was carried out using an expert standard-setting panel[1] in July 2014. The technical details of the standard-setting process are described elsewhere. Suffice to say, EF now has an IRT-based scale for Listening, θ_L, and another for Reading, θ_R, with cut scores on the IRT scales corresponding to the values shown in Table 1.

TABLE 1. NEW EF CUT SCORES FROM JULY 2014 STANDARD SETTING

	Cut Scores (Theta or θ)	
CEFR Categories	**Listening**	**Reading**
A1/A2	-1.142	-0.890
A2/B1	-0.457	-0.223
B1/B2	0.109	0.132
B2/C1	0.752	0.667
C1/C2	1.118	1.170

It is important to understand that: (a) θ_L and θ_R are the "official" EF score scales in a technical sense; and (b) the cut scores in Table 1 reflect the "official" CEFR classification boundaries for Listening and Reading. The discussion to follow reflects transformations of the estimated θ_L and θ_R scores to what are commonly called reported scale scores, under the assumption that the reported scale scores may be easier to understand for non-psychometricians.

Transformation Options

There were two score transformation options for EF to consider. The first, termed the RTF Option, reflects the process of using Listening and Reading reference test forms (RTFs) to transform the θ_L and θ_R score estimates to an expected percent correct (EPC) scale that goes from 0 to 100. The second, termed the LTS Option, employs direct [piece-wise] linear transformations (LT) of the θ score estimates to a score scale that likewise ranges from 0 to 100 points[2], with ten points allocated to each of the mid-CEFR levels A2 to C1 (scores points of 41 to 70), and the remaining 61 points allocated to the A2 level (0 to 30 points) and the C2 level (71 to 100 points).

After careful consideration of the options available to EF, the decision was made to proceed with reporting score scales using the LTS option. It is important to realize that there is NO BEST option because the criteria for choosing one option over another may be differentially considered. The simple fact is that any scale transformation carries with it certain practical as well as technical advantages and disadvantages, including operational considerations.

[1] All were highly experienced English language teachers with English as their first language.

[2] There is no explicit requirement to use a scale from 0 to 100. Many testing programs employ scales that may go from 200 to 800 (the SAT in the U.S.), or any other desired values (e.g., 0 to 50 or 100 to 400).

The RTF Option

The RTF Option employs a non-linear transformation of the corresponding IRT θ scale to a scale that reflects expected performance on a manufactured test form known as the reference test form (RTF). The RTF is constructed by test development experts to reflect a reasonable spread of item difficulty and other desirable measurement properties. In a practical sense, the transformation of estimated θ scores to what are called expected raw scores are saying, "If you [the examinee] had taken the RTF, here is how you would be likely to perform on that test, based upon your actual test performance on the [secure] form that you DID take." Since all examinees' expected raw scores are computed using the same RTF, their scores are directly comparable.

In turn, the expected percent-correct (EPC) score[3] can be computed from the expected raw scores on the RTF. To reiterate, the EPC score scale essentially predicts what score an examinee with an estimated θ score would be expected to obtain on the well-chosen RTF set of test items of varying difficulties calibrated to the corresponding θ scale. Given the calibrated item difficulties, bi, for i=1,…,n items, and any associated, calibrated item threshold parameters, dik (for k=0,…,xi points), an EPC can be formally computed as

$$EPC = ROUND_{\text{int}}\left\{100\left[\frac{1}{X^{\max}}\sum\nolimits_{i=1}^{n}\sum\nolimits_{k=0}^{x_i} X_{ik}P_{ik}\left(\hat{\theta}\right)\right]\right\}$$

(Equation 1)

where Pik is the category response probability under Master's (2010) partial-credit model , Xik are the possible score points on each item and Xmax denotes the maximum score points on the RTF. Masters underlying (2010) partial-credit model (PCM) can be written as follows:

$$P\left(x = X\middle|\theta;b_i,\mathbf{d}_i\right) \equiv P_{ix}\left(\theta\right) = \frac{\exp\left[\sum\nolimits_{k=0}^{x}\theta-\left(b_i+d_{ik}\right)\right]}{\sum\limits_{j=0}^{m}\exp\left[\sum\nolimits_{k=0}^{j}\theta-\left(b_i+d_{ik}\right)\right]} = \frac{\exp\left[\sum\nolimits_{k=0}^{x}\theta-b_i-d_{ik}\right]}{\sum\limits_{j=0}^{m}\exp\left[\sum\nolimits_{k=0}^{j}\theta-b_i-d_{ik}\right]}$$

(Equation 2)

where θ is the examinee's proficiency score, bi denotes an item difficulty or location (on the θ scale) for item i, and dik denotes two or more threshold parameters associated separations of the category points for items that use three or more score points (k=0,…,xi). For items scored correct-incorrect, the dik parameters disappear from the model, reducing the model to what is typically called the "Rasch model".

The bottom line is that, because the EPC scores are based on the RTF items, which in turn are calibrated to the same scale used to estimate the examinees' θ scores, all reported EPC scores are on the SAME scale. As percent-correct scores, the EPC scale retains the intuitive appearance and understanding of more typical percent-correct scores used in classroom and other tests. For example, teachers and students seem to intuitively understand what a percent-correct score of 80 percent means . However, unlike observed number-correct or percent-correct scores, RTF-based EPCs can also be used with adaptive tests where the difficulty of each test form is tailored to every examinee's apparent level of proficiency.

Another advantage of the RTF Option and using EPC scoring is the automatic handling of lowest obtainable scale scores (LOSS) and highest obtainable scale scores (HOSS) issues. Many testing programs that employ a singular linear transformation must contend with LOSS and HOSS issues that usually require truncation of the computed scale scores near the tails of the score distribution. For example, if the LOSS is set at 0 and HOSS at 100, any computed scores below or above that range—which is indeed possible—are truncated to 0 or 100.

The EPC scale scores also deal with a technical complication of very large estimation errors near the tails of the distribution when IRT scores are computed. All scores contain some error. The EPC is a robust method of scaling that retains the intuitive value of percent-correct scoring without sacrificing technical quality or comparability of the reported scores.

There are two disadvantages of implementing the RTF option—that is, reporting EPC scores. The first is operational complexity. The software resources and quality control steps needed to implement an RTF scoring framework is extremely complicated for adaptive multistage tests like the EF SET. Although it has been successfully done at EF for the past year, maintaining that

[3] An EPC is a special case of what previous IRT literature has called a 'domain score'.

system for new test forms is complicated and expensive. The second disadvantage is that the cut scores do not resolve to easily interpretable numbers —for example, a particular RTF-based cut score on the EPC value of EPC(θ_{cut}) = 46.58 may seem strange and difficult to understand for the test takers and other non-technical score users. Both disadvantages ultimately drove the decision to move to the LTS option described in the next section.

The LTS Option

The reported score scale using the LT can be expressed as a rounded integer value between 0 and 100 points (see footnote #2), using the equation for a line,

$$y = round\left(A_c + B_c \cdot \hat{\theta}_{j|c}\right)$$

(Equation 3)

with linear coefficients consisting of the slope, Bc, and intercept Ac. The two coefficients are specific to each of the six CEFR categories (c=1, 2,…,6) where the estimated θ (denoted by the "^") falls with the interval bounded by the cuts scores for adjacent CEFR categories (see Table 1). We can further conveniently specify the desired minimum and maximum scale scores for each CEFR interval (e.g., 0-30 for A1, 31-40 for A2, etc.).

The piece-wise linear transformation constants in Equation 3 are simple to compute. The slope is computed as

$$B_c = \frac{\max\left(y_c\right) - \min\left(y_c\right)}{\left(\hat{\theta}_{c+1} - \varepsilon\right) - \hat{\theta}_c}$$

(Equation 4)

and the intercept is calculated as

$$A_c = \min\left(y_c\right) - B_c \cdot \hat{\theta}_c$$

(Equation 5)

where min ($\hat{\theta}_c$) is the cut score on the estimated θ_L or θ_R scale (Table 1), c is the CEFR category index (1=A1, 2=A2, etc.), within each CEFR interval and max($\hat{\theta}_{c+1} - \varepsilon$) denotes a value just below the cut for the next higher category (for example, ε=0.0001). Table 2 shows the corresponding slope and intercept terms that would result if we want to retain a LTS scale with point values ranging from 0 to 100 (e.g., with 10 points within the intervals for A2 to C1, 31 points allocated within the A1 interval and 30 points for C2 examinees[6] as shown in these examples). Note that the decision to increase the number of points within the A1 and C2 categories—rather than having approximately equal points in every category—was intentional and allows for more score flexibility in the scale near the tails. The minimum and maximum values on the corresponding θ scales are set at –7.5 and +7.5, respectively. Practically speaking, that range would cover virtually any and all estimated θ scores and any needed truncation beyond that range would be completely trivial. The implication is that the transformation coefficients shown in Table 2 COULD be operationally implemented for the EF tests.

[6] To keep the minimum and maximum scores at integer values—for example, having category A2 end at 33 and B1 start at 34 points, one of the interval sizes had to be reduced to 15 points.

TABLE 2. LTS INTERCEPTS AND SLOPES FOR LISTENING AND READING

				Scaling Constants			
	Piece-wise Linear Transforms			Listening		Reading	
CEFR Label	Max(y)	Max(y)	Interval Size	Slope	Intercept	Slope	Intercept
A1	0	30	31	4.7188	35.3909	4.5386	34.0394
A2	31	40	10	13.1352	46.0049	13.5014	43.0150
B1	41	50	10	15.8890	48.2622	25.3643	46.6616
B2	51	60	10	14.0115	49.4662	16.8224	48.7841
C1	61	70	10	24.6090	42.4965	17.8862	49.0731
C2	71	100	30	4.5438	65.9213	4.5814	65.6393

Computationally, the LTS approach is far less intensive than the RTF option described earlier because there are fewer variables to deal with (e.g., no IRT-based RTF item statistics needed). A simple look-up can be used to determine which slope and intercept pair to use in the computations using only the examinee's estimated θ score and the corresponding standard-setting cut scores needed for classifying him or her into the corresponding CEFR category. This LTS method can also be readily applied with IRT-based "score look-up" tables (as is now done).

Implementation mechanics aside, the unit size for the scaling does differ across the six CEFR intervals (see Slope columns in Table 2). For example, two examinees with θ_L score estimates that are 0.1 units apart on the θ scale within category A2 would be about 2.3 points apart on the reported LTS scale for Listening (slope of 23.352 × 0.1 ≈ 2.3). The same difference of 0.1 estimated "θ units" would, however produce a 4.4 point difference in scale score units within category C1, where the slope is larger (43.749). Note that the same could be said of the RTF scores, but for different reasons. With the LTS approach we are artificially stretching or pulling the score points to achieve an equal-sized intervals on the reported scale across categories, even though the standard setting study produced cut scores on the θ scales that were NOT equally spaced.

Provided that estimated IRT scores are truncated to the range $-7.5 \leq \theta \leq 7.5$, there are no LOSS or HOSS issues with the LTS approach, nor is there any possibility of "overlap" of the reported scores across CEFR categories, provided that the classifications of examinees to intervals is accurately done (i.e., the correct linear transformation scaling coefficients are used).

One important consideration to consider using the LTS Option is that, should the cut scores change in the future, a decision would be needed as to whether to continue to use the LTS transformation coefficients from Table 2 or compute new values that took those new points into account. The coefficients could certainly be changed to hold constant the boundary cut points on the LTS reporting scale at the current boundary values (A1 = 0–30, A2 = 31–40, etc.; see Table 2), with the important behind-the-scenes technical recognition that the substantive and normative meaning of those score intervals could change if the change in cut scores were likewise substantial.

Some Empirical Results for the LTS Option

Two very large samples of examinees have taken the Listening and Reading examinations to date, with counts (N) of NL=37,059 and NR=37,160. Theta (θ) score estimates were obtained for each examinee on the corresponding IRT scale. Those θ-estimates were then transformed to LTS scores (i.e., applying the LTS option, using the coefficients in Table 2). The Listening and Reading results are presented in Table 3.

TABLE 3. DESCRIPTIVE STATISTICS FOR LISTENING AND READING

	Listening		Reading	
Statistics	**IRT θ_L**	**LTS_L Score**	**IRT θ_R**	**LTS_R Score**
Minimum	-7.000	2.359	-6.380	5.083
Maximum	4.410	85.960	4.750	87.401
Mean	-0.162	47.217	-0.099	46.657
Std. Deviation	1.141	14.623	1.091	15.189

The means (average scores) and standard deviations for LTS scores are comparable for the Reading and Listening tests. This suggests that any normative interpretations will likewise be comparable.   The θ_L score estimates are somewhat "normal" (bell-shaped in appearance) with the smallest concentration of examinees near the tails, as we might expect with large-sample assessment data drawn from a general population.

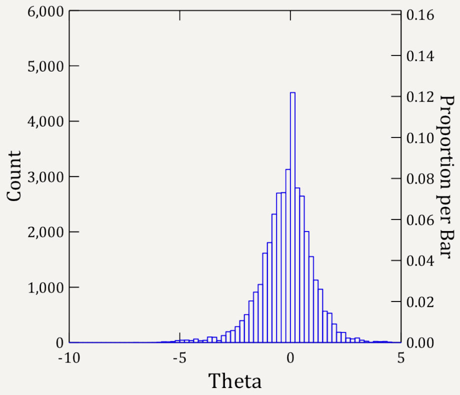


Figure 1. Histogram of IRT θ_L Estimates (N_L=37,059 EF Examinees)

The histogram for the LTS scores is displayed in Figure 2. Figures 1 and 2, jointly considered, demonstrate how the piece-wise LTS transformation spreads out the rescaled θ_L scores. That larger spread should not be surprising considering that the piece-wise LTS Option applied differential transformations across the six CEFR categories, with some of the slopes being quite large.

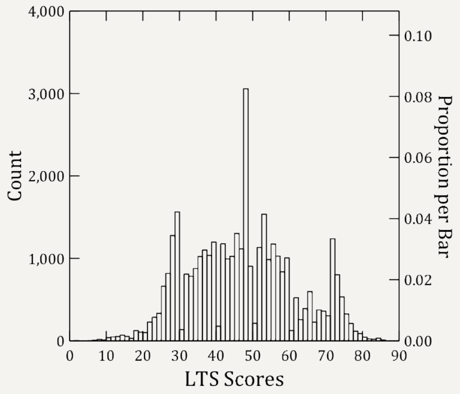


Figure 2. Histogram for Listening LTS Scores (NL=37,059 Examinees)

Figures 3 and 4 correspondingly show the histograms for the θ_R and LTSR (transformed scores). The distributions are similar to the Listening results discussed above.

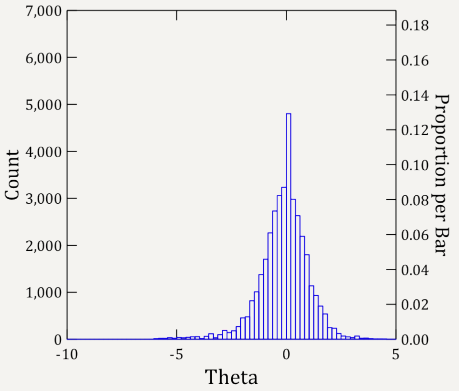


Figure 3. Histogram of IRT θ_R Estimates (N_R=37,160 EF Examinees)

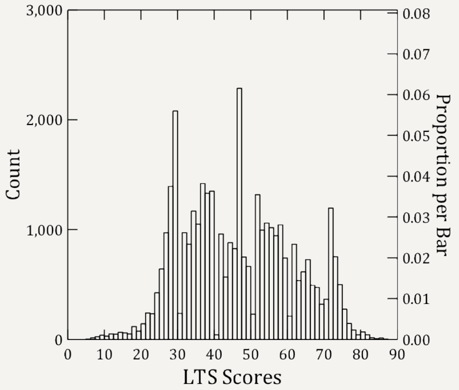


Figure 4. Histogram for Reading LTS Scores (N_R=37,160 Examinees)

Another important consideration is the impact of the scaling on estimation errors (or measurement errors). Figure 5 shows the side-by-side summary dot plots of the standard errors for Listening score estimates on the θ_L scale (left side) and for the LTS scores (right side).

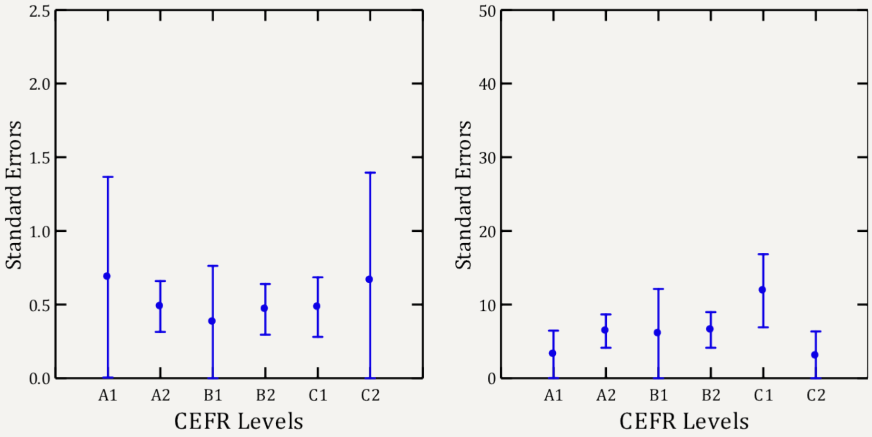


Figure 5. Standard Errors by CEFR Level, SE(θ_L) on Left; SE(LTSL) on Right

Each dot represents the mean or average standard error within each of the CEFR categories, based on the 37,059 EF Listening examinees' –standard errors of the θ_L score estimates on the left; standard errors of the LTS scores on the right.. The error bands reflect the sampling distribution of the standard errors within CEFR categories (the bands capture one standard deviation or 95% of the standard errors). Note that standard errors for the IRT θ estimates are more "U-shaped" across the six categories, with the largest errors near the tails. In contrast, the LTS scores have somewhat uniform standard errors across the categories. For some measurement practitioners, having uniform errors across the scale is highly desirable. The LTS option helps in that regard.

Figure 6 shows corresponding Reading standard errors of the θ_R score estimates on left; for the LTSR scores on the right (95% standard deviation bands shown). It should be apparent that the "U" shaped pattern common to IRT-based SE(θ) estimates is offset by the LTS piece-wise transformation.

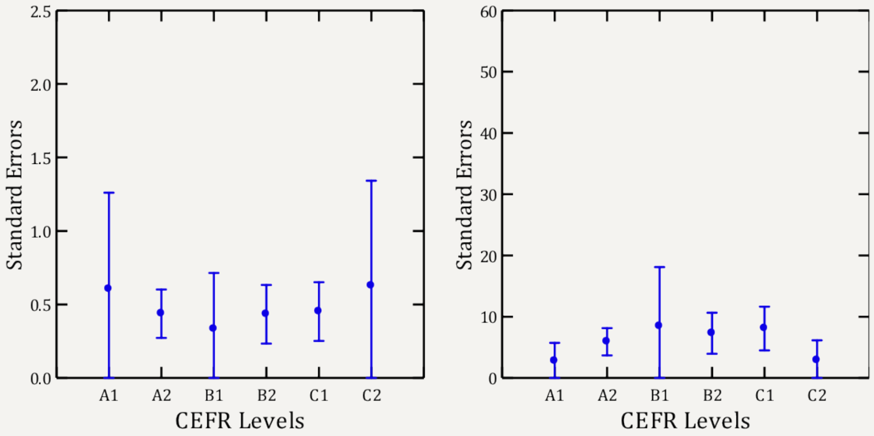


Figure 6. Standard Errors by CEFR Level, SE(θ_R) on Left; SE(LTSR) on Right

To summarize, the LTS Option creates a more uniform spread of scores within each CEFR category and an overall "flatter" distribution of standard errors across the entire score scale.

Some Final Comments

Many testing programs want point values on the reported scale to be meaningful. For example, taking into account that there is error in all test scores, we might want to conclude that a 5 or 6 point difference on the scale is "real" insofar as talking about one set of scores as being statistically higher than other scores, regardless of where along the scale we happen to be making such comparisons. That would argue for a scale like the EPC with somewhat uniform standard errors across the scale.

While there are a number of ways to represent a reported scale, a consensus was agreed to implement a scale that is easily interpretable, operationally less complex to generate, and provides reasonable scoring quality opportunites. The LTS option gives us that.

Adopting the LTS option will entail, by default, the need to recompute the linear transformations within each of the CEFR categories in the future if the Listening or Reading cut scores change, as noted earlier. This would be necessary to retain the same equivalent ranges. It is not an overwhelming problem, but would effectively create a discontinuity between the "older" and "newer" LTS Option reported score scales. The implications of those changes on interpretative materials and uses of the scores over time would need to be considered as well.

A final consideration is the interpretation of performance relative to the reported scale. The six CEFR levels already provide a general level of interpretation regarding English language proficiency. For example, most teachers and others can ascertain an examinees levels of English language listening and reading proficiency just by knowing that the examinee is an "B1" or "C1".

Similar to the an "advantage" mentioned for the RTF option, the LTS pption can also be tied to likely performance by a technique known as "scale anchoring". Scale anchoring chooses a reasonable level of expected successful performance such as having an 80% chance of answering a particular item. Select items are then anchored or located along the θ scale (or the corresponding LTS Option scale) to show the examinees the types of items they are likely to perform successfully. By considering items somewhat higher than their own scores, the examinees can also see a progression of the types of tasks they need to master to improve their scores. Scale anchoring has been effectively used for many years with the National Assessment of Educational Progress (NAEP) in the U.S.

References

Masters, G. (2010). The Partial Credit Model. In M. Nering and R. Ostini (Eds.), Handbook of Polytomous Item Response Theory Models, pp. 109-122. New York: Routledge Academic.

12. COMBINED SCORES ON EF SET PLUS: THE SCORING MATRIX

For language learners, reading and listening present different challenges. Performance on test questions that use one modality will often not be in the same place on the score scale as performance in the other with sometimes great disparities. For this reason, caution was taken when deciding when to report final combined scores for specific combinations of listening and reading EF SET PLUS results, and when not to report combined scores at all.

The ‘scoring matrix’ below shows all possible CEFR band scores for listening and reading in EF SET PLUS – the top row and first column represent listening or reading CEFR classifications. The three shaded areas or ‘zones’ represent different combined-score logic scenarios. The green zone represents the possible CEFR band score combinations that allow reporting a combined reading and listening tests score. The yellow cells represent the CEFR combinations where a combined EF score will be reported, but this score should be interpreted with caution. A cautionary note is displayed to the test taker alongside his/her test results if the EF SET PLUS listening and reading results combination falls in the yellow zone. The red area indicates the combination of CEFR band scores for listening and reading that are too disparate to warrant a combined score that is meaningful. Thus, in this particular area, a combined score is not reported because it cannot be reasonably estimated.

Also, if either the listening or reading test is not attempted (a score of 0 is achieved), then no combined score is reported in this case.

CEFR	A1	A2	B1	B2	C1	C2
A1	🟩	🟩	🟩	🟨	🟥	🟥
A2	🟩	🟩	🟩	🟩	🟥	🟥
B1	🟩	🟩	🟩	🟩	🟩	🟩
B2	🟨	🟩	🟩	🟩	🟩	🟩
C1	🟥	🟥	🟩	🟩	🟩	🟩
C2	🟥	🟥	🟩	🟩	🟩	🟩

APPENDIX A
EXAMPLES OF ORIENTATION MATERIALS SENT TO PANELISTS

EF Education First

EF STANDARD SETTING MEETING

SATURDAY, JULY 5TH 2014

Evaluating the test material

On Saturday, 5th July, you will be evaluating actual test takers' responses to listening and reading tasks. In each 'work session' on the 5th July, you will:

- be assigned to one of five groups (Group A, B, C, D, or E) - you will be informed which group you will assigned to before you arrive.
- rate actual test takers' responses to four listening tasks and four reading tasks (of varying difficulty).
- classify each test taker (based on their overall response patterns) into one of the CEFR levels based on your expert judgment.

For you reference, you can download the Common European Framework of Reference for Languages: Learning, Teaching, Assessment document here:

http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf 6

Preparation before the meeting

- In the **week prior** to the Standard Setting meeting we expect you to look at and engage with four listening tasks and four reading tasks that you will be rating on the day of meeting as a 'test taker' yourself. The purpose of doing so is to give you a better understanding of how the test takers interact with the tasks in a testing environment.
- While all the tasks you will see in the form are actual test tasks that have been taken by students we have re-combined them in a form that is specific to your assigned group. It should take you no longer than 2 hours to answer all the questions (for the four listening and four reading tasks). Please remember that it is important that every participant simulates being a test taker.

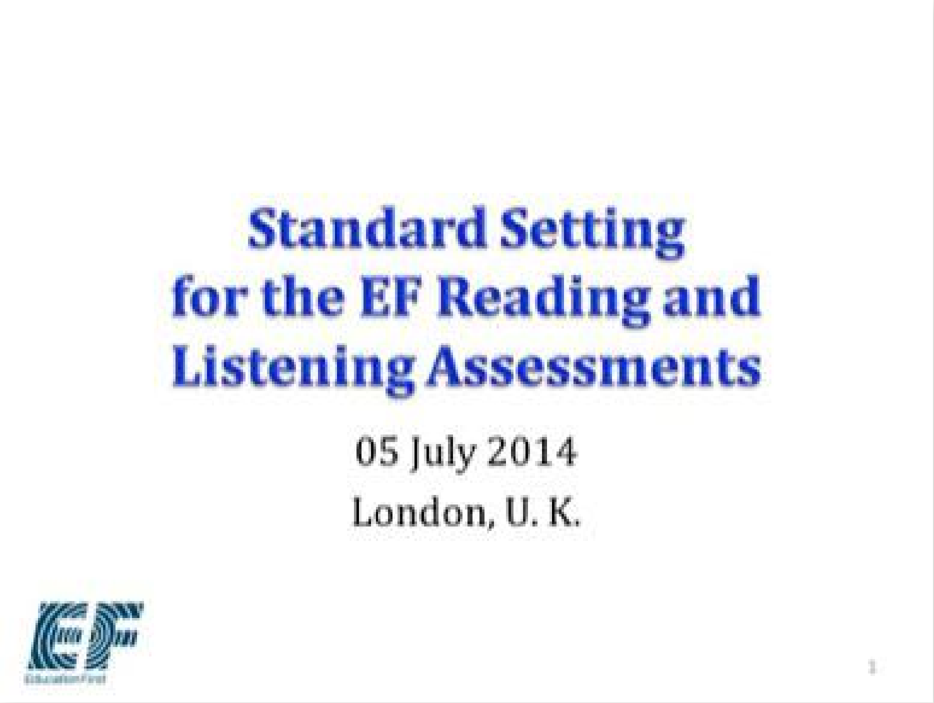
  - We will be sending you each an individual email which will include:

  - The URL to access the 'test' form

  - Your unique username and password that you need to enter

  - A code which you will need to access your assigned group's tasks

APPENDIX B
OPENING PLENARY SESSION PRESENTATION



Standard Setting for the EF Reading and Listening Assessments

05 July 2014

London, U. K.

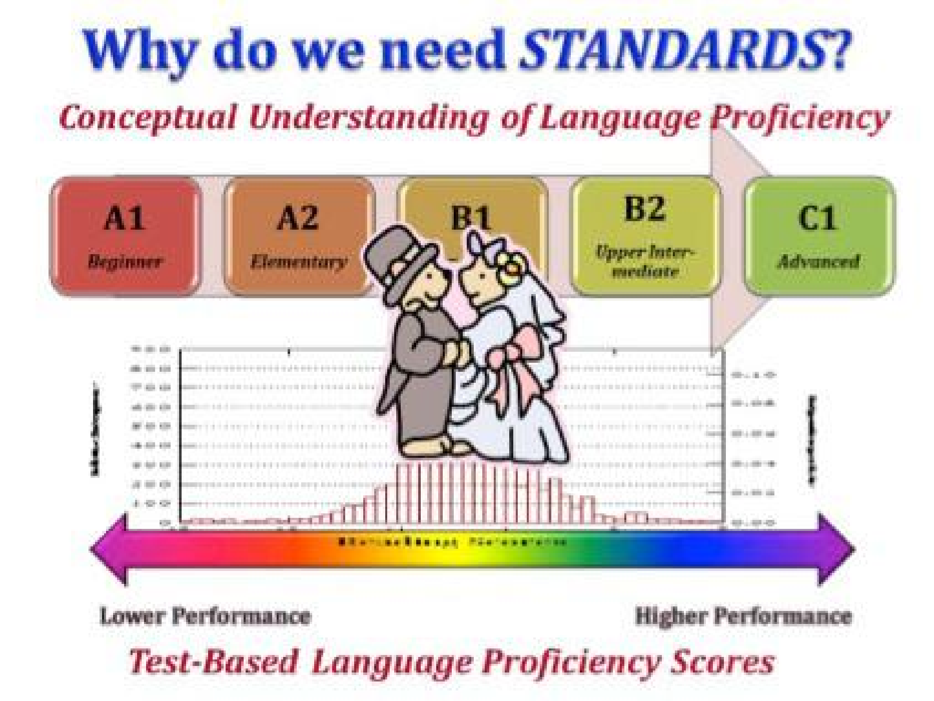
Education First

1

Welcome!

2

APPENDIX B
OPENING PLENARY SESSION PRESENTATION (CONTINUED)



Why do we need *STANDARDS*?

Conceptual Understanding of Language Proficiency

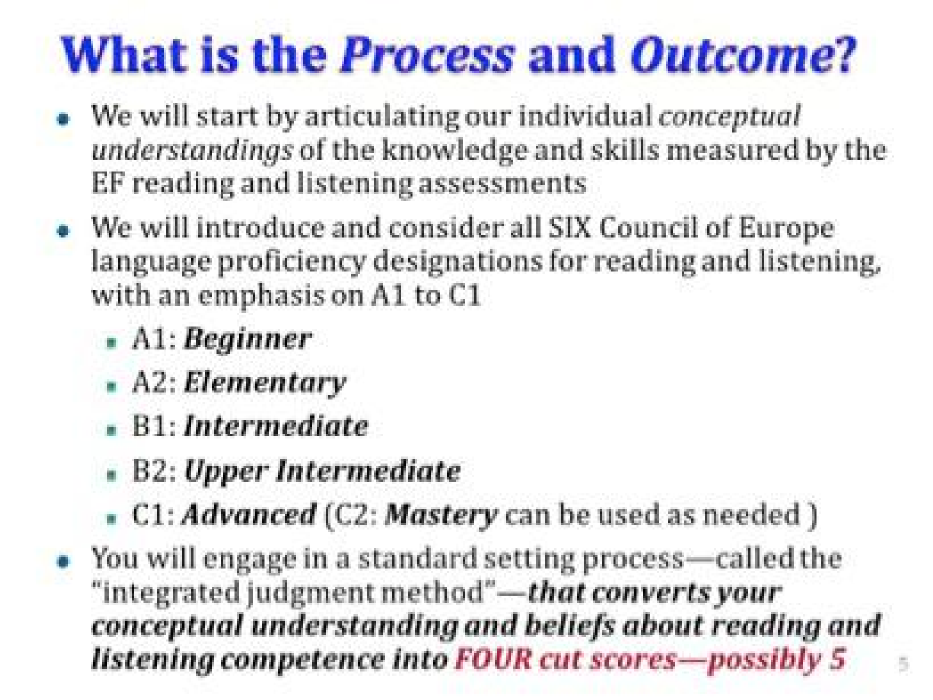
Lower Performance

Higher Performance

Test-Based Language Proficiency Scores

YOU are [collectively] the *EF Standard Setting Panel*!

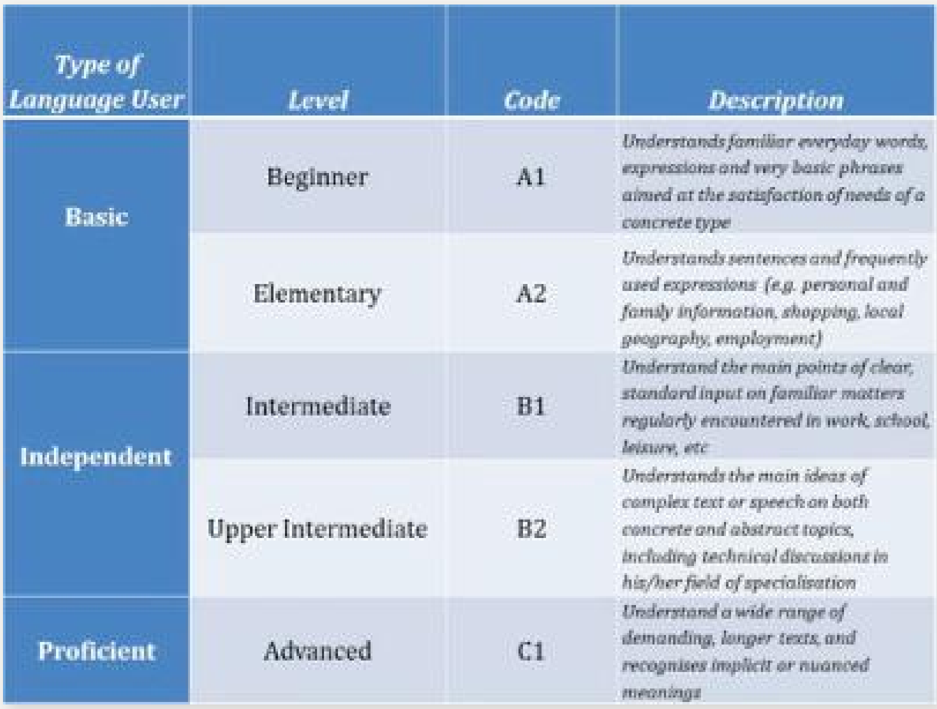
APPENDIX B
OPENING PLENARY SESSION PRESENTATION (CONTINUED)



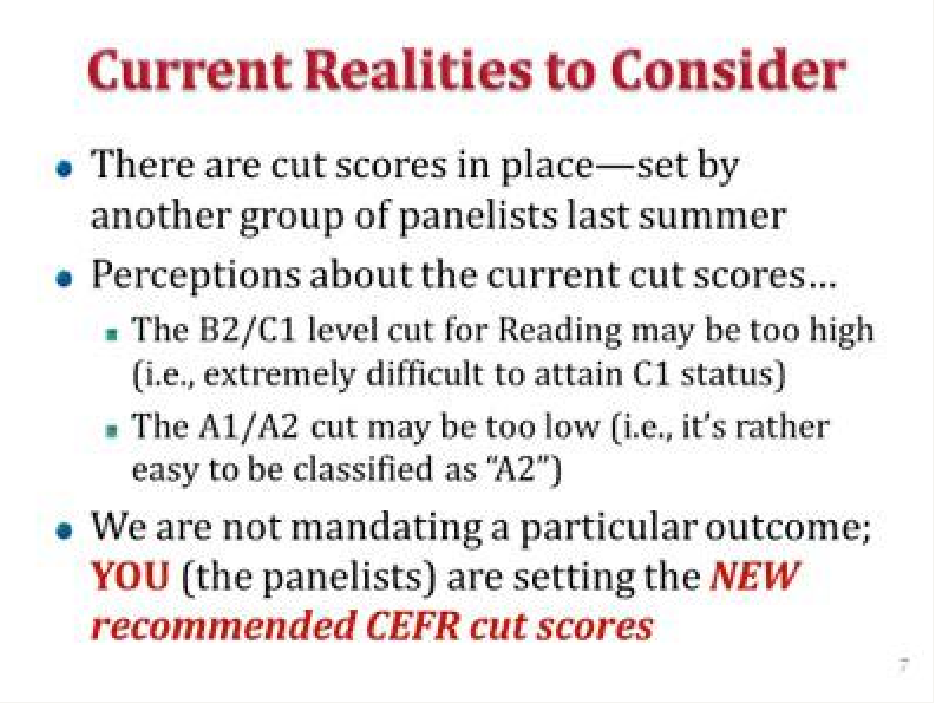
What is the *Process* and *Outcome*?

- We will start by articulating our individual *conceptual understandings* of the knowledge and skills measured by the EF reading and listening assessments
- We will introduce and consider all SIX Council of Europe language proficiency designations for reading and listening, with an emphasis on A1 to C1
  - A1: ***Beginner***
  - A2: ***Elementary***
  - B1: ***Intermediate***
  - B2: ***Upper Intermediate***
  - C1: ***Advanced*** (C2: ***Mastery*** can be used as needed)
- You will engage in a standard setting process—called the "integrated judgment method"—***that converts your conceptual understanding and beliefs about reading and listening competence into FOUR cut scores—possibly 5***

5



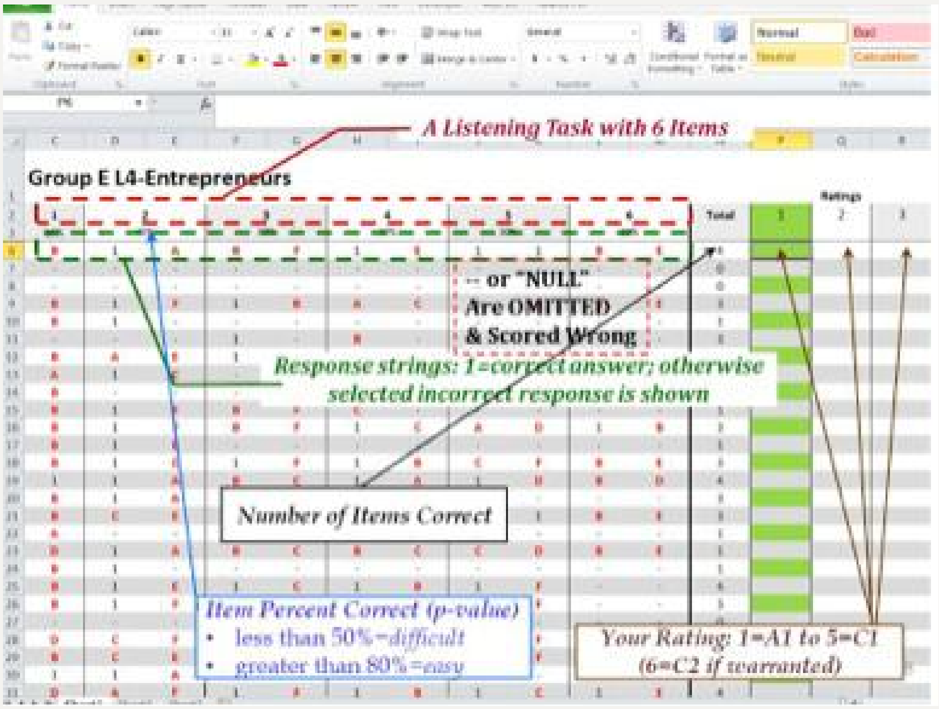
Type of Language User	*Level*	*Code*	*Description*
Basic	Beginner	A1	*Understands familiar everyday words, expressions and very basic phrases aimed at the satisfaction of needs of a concrete type*
	Elementary	A2	*Understands sentences and frequently used expressions (e.g. personal and family information, shopping, local geography, employment)*
Independent	Intermediate	B1	*Understand the main points of clear, standard input on familiar matters regularly encountered in work, school, leisure, etc*
	Upper Intermediate	B2	*Understands the main ideas of complex text or speech on both concrete and abstract topics, including technical discussions in his/her field of specialisation*
Proficient	Advanced	C1	*Understand a wide range of demanding, longer texts, and recognises implicit or nuanced meanings*



Current Realities to Consider

- There are cut scores in place—set by another group of panelists last summer
- Perceptions about the current cut scores…
  - The B2/C1 level cut for Reading may be too high (i.e., extremely difficult to attain C1 status)
  - The A1/A2 cut may be too low (i.e., it's rather easy to be classified as "A2")
- We are not mandating a particular outcome; **YOU** (the panelists) are setting the ***NEW recommended CEFR cut scores***

7



A Listening Task with 6 Items

Group E L4-Entrepreneurs

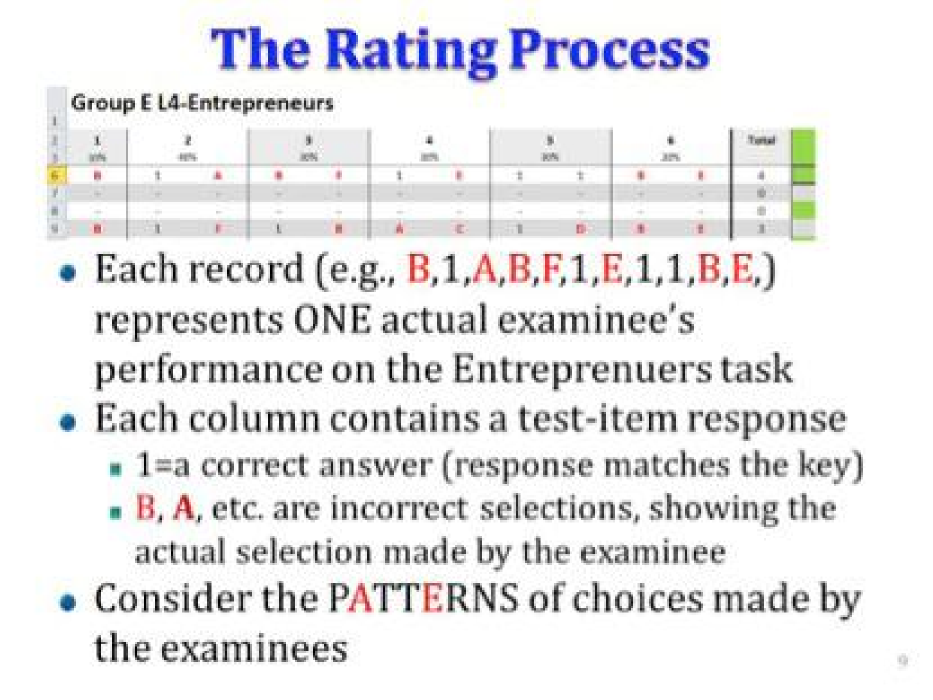
-- or "NULL" Are OMITTED & Scored Wrong

Response strings: 1=correct answer; otherwise selected incorrect response is shown

Number of Items Correct

Item Percent Correct (p-value)
- *less than 50%=difficult*
- *greater than 80%=easy*

Your Rating: 1=A1 to 5=C1 (6=C2 if warranted)



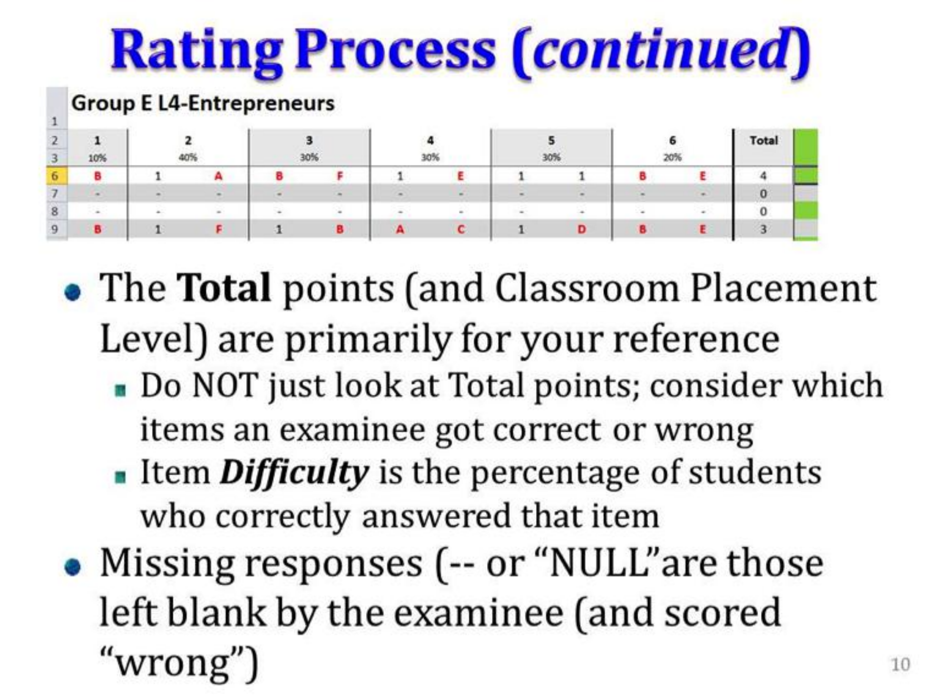
The Rating Process

Group E L4-Entrepreneurs

1													
2	1	2		3		4		5		6		Total	🟩
3	10%	40%		30%		30%		30%		20%			🟩
6	B	1	A	B	F	1	E	1	1	B	E	4	🟩
7	-	-	-	-	-	-	-	-	-	-	-	0	
8	-	-	-	-	-	-	-	-	-	-	-	0	🟩
9	B	1	F	1	B	A	C	1	D	B	E	3	

- Each record (e.g., B,1,A,B,F,1,E,1,1,B,E,) represents ONE actual examinee's performance on the Entreprenuers task
- Each column contains a test-item response
  - 1=a correct answer (response matches the key)
  - B, A, etc. are incorrect selections, showing the actual selection made by the examinee
- Consider the PATTERNS of choices made by the examinees

9



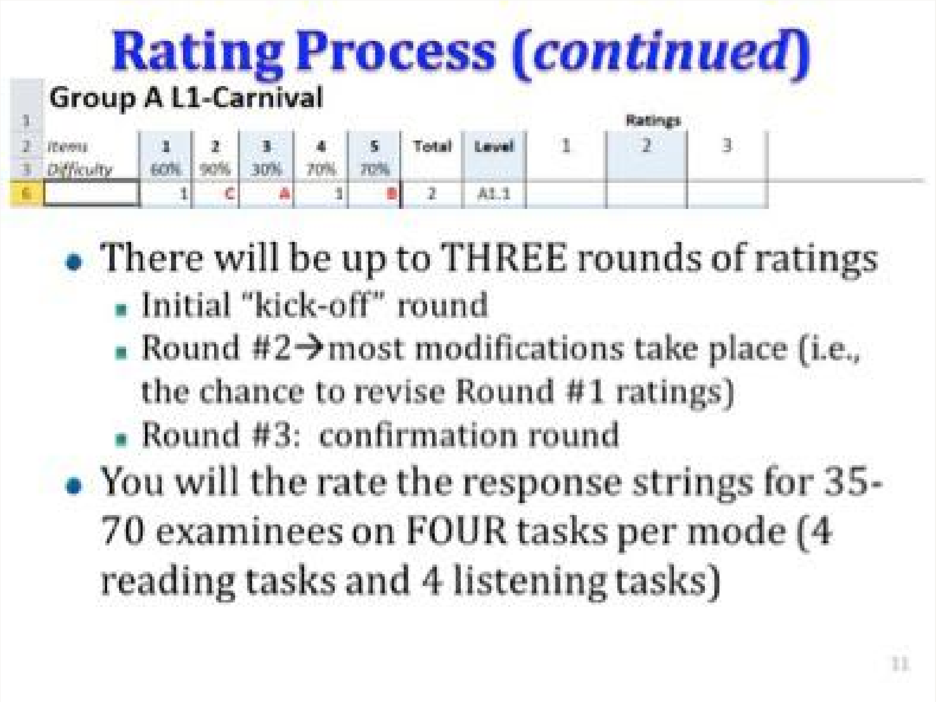
Rating Process (*continued*)

Group E L4-Entrepreneurs

1													
2	1	2		3		4		5		6		Total	🟩
3	10%	40%		30%		30%		30%		20%			🟩
6	B	1	A	B	F	1	E	1	1	B	E	4	🟩
7	-	-	-	-	-	-	-	-	-	-	-	0	
8	-	-	-	-	-	-	-	-	-	-	-	0	🟩
9	B	1	F	1	B	A	C	1	D	B	E	3	

- The **Total** points (and Classroom Placement Level) are primarily for your reference
  - Do NOT just look at Total points; consider which items an examinee got correct or wrong
  - Item ***Difficulty*** is the percentage of students who correctly answered that item
- Missing responses (-- or "NULL"are those left blank by the examinee (and scored "wrong")

10



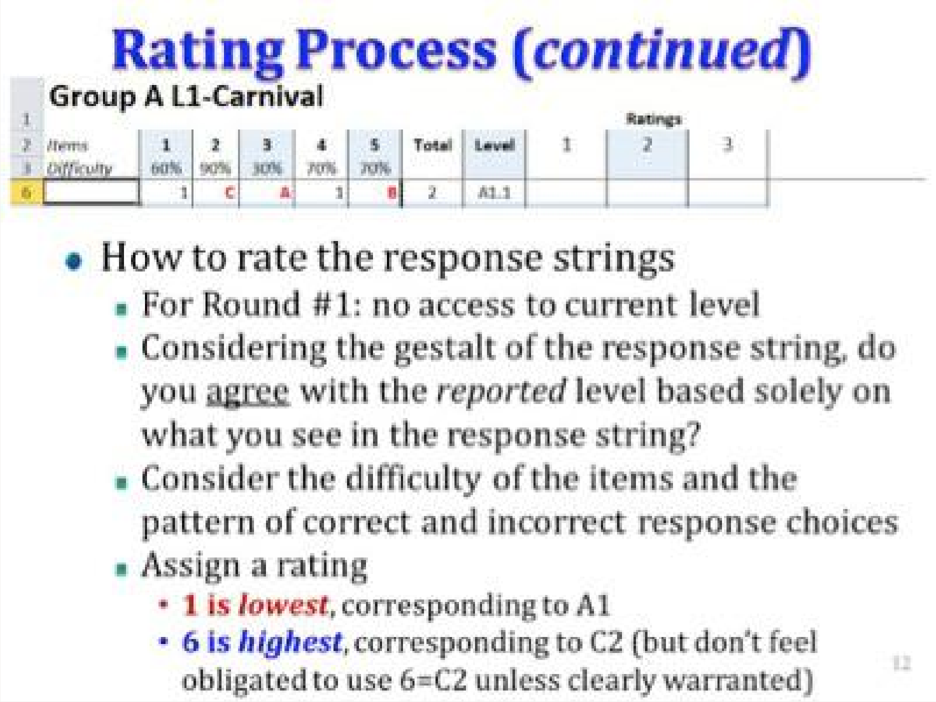
Rating Process (*continued*)

Group A L1-Carnival

1										Ratings	
2	*Items*	1	2	3	4	5	Total	Level	1	2	3
3	*Difficulty*	60%	90%	30%	70%	70%					
6		1	C	A	1	B	2	A1.1			

- There will be up to THREE rounds of ratings
  - Initial "kick-off" round
  - Round #2→most modifications take place (i.e., the chance to revise Round #1 ratings)
  - Round #3: confirmation round
- You will the rate the response strings for 35-70 examinees on FOUR tasks per mode (4 reading tasks and 4 listening tasks)

11



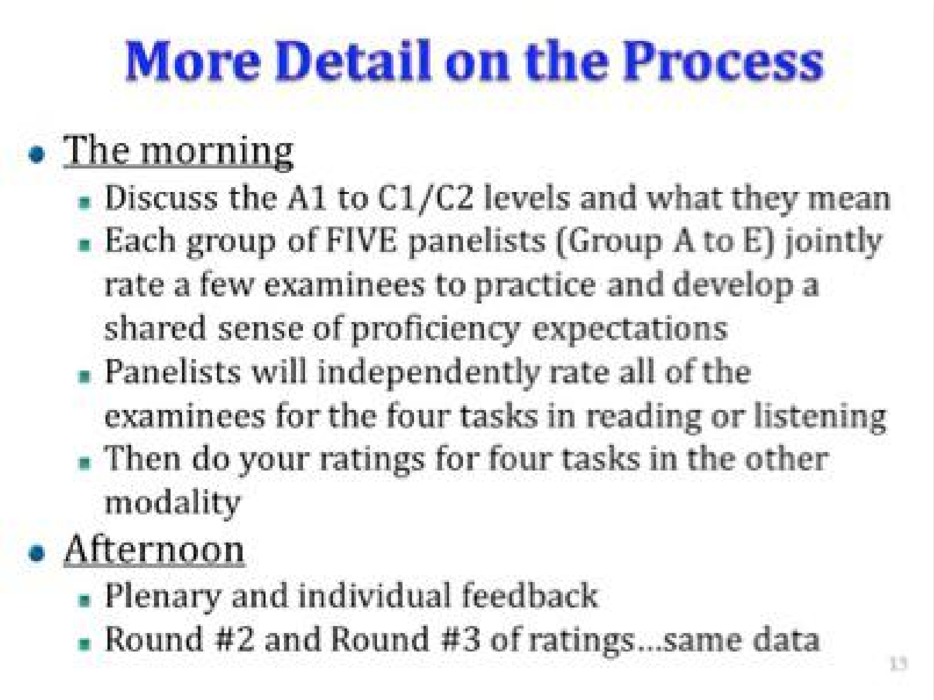
Rating Process (*continued*)

Group A L1-Carnival

1										Ratings	
2	*Items*	1	2	3	4	5	Total	Level	1	2	3
3	*Difficulty*	60%	90%	30%	70%	70%					
6		1	C	A	1	B	2	A1.1			

- How to rate the response strings
  - For Round #1: no access to current level
  - Considering the gestalt of the response string, do you agree with the *reported* level based solely on what you see in the response string?
  - Consider the difficulty of the items and the pattern of correct and incorrect response choices
  - Assign a rating
    - **1 is *lowest***, corresponding to A1
    - **6 is *highest***, corresponding to C2 (but don't feel obligated to use 6=C2 unless clearly warranted)

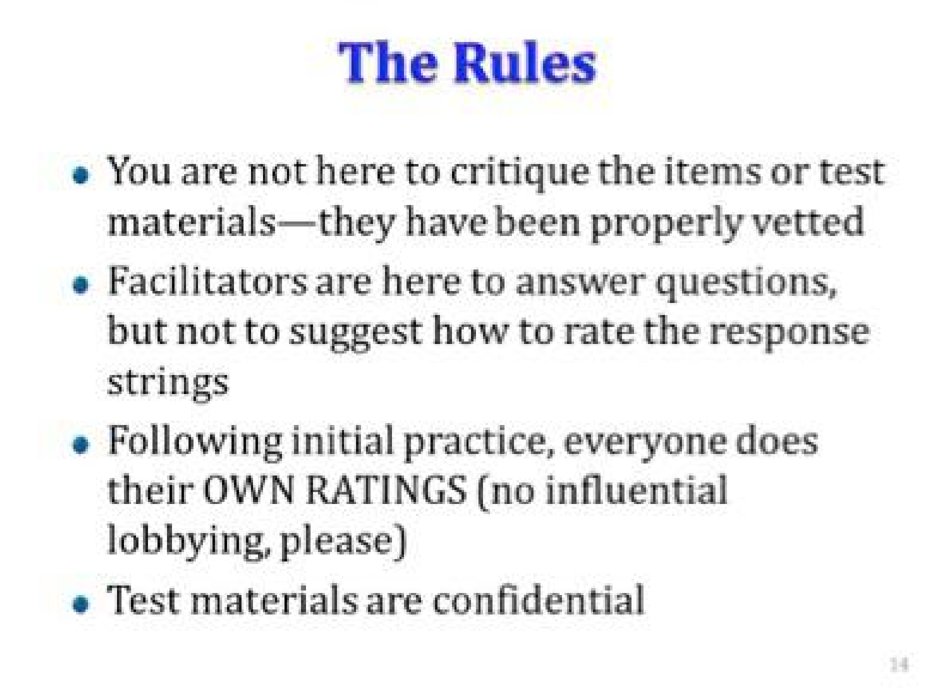
12



More Detail on the Process

- The morning
  - Discuss the A1 to C1/C2 levels and what they mean
  - Each group of FIVE panelists (Group A to E) jointly rate a few examinees to practice and develop a shared sense of proficiency expectations
  - Panelists will independently rate all of the examinees for the four tasks in reading or listening
  - Then do your ratings for four tasks in the other modality
- Afternoon
  - Plenary and individual feedback
  - Round #2 and Round #3 of ratings…same data

13



The Rules

- You are not here to critique the items or test materials—they have been properly vetted
- Facilitators are here to answer questions, but not to suggest how to rate the response strings
- Following initial practice, everyone does their OWN RATINGS (no influential lobbying, please)
- Test materials are confidential

14

Díky

Vielen Dank

תודה

Gracias

Merci

Köszönettel

Thank You

Ευχαριστώ

شكراً

Teşekkürler

Hvala

ขอบคุณ

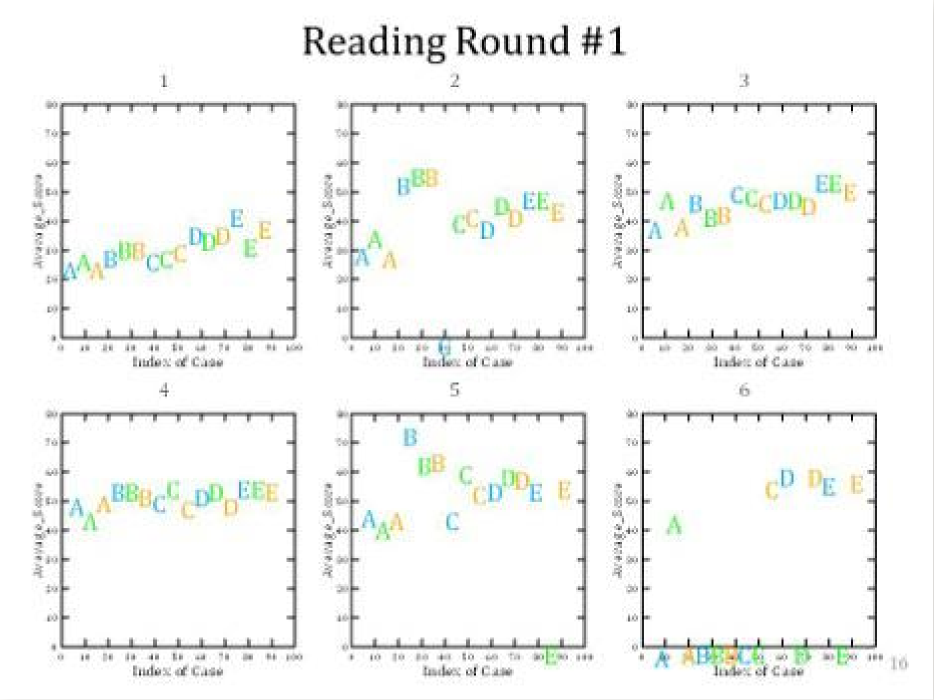
Bedankt

Grazie

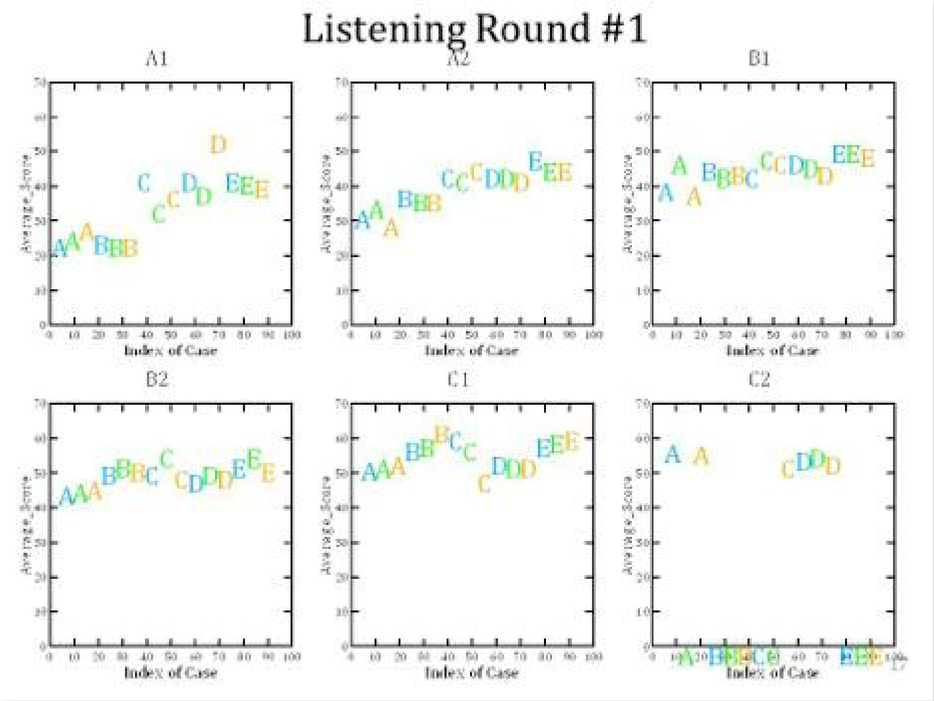
15

APPENDIX C

SLIDES PROVIDED TO PANELISTS BETWEEN ROUNDS (SEE TEXT FOR DESCRIPTION)



Reading Round #1



Listening Round #1

WWW.EFSET.ORG

Education First