Dynamic Placement Test Standard Setting and Benchmarking

9-11 May 2019

ClarityEnglish telc LANGUAGE TESTS

Dynamic Placement Test (DPT) Receptive Skills Standard Setting

Hong Kong, 9 – 11 May 2019

Introduction

This document reports on a Standard Setting workshop run to establish and improve the quality of items used in the ClarityEnglish Dynamic Placement Test. The objective of this published report is to describe the workshop setup and process and to give the reader an overview of what was achieved. The report does not compromise the security of test items; those readers with a legitimate need to examine the output of the analyses should contact the author.

Introduction 1

The test 2

- Part 1 - Gauge 3
- Part 2 - Track 3
- Tasks and the CEFR 3

The procedure for the DPT Standard Setting 4

- Thursday: introduction and orientation 4
- Friday: CEFR familiarisation and standard setting 4
- Saturday: Standard setting 5

The judges 5

Conceptualisation of the Minimally Competent Person (MCP) 6

- MCP conceptualization 1: Work with the Global Scale 6
- MCP conceptualization 2: Work with the Reading and Listening scales 6
- MCP conceptualization 3: MCP's performance in an existing examination 6
- MCP conceptualization 4: Thinking about ETS's MCP definitions 8

The data 9

Standard setting: Method 9

- DPT item voting 10
- Voting mechanism 13
- Anchor Items 14

Standard setting: Results 15

Annexes **16**
A Schedule of workshop 16
B The judges 17
C Judges response to the Standard Setting 18
D The test - list of items and sets 21
E Gauge item types 22
F Track item types 22
G The pretesting candidate samples 23
H Selected items and judges comments 24
B2/C1 Item – Listening (Track - Anchor Item) 24
C1 Item – Reading (Track) 25
A2 Item – Text Organization 26
B1 Item – Text Organization (Gauge) 27
A1/A2 Listening Comprehension (Track) 28
B1 Item Reading Comprehension (Track) 29
I Language Elements (Vocabulary and Grammar) items - Difficulty parameters 30
J MCP conceptualization: Judges' views on the ETS MCP receptive skills descriptors 31
K Listening/Reading MCP characteristics, from CEFR and Tannenbaum/Wylie 2008, judges' choice 35
L Standard setting, judges deviations and mean 36
M Judges' comments on tasks 37
N References 40

The test

The Dynamic Placement Test (DPT) is a non-compulsory online placement test targeted primarily, but not exclusively, at students at the end of secondary education or at the beginning of university. It is designed to be taken by large groups of incoming students to properly place them in language learning classes, although it may also be taken by individual students on an *ad hoc* basis. A result based on the receptive skills is reported in Common European Framework of Reference (CEFR) level at the end of the test. Possible outcomes are the CEFR levels from A1-C2. Raw score points will be communicated as a further indicator of success within each band.

The Dynamic Placement Test has two parts, the Gauge which looks at linguistic range and accuracy, and the Track which offers more task-oriented test items, a mix of Reading, Listening and language tasks.

Part 1 - Gauge

Structure: The gauge has 3 item types, with a selection of these items at each level. The gauge has items from A1 level to C2 which are presented adaptively.

Objective: The focus is on language in functional terms. The gauge tests vocabulary and grammar in the context of specific goals or skills (describing future wishes, talking about past habits etc), rather than testing knowledge of individual words or sentences based on tenses.

Rationale for item types: These task types have a higher number of possible answers which reduces the chances of guessing the correct one. Test takers cannot look up the right answer in an online dictionary or grammar reference. In addition, they have to interact with the item and not simply tick a box a, b or c.

Part 2 - Track

Structure: When test takers have completed the Gauge, the language skills section, they are placed within one of three level bands for part 2, namely track A, B or C. This part then focuses on real-life tasks, and items include reading emails and articles, and listening to conversations and speeches. The design goal here is to decide where within the level band or range the candidate is. In some cases, for those at the lower or upper end of the band, bonus questions help decide if the candidate can break out of the band, downwards or upwards.

The Track is made up of a mix of Listening, Reading, Vocabulary and Grammar items. Each Track represents a CEFR bandwidth and contains items from each of the CEFR levels within that bandwidth. Thus, Track B has the following structure:

- 3 Listening Items at B1
- 3 Reading Items at B1
- 5 Vocab/Grammar items at B1
- 3 Listening Items at B2
- 3 Reading Items at B2
- 5 Vocab/Grammar items at B2

It is the test taker's success in the track which determines their final CEFR level. Borderline test takers are given 3 additional items to make a final determination.

Tasks and the CEFR

The tasks in part 2, the Reading, Listening and Language Elements sections, are based on a number of CEFR descriptors, such as

- Reading for information and argument
- Processing text
- Transactions to obtain goods and services

- Listening to announcements and instructions
- Correspondence
- Sociolinguistic appropriateness

Additionally, items in the Language Elements — along with all part 1 items — were cross-referenced with writing from telc's own vast test selection of samples from levels A1 to C2 and also with the *Cambridge Grammar Profile*.

Annex D contains a list of items, test-sets and tasks.

The procedure for the DPT Standard Setting

Thursday: introduction and orientation

The event started with an introduction to the CEFR, the Council of Europe's principle of plurilingualism and integration, and a short review of different types of language tests. We discussed the differences between proficiency, diagnostic, achievement, and placement tests and focused especially on the purpose of placement tests. We then looked at problematic examples of item types in use in other placement tests, to highlight the inherent difficulties in designing valid and reliable test items.

Friday: CEFR familiarisation and standard setting

We looked at the CEFR philosophy of transparency and coherence, through the use of Can-do descriptors, the positioning of the learner as a social agent as reflected in the focus on reception, production, interaction and mediation, the celebration of plurilingualism and the belief that language learning is a discovery of other cultures and one's own.

After discussing the uses of the CEFR for Learning, Teaching and Assessment, we moved on to the CEFR levels and used common telc activities such as card games and mind maps to consolidate our understanding of what learners at each level can do. The purpose of this section was to establish a common understanding of the goals and levels of the CEFR among participants.

To do this, familiarization activities were carried out. We began by reviewing the CEFR scales and determining CEFR levels using video material. Although the DPT does not have an oral module, this is a useful exercise to familiarize the judges with the CEFR and the corresponding expectations.

The next segment was used for defining the Minimally Competent Person (MCP) for levels A2 - B2. For this activity, the descriptors for A2, B1, and B2 are arranged side-by-side for each scale, creating a chart for use. We determine the MCP using the CEFR, *Profile English* and Council of Europe (COE) material. Additional items come from telc – language tests and the material from the DPT.

Tasks of other providers are also used to be able to approach the location with fresh eyes — unclouded by the knowledge of which level the tasks were designed for, which can happen with an experienced practitioner. The tasks and their sources are given after the event. Activities included:

- Puzzle - Listening to Announcements and Instructions
- Puzzle - Reading Instructions

Then the rounds of standard setting began with everyone working through packet 1 in detail with a round 1 voting, discussion and results, then round 2 voting with discussion and results.

As the test is delivered digitally, the judges were provided with ten laptops on which to view it. As there were 22 judges (excluding leaders and moderators), the laptops had to be shared.

Saturday: Standard setting

A session for reviewing receptive skills at various CEFR levels was set up. Work on the different skills was organized in parallel working groups. The findings for the working groups on is reproduced in annexes J and K.

Then the remaining packets were unleashed with round 1 voting, discussion and round 2 voting for each. Some packets were only focussed on by some groups of judges.

The judges

A group of 22 judges took part in the Standard Setting. All of them were actively involved in either English language teaching or testing, at school or in adult education. They provided the following personal information (collected by means of a judge information sheet, multiple answers possible):

I am a teacher of English at a school.	16
I am a teacher of English in adult education.	14
I am an oral examiner for telc (or another examination board — please specify).	3
I am a rater for telc writing tasks (or another examination board — please specify).	2
I am involved in curriculum development.	14
I am an author of learning materials.	12

Judges were also asked to assess their familiarity with the CEFR on a scale coded 1 to 4, in the following four areas:

	Mean
I am familiar with the CEFR.	3.25
I have read the CEFR.	3.05

I have worked with CEFR descriptors.	2.95
The CEFR is part of my everyday work.	1.90

A complete list of the judges and some of their feedback can be seen in annexes B and C.

Conceptualisation of the Minimally Competent Person (MCP)

As Buckendahl (2005:219) put it, 'The challenge for all standard-setting methodologies is to effectively translate a participant's mental model of the target examinee (e.g. barely proficient student) into judgments that communicate the participant's recommendation of a value that characterizes the point of separation between one or more categories.' This is not always easy to do, as the participants in a standard setting, i.e., the judges may have different interpretations of the standard itself (in this case, the CEFR levels), and of the concept of 'mastering' a standard.

The conceptualization of the MCP is thus an essential part of Standard Setting and requires careful planning. In this workshop, it was done in four stages.

MCP conceptualization 1: Work with the Global Scale

As a warm-up activity, judges were asked to sort the six descriptors of the Global Scale of the CEFR into ascending order. None of the judges had any difficulty with this. The activity was nevertheless appreciated, as it prepared the way for the coming tasks.

MCP conceptualization 2: Work with the Reading and Listening scales

Judges were then given the Pre-A1 up to and including the B2+ Overall Reading and Listening scales from the 2017 Companion volume to the CEFR, with some of the descriptors (the 'puzzle bits') deleted, and a numbered list of puzzle bits. Their task was to write the number of a puzzle bit into each of the gaps, working in groups of two. This task proved to be somewhat more difficult and engendered some valuable discussion on the exact nature of each level.

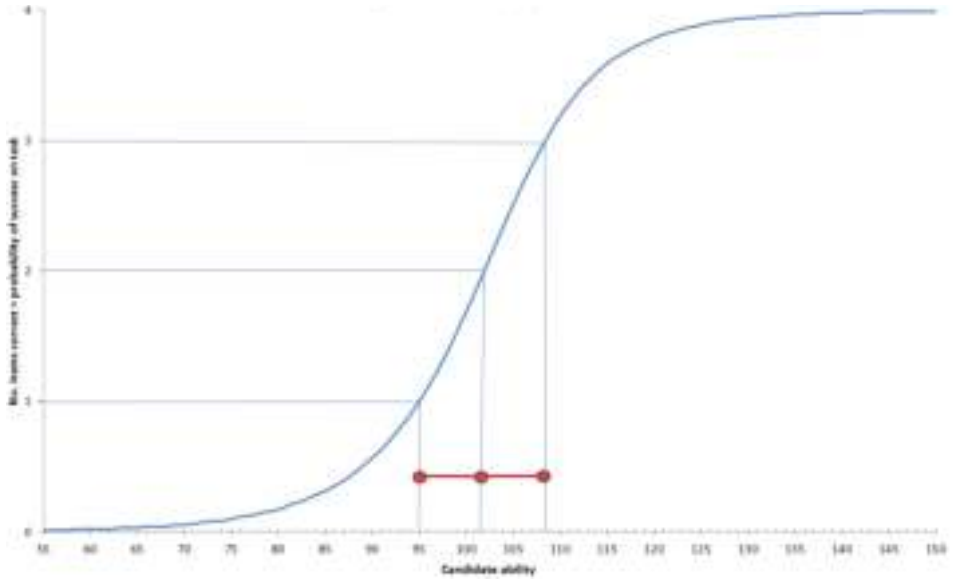
MCP conceptualization 3: MCP's performance in an existing examination

As the levels had now been internalized as theoretical concepts, judges were asked in a next step to look at actual candidates' work to see what certain candidates were able to do or not to do in practice. For this task, data from an existing paper-and-pencil examination targeting the same levels and the same population (the *telc English* examination) were used.

A task-centered method was chosen that was modelled on the standard setting conducted in the SurveyLang project (Jones et al. 2012), which is again indebted to the Van der Schoot method (Van der Schoot 2009). This is due to the fact that, as in the

SurveyLang case, the test contains cluster items where item dependencies can be expected, and that therefore a partial credit Item Response Theory (IRT) model was used.

The task was thus seen as the basic unit, and the number of items that were solved correctly by a test taker as that test taker's score on the task. For each of the tasks, the task response function was constructed. The diagram below shows the task response function for one of the tasks, with the candidates' ability at each of the possible scores for this task. Candidate ability was scaled to avoid negative values and decimals.



– figure 1 –

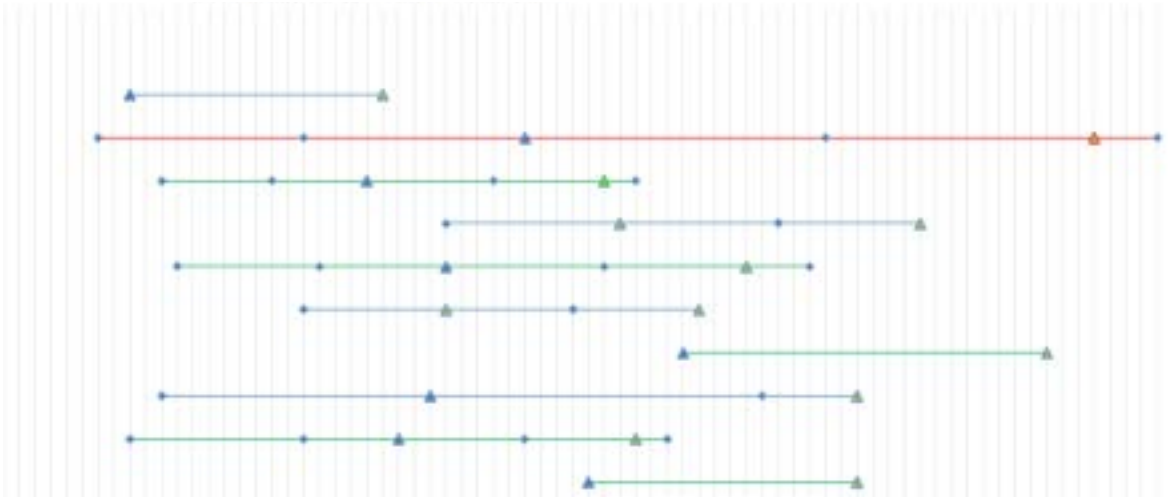
The concept of task response curves was explained to the judges, and it was pointed out that each task response curve can be summarized by a line like the red (horizontal) one in figure 1, which shows candidate abilities of a candidate who can solve 1, 2 or 3 items. For the candidate who is able to solve four, i.e. all items on this task, no ability value can be shown as this candidate may be of an ability that is just sufficient to solve the whole task, or any ability above that. Similarly, the ability of a candidate who can answer none of the items correctly cannot be shown.

Before starting on the actual standard setting, a phase of familiarization with the concepts of task response curves and abilities was run. This was done by way of showing a (slightly simplified) task diagram like the one that was to be used later and asking the judges a number of questions to make sure that they were confident in applying the concepts.

With this data we can further develop the following task response curves, as seen in figure 2. The dots demonstrate candidate abilities at the scores, with indicators showing which ability was needed to have a 50% chance and an 80% chance of solving the task at the given CEFR level. These indicators reflected the actual scores in some cases from the sample date, in other cases they were derived from hypothetical fractional score values derived by calculating 50% and 80% of the maximum possible score, and the

corresponding ability value. This can then be used to facilitate comparison between the tasks and the scores given by the judges. The values of 50% and 80% are arbitrary, yet they have a certain plausibility: a person who has a 50% chance of getting an item of a task right, can be said to have 'moderate mastery' of the task (to use Jones' (2012) terminology), a person who has an 80% chance can be said to display 'mastery'. An expected score of less than 50% was regarded as 'non-mastery'.

Figure 2 shows a section of the results. Each horizontal line is an item, the colour indicating a Listening (blue), Reading (green) or Language Elements (red) task. Vertical lines are candidate ability. Dots are actual scores, triangles are the 50% and 80% points, blue where they coincide with actual scores and green with a coloured outline where they are extrapolated. Tasks are shown in the order in which they occur on the test rather than ordered by difficulty as in Jones et al (2012), to save judges the work of jumping backwards and forwards in the test.



– figure 2 –

MCP conceptualization 4: Thinking about ETS's MCP definitions

In 2008, ETS conducted a standard setting in order to connect the TOEFL iBT, the TOEIC and the TOEIC Bridge tests to the levels defined in the CEFR. In order to familiarize the panelists with the CEFR scales, they were given the preliminary task to review selected tables from the CEFR for each language modality and to 'write down key characteristics or indicators from the tables that described an English language learner (candidate) with *just enough skills* to be performing at each CEFR level' with an explicit reminder that 'the CEFR describes the abilities of someone who is *typical* of a particular level', so that the characteristics of borderline ability have to be extrapolated (Tannenbaum/Wylie 2008, 7f). These indicators were further discussed during the ETS standard setting, and a set of 'MCP descriptors' was produced.

A selection of these descriptors was used in the final phase of MCP conceptualization. Judges were asked to consider the descriptors and indicate which ones they found useful for describing borderline ability.

While the primary aim of this task for the judges was to stimulate reflection of the minimally acceptable performance for the target levels, A2 and B1, from yet another perspective, the results may be useful for further standard setting activities, as they can be used to establish a very concise list of the most relevant descriptors. Judges' answers are given in annex K, as well as a collation of the descriptors that were found to be the most useful for characterising the minimally competent candidate.

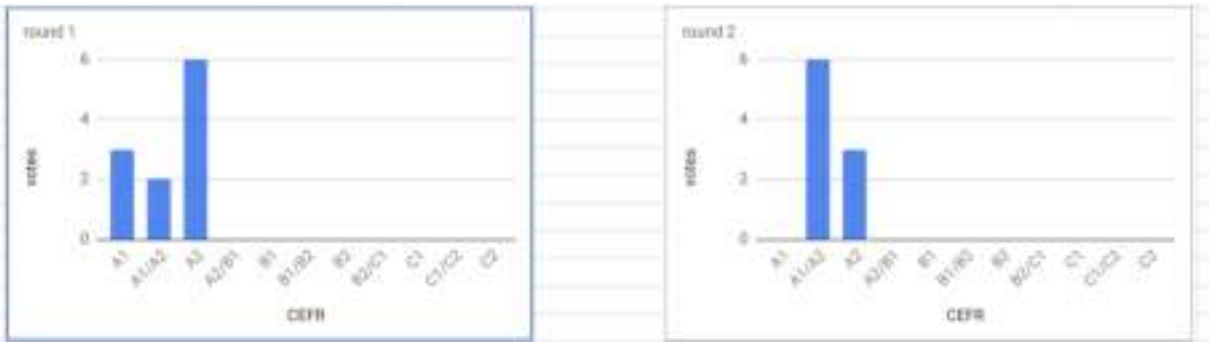
The data

For the standard setting, Clarity made data gathered from 2018/2019 available. 3,033 tests were recorded and anonymised. The data was gathered across various institutions from a variety of regions globally. The sample was reasonably representative of the expected test population. Details are described in annex G.

Standard setting: Method

For the standard setting 66 test packets were developed, with each test packet containing 3-7 items. The judges' task was to work through the exam (they were provided with laptops and headphones for this purpose), and to consider each task to determine the most appropriate CEFR level for the corresponding MCP, or how an MCP at any CEFR level could be expected to answer correctly. In order to demonstrate the outcome of the standard setting, a dedicated spreadsheet was developed to project the achieved score and agreed standard. The work of the judges alternated between working through the items in each packet and then discussing the results in the group. After discussion a second round was done by the judges. The results gave the judges an opportunity for judges to argue/defend their decisions, but generally led to an alignment, as shown in figure 3:

Item: ***ff4ce5***



– figure 3 –

Each judge was provided with a selection of items displayed in a native context on the laptops provided. The judges were presented with the exam material in the exact same manner as the test taker would see it. All item types from the DPT were presented:

From the Gauge:

- Word placement
- Sentence reconstruction
- Text organization

From the Track, we reviewed the traditional Listening and Reading skills as delineated in the CEFR. Additionally, we also considered Language Elements. All items were presented in the DPT Viewer on the laptops in their native environment. We also provided the following additional information along with the transcript: correct answer, percentage of those candidates who reached A2 in the exam and were able to answer the item correctly, percentage of those candidates who reached B1 in the exam and were able to answer the item correctly, percentage of A2 MCPs able to solve the item and percentage of B1 MCPs who were able to solve the item, plus a distractor analysis of the A2 and B1 MCPs. Candidates had been identified as MCPs by their only gaining just enough points to get the A2 or B1 result respectively.

DPT item voting

During the DPT standard setting, the judges voted on each item with the CEFR level of the MCP. ClarityEnglish developed a tool to make this straightforward and beautiful.

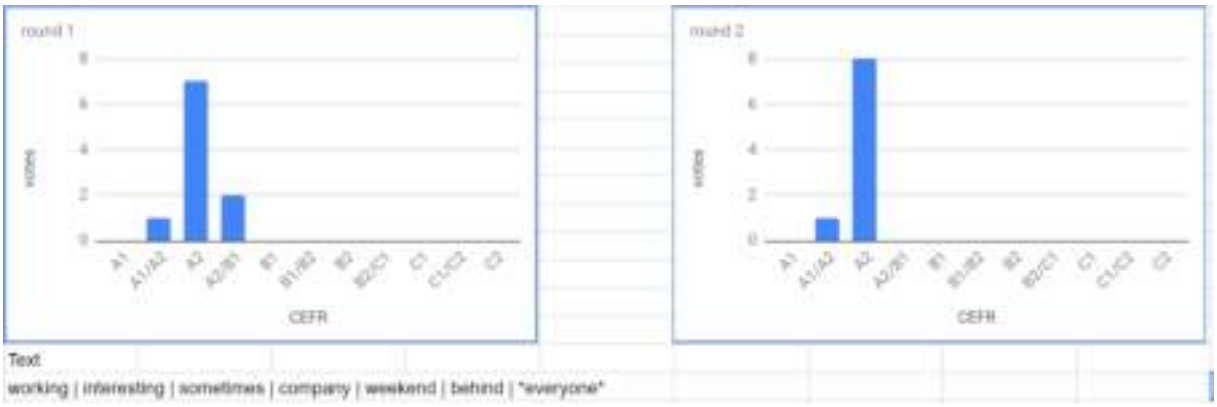
To place each item in its native setting, ClarityEnglish built extra code into the item renderer to display additional code/content in any question. This allowed us to use config files to add the voting for this project, add debug ids and editing tools if we use the renderer for item reviewing, with a slider representing the CEFR Levels A2-C1.

– figure 4 –

Each slider was linked to an event function that saves the vote to a database. We also provided information with a description of the test population and a key on how to read

the data provided for each item. In the 'Items for Viewer', a spreadsheet contains one sheet for each 'packet' of items. The first 6 packets are a handmade mix of Gauge items and a special packet1.html was created. The other 60 packets are the actual exercises from the Track, mixed around. Each sheet listed the item_id and the item_text (taken from the item analysis spreadsheet). Summary functions were created to read through this control sheet and use it during item discussion between and after voting rounds..

Opening the packet, we could then run getVotes and it will pull data for each item and put it into named ranges, which should then update each chart. Round 1 and Round 2 pull different data from the database. See figure 5.



– figure 5 –

'*dptss setup data collection*' is a script that reads through this control sheet. For each packet it copies 'data collection template' and makes a sheet for each item. It fills in the item_id and the item_text.

The judges were invited to think about possible reasons for the distribution of answers found, about features of item difficulty and the MCPs' capacity of dealing with them. Their comments are given below each item in annex H. Several parameters that influence item difficulty were identified. These fall roughly into two classes relating to the amount of knowledge already acquired, and the amount of information to be processed. The latter influences the capacity to understand texts by using up a part of the cognitive resources needed to access meaning.

Knowledge parameters

1. Vocabulary, including idiomatic language and chunks
2. Grammar (especially use of passive voice, tenses other than the present tense)

Processing parameters

3. Width of context needed to find the correct answer (word, phrase, sentence, paragraph, whole text)
4. Position of relevant information in a Listening text (beginning, middle, end, multiple places)
5. Amount of (additional/irrelevant) information, 'too much information'

6. Counter-intuitiveness
7. Parallel processing (logical thinking, interpretation)
8. Information is put in a different way in text and item (paraphrase)
9. Word overlap (text with correct answer / text with distractors / no word overlap)

One additional dimension was mentioned, namely

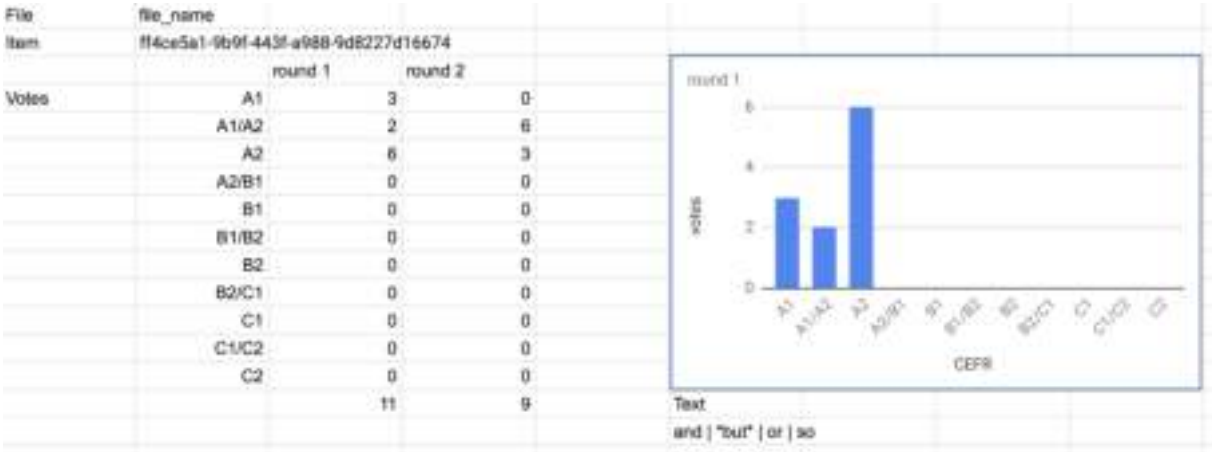
10. sound/script matching

which is required for the Listening items, and where mistakes may lead to a misunderstanding of the item (e.g. misinterpreting 'weight' as 'wait').

A possible interference of world knowledge was spotted in one item which relies on matching "Nobel-prize winning' with 'famous'.

These parameters provided a framework for assessing the difficulty of the items in the standard setting. It is however the pertinence, e.g. the frequency of the vocabulary in question and its relevance for finding the answer, rather than the mere presence of one or several of these factors that determines item difficulty. No direct relationship between any of them and item difficulty was found.

While going through the packets, judges were required to enter their score using the slider, (*figure 4*), which was then exported into the spreadsheet, which then correlated the input and made it available for comparison, figure 6.



– figure 6 –

The judges were given the task of placing the item at the CEFR level they feel best reflects the candidate's ability. In effect, the judges provided two assessments, one prior to group discussion and one after. Each of these assessments is a standard of its own, as each is an estimation of the target MCP's ability. If the test worked perfectly, and if each judge were perfectly consistent in his or her interpretation of the target level, and if all judges were of equal strictness, the ability found would be the same for each task. Perfection is however not to be expected in any human activity, so that ability values can be expected to vary across tasks as well as across judges. A compromise therefore has to

be found. This was done by calculating the mean MCP ability first per judge across all tasks, then across judges. The abilities were re-translated into raw score points, as the result of the exam is to be reported in raw score points.

The result was shown to the judges and discussed. It was also shown what impact the cut score found would have on the pretesting group's grades. After the first round, judges found that their standards were probably too strict. They were then given the opportunity to go through the exam again and to reassess and modify their first judgements in the light of the discussion, and hand in their reassessments in the same way as in round one. This led to a lowering of the cut scores by one point for A2 as well as for B1.

Judges were given the opportunity to write any comments they might have on any of the items, into their item booklet. Some of these comments are reproduced in annex H.

Voting mechanism

We took the items from the DPT and placed them into packets, with each packet containing between 3 and 7 items, depending on difficulty and the time required to complete each one. In total the judges reviewed a total of 129 items, representing a cross-section of the DPT database. Due to the large database of items in the DPT, it was not within the sessions' scope to review each individual item, but rather to review a cross section of the items.

Items for the DPT are drawn randomly from a database of available items for each section. Items are categorized in the following manner:

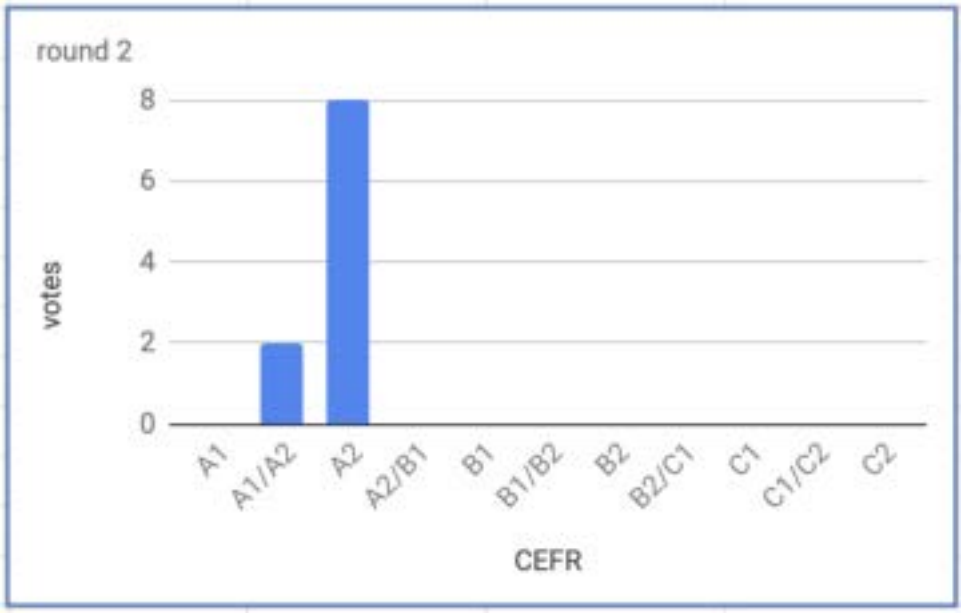
- CEFR level (A2-C2)
- Item Type (Sentence Reconstruction, Text Organisation, Word Placement)
- Skill (Reading, Listening, Language Elements)
- Gauge or Track

Items are chosen according to the scheme as defined above in DPT. Once the items from the standard setting have been vetted, we can use the data gathered to gauge the performance of similar items.

For example:

Item: 83e249, a Word Placement item taken from the Gauge.

The qualitative analysis from the panel determined that this is a mid/low A2 item as shown in figure 7.



– figure 7 –

Firstly, we can compare the qualitative analysis of this item with the quantitative analysis of the same item. From the sample of 3,033 test takers, we see that the analysis supports this outcome, i.e. the results from the minimally competent test taker reflect the outcome of the judges' assessment. In other words, looking at the data from the DPT, a test taker who can correctly solve this item is likely to be awarded a CEFR level of A2 or above. The correlation is stronger as the awarded CEFR level rises.

Furthermore, we can see that items similar in properties to this item (Gauge, Word Placement, Language Elements, A2) perform in a similar way. Thus, each item which has been vetted by both a quantitative and qualitative analysis can serve as an anchor item for the remainder of the items available in the DPT database.

Anchor Items

Anchor items are used throughout the test in order to secure quality and appropriate level award. With the DPT, anchor items are a common set of items administered in combination with two or more alternative forms of the test with the aim of establishing the equivalence of the test scores. Ideally, a test item should always deliver the same results. For "validity", or for the test to be valid, we need to be sure that the DPT measures what it is supposed to measure, or to answer the question: "is a B2 from DPT really a B2?". The purpose of the anchor item is to provide a baseline for an equating analysis between different variations of the item, manifesting itself in different "versions" of the test.

When all the items of different randomly generated "versions" produce similar scores, then the test can be said to be reliable. The test validity can itself be tested/validated using the test's reliability, repeatability (test-retest reliability), and other traits, usually via multiple runs of the test whose results are compared. Statistical analysis helps determine

whether the differences between the various results either are large enough to be a problem or are acceptably small.

Anchor items are not intended to test the individual's ability to take tests, interpret questions, or understand concepts unrelated to the test questions. Instead, our goal is to eliminate the incongruence between what the DPT is designed to assess and what it actually assesses. In this way we assessed items (requiring the same knowledge and linguistic skills) in multiple ways, both qualitatively and quantitatively. Like all language exams based on the CEFR the DPT (and all the items) are intended to find out what an individual is able to do rather than what an individual is unable to do — i.e. "Can-do" statements.

Thus, the validation process for the DPT is very robust. Ideally, this will also reflect in the reliability. Reliability in testing means consistency: a test with reliable scores produces the same or similar results on repeated use. This means that a test would always rank-order a group of test takers in nearly the same way. This is particularly important for the DPT, as the algorithmic test constructor automatically generates different "versions" of the same test, as described above. If each item is determined to have the same value (as compared to the anchor item) then the overall test can be said to be reliable. A single test taker could take the DPT several times and they must always receive the same results, relative to their CEFR abilities.

The validity of a cross section of items was proven through qualitative analysis at the standard setting event. The quantitative analysis reinforced this assessment, so that we have two very strong indicators for the items reviewed. These items can be used as anchor items then to prove the validity of similar items. This scheme validates the other items in the DPT database. The DPT is dynamically generated from valid items in the DPT database, producing individual, valid and reliable tests.

Standard setting: Results

As the annex L demonstrates, the judges, through independent work, were able to correctly identify each item according to the relevant CEFR levels. The scale of the event, with 22 judges, allowed for small variances or discrepancies among individual judges. The large group of practitioners was found to be reliable to form a consensus on each item's validity. *Note that for reasons of test security, actual results are not included in this report. Interested parties should contact the author.*

Annexes

A Schedule of workshop

Day 1

Round of introductions

Agenda

Presentation of Pretesting Sample

MCP Conceptualisation 1: Work with the Global Scale

MCP Conceptualisation 2: Work with the Overall Reading and Listening scales

MCP Conceptualisation 3: MCP's performance in existing examination

MCP Conceptualisation 4: Thinking about ETS's MCP definitions

Standard setting: Explanation of method

Standard setting: Round 1 packet 1

Standard setting: Presentation of Round 1 results, discussion

Standard setting: Round 2 packet 1

Standard setting: Presentation of Round 2 results, discussion

Day 2

Receptive Skills CEFR workshop

Standard setting: Round 1 and 2 for packets 2-24

Standard setting: Presentation of results, discussion

Judge feedback

Conclusion

Feast

B The judges

The seminar leaders and moderators of the standard setting were:

1.	Laura Edwards	telc – Language Tests
2.	Charlotte Kwok	ClarityEnglish
3.	Sieon Lau	ClarityEnglish
4.	Sean McDonald	telc – Language Tests
5.	Adrian Raper	ClarityEnglish
6.	Andrew Stokes	ClarityEnglish

The following list of practitioners were active and served as judges in the Standard Setting:

7.	Michelle Raquel	The University of Hong Kong, Hong Kong
8.	Chi Lai Tsang	St Joseph's College, Hong Kong
9.	Kima Huang	The Winhoe Company, Taiwan
10.	Mei-Hua Chen	Wenzao Ursuline University of Language, Taiwan
11.	Tun-Whei Chuo	Wenzao Ursuline University of Language, Taiwan
12.	Ling-Ying Chou	Wenzao Ursuline University of Language, Taiwan
13.	Ching-Hsien Hung	MCU English Language Center, Taiwan
14.	Mia Aghajari	telc Language Tests, Germany
15.	Zhao Ming Gao	National Taiwan University, Taiwan
16.	Santi Budi Lestari	University of Lancaster, United Kingdom
17.	Elinor Stokes	AtlasEnglish, United Kingdom
18.	Thomas Jones	Brock Solutions Agency, United Kingdom
19.	Matthew Patrick Wallace	University of Macau, Macau
20.	Christina Au	EduWise, Macau
21.	Brenda Pui Lam Yuen	National University of Singapore, Singapore
22.	Paul Rogers Barney	Higher Colleges of Technology, United Arab Emirates
23.	Huỳnh Thị Ái Nguyên	Vietnam USA Society English Centers, Vietnam
24.	Nguyên Thi Ngoc Quynh	Vietnam National University, Vietnam
25.	Ervida Lin	Solusi Education, Indonesia
26.	Sisilia Setiawati Halimi	Universitas Indonesia, Indonesia
27.	Gunadi Harry Sulistyo	State University of Malang, Indonesia
28.	Kun Aniroh	Universitas Merdeka Malang, Indonesia

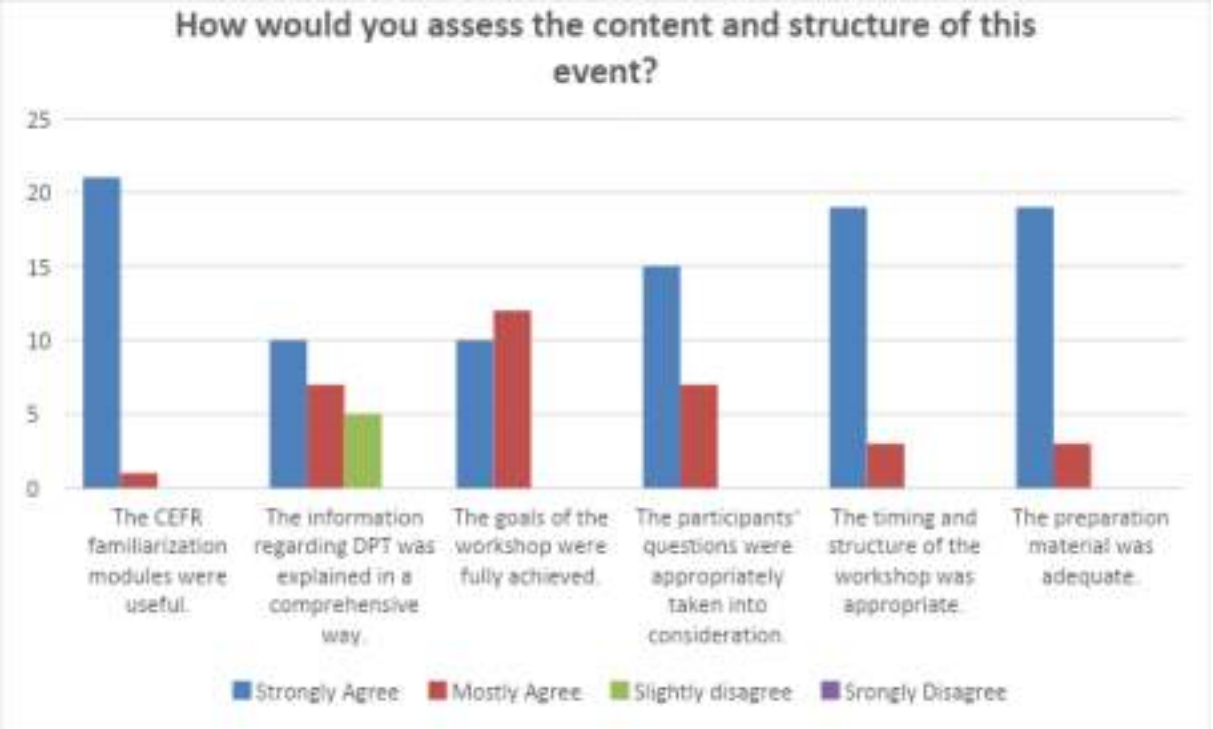
C Judges response to the Standard Setting

At the end of the workshop, we asked the judges to anonymously complete a questionnaire.

In the first part, we asked as to what degree they agree with the following statements:

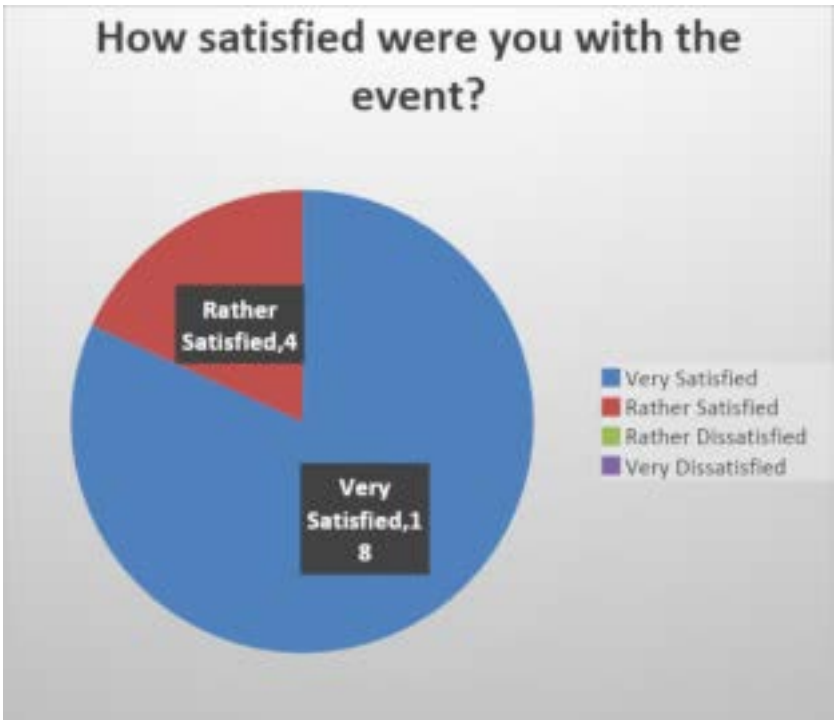
- The goals of the workshop were fully achieved.
- The information regarding DPT was explained in a comprehensive way.
- The CEFR familiarization modules were useful.
- Questions coming from judges/participants were appropriately taken into consideration.
- The timing and structure of the workshop was appropriate.
- The preparation material was adequate.

The results were overwhelmingly positive:



– figure 8 –

We also asked our academic panel of practitioners if they were satisfied with the event:



– figure 9–

Why were you satisfied with the event?

1. Got more familiar with CEFR & Clarity's new approach to testing. *-M. Chan*
2. The event is structure clearly with a particular emphasis in each section *-Gunadi H Sulistyo*
3. This event is very enlightening and fruitful. I have benefited a lot from it. Thanks for organising this wonderful workshop. *-Anonymous*
4. Time management on different tasks was just perfect. *-Christina Au*
5. The board offers a very decent introduction to the CEFR and plenty of opportunities and freedom for us to discuss the items and evaluate the test *-Chi Lai Tsang*
6. It's organised so well. The program arranged so appropriately. *-Anonymous*
7. I learnt a lot of new things related to the CEFR and I have understood the CEFR and the test well. *-Anonymous*
8. I learned a lot about CEFR, objectives of the event are met *-E. Lin*
9. A chance to interact with professionals in different areas of the world in terms of testing. *-Tun-Whei Isabel Chuo*
10. Well organised with deep discussion *(Anonymous)*
11. Great organisation, well thought out, very informative *-E. Stokes*
12. All aims achieved plus a good time and good collaboration *-T. Jones*

What did you especially like?

1. The discussion on test items *-M. Chan*
2. CEFR familiarisation *-Gunadi H Sulistyo*

3. The idea of working in pairs to identify the level. This approach has encouraged fruitful discussion by justifying our own decisions. *-Anonymous*
4. CEFR familiarisation, presentation of items, early discussion of items as a group. *-Anonymous*
5. The discussion about the test items which we can share different opinions. *-A. Chou*
6. The vibrant atmosphere in the workshop pair work *-C. Au*
7. How the sessions were structured and the fact that the items (of different CEFR levels) were presented in a mixed fashion in some sessions rather than presenting them homogeneously as indicated in the schedule. *-Anonymous*
8. The grouping arrangement and the voting mechanism *-Anonymous*
11. The way the level setting was conducted *-Anonymous*
12. The diversity in panelist representation and perspectives in discussions *-Anonymous*
13. The CEFR familiarization *-Anonymous*
14. CEFR familiarisation, food, number of participants *-Anonymous*
15. The exchanges of friends in the group we worked with *-Anonymous*
16. The schedule arrangement, the discussion part, the item viewer *-E.Lin*
17. The discussion *-Tun-Whei Isabel Chuo*
18. The way the organisers and panelists shared their ideas. *-Thi Ai Nguyen Huynh*
19. The lively informal atmosphere that encouraged discussion *-E. Stokes*
20. Range of experience of attendees, dialogic nature of response examination *-M. Wallace*
21. Sean and Laura's style and attitude *-T. Jones*

D The test - list of items and sets

- redacted for publication -

E Gauge item types

The Gauge has the following item types for levels A1-C2

- Sentence reconstruction
- Word placement
- Text organization

F Track item types

The Track has the following item types for levels A1-C2

- Reading
- Listening
- Language Elements (Vocabulary and Grammar)

G The pretesting candidate samples

Candidates were mainly school leavers entering university, with a few younger and older students. There were slightly more male than female students. First languages included German, Chinese, Spanish, Indonesian, Arabic.

There were 3,033 candidates who completed DPT and whose anonymised results were used in item analysis prior to conducting this standard setting.

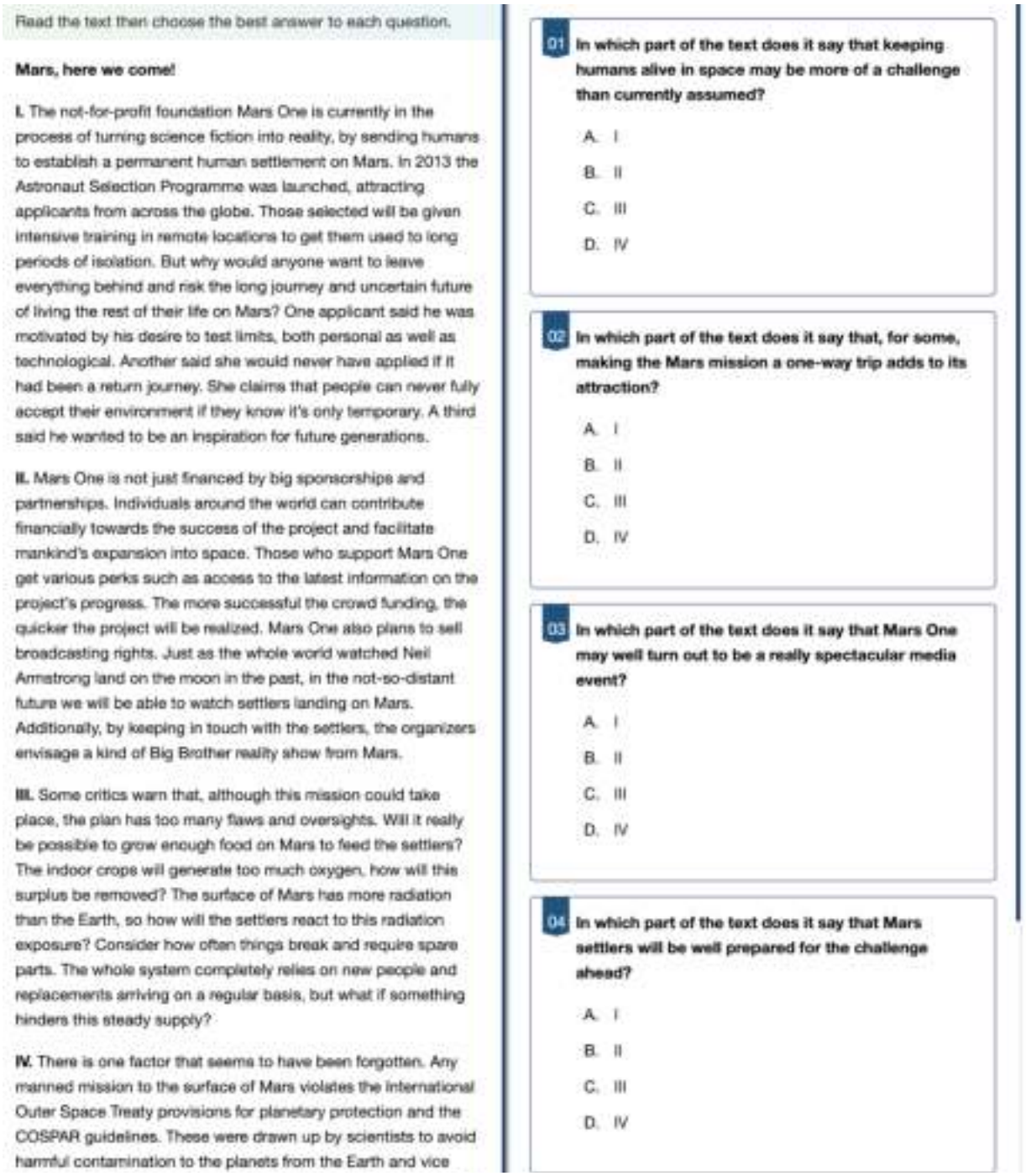
H Selected items and judges comments

B2/C1 Item – Listening (Track - Anchor Item)

Judges' comments

- B2: too much info. thought question implied another answer.
- Last option too easy to guess
- Lower levels: not expecting "genre"; hear "gender" and stop listening
- Position of correct answer deeply embedded
- Too much paraphrasing (typically used at C1, is CEFR descriptor)
- far too much information. Vocabulary --> last sentence:
- Idea not straightforward, over 300 words too long, 100 words max! --> C1-item!

C1 Item – Reading (Track)



Read the text then choose the best answer to each question.

Mars, here we come!

I. The not-for-profit foundation Mars One is currently in the process of turning science fiction into reality, by sending humans to establish a permanent human settlement on Mars. In 2013 the Astronaut Selection Programme was launched, attracting applicants from across the globe. Those selected will be given intensive training in remote locations to get them used to long periods of isolation. But why would anyone want to leave everything behind and risk the long journey and uncertain future of living the rest of their life on Mars? One applicant said he was motivated by his desire to test limits, both personal as well as technological. Another said she would never have applied if it had been a return journey. She claims that people can never fully accept their environment if they know it's only temporary. A third said he wanted to be an inspiration for future generations.

II. Mars One is not just financed by big sponsorships and partnerships. Individuals around the world can contribute financially towards the success of the project and facilitate mankind's expansion into space. Those who support Mars One get various perks such as access to the latest information on the project's progress. The more successful the crowd funding, the quicker the project will be realized. Mars One also plans to sell broadcasting rights. Just as the whole world watched Neil Armstrong land on the moon in the past, in the not-so-distant future we will be able to watch settlers landing on Mars. Additionally, by keeping in touch with the settlers, the organizers envisage a kind of Big Brother reality show from Mars.

III. Some critics warn that, although this mission could take place, the plan has too many flaws and oversights. Will it really be possible to grow enough food on Mars to feed the settlers? The indoor crops will generate too much oxygen, how will this surplus be removed? The surface of Mars has more radiation than the Earth, so how will the settlers react to this radiation exposure? Consider how often things break and require spare parts. The whole system completely relies on new people and replacements arriving on a regular basis, but what if something hinders this steady supply?

IV. There is one factor that seems to have been forgotten. Any manned mission to the surface of Mars violates the International Outer Space Treaty provisions for planetary protection and the COSPAR guidelines. These were drawn up by scientists to avoid harmful contamination to the planets from the Earth and vice

01 **In which part of the text does it say that keeping humans alive in space may be more of a challenge than currently assumed?**

A. I
B. II
C. III
D. IV

02 **In which part of the text does it say that, for some, making the Mars mission a one-way trip adds to its attraction?**

A. I
B. II
C. III
D. IV

03 **In which part of the text does it say that Mars One may well turn out to be a really spectacular media event?**

A. I
B. II
C. III
D. IV

04 **In which part of the text does it say that Mars settlers will be well prepared for the challenge ahead?**

A. I
B. II
C. III
D. IV

Judges' comments

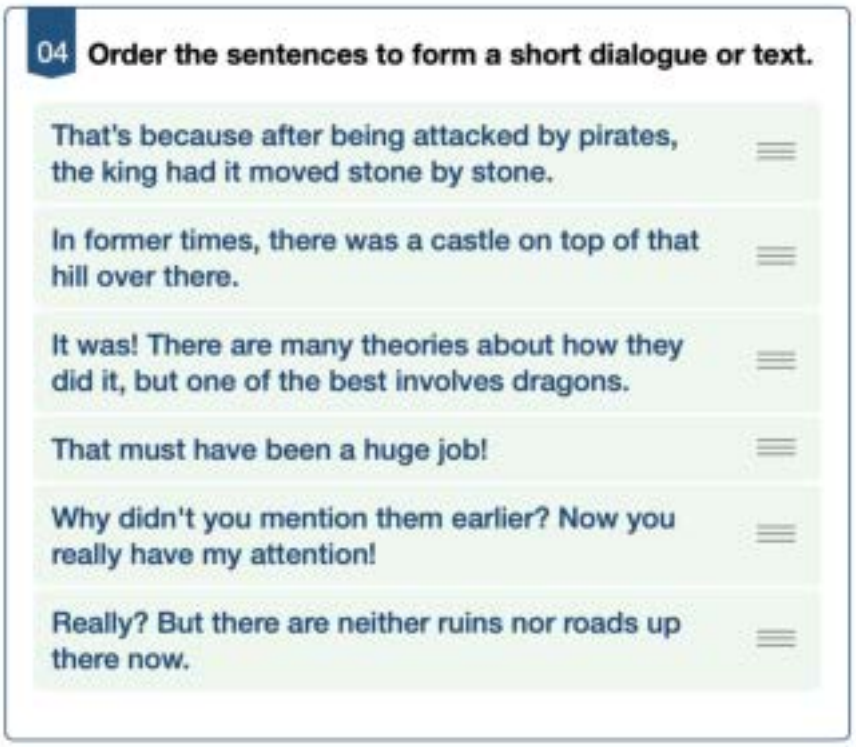
- Extremely difficult
- "Spectacular event" subjective
- Too scientifically oriented – general topic?
- Complex sentences -> C1 Item, can be supported by CEFR Descriptors

A2 Item – Text Organization

Judges' comments

- Surname vs First Name
- Other order possible -> no other order possible!
- Unrealistic
- Nationality – A2 word?
- Too difficult for A2 MCP - > Why?
- went for this one
- Too much info – focuses on various items; vocab + variety of items; length

B1 Item – Text Organization (Gauge)



Judges' comments

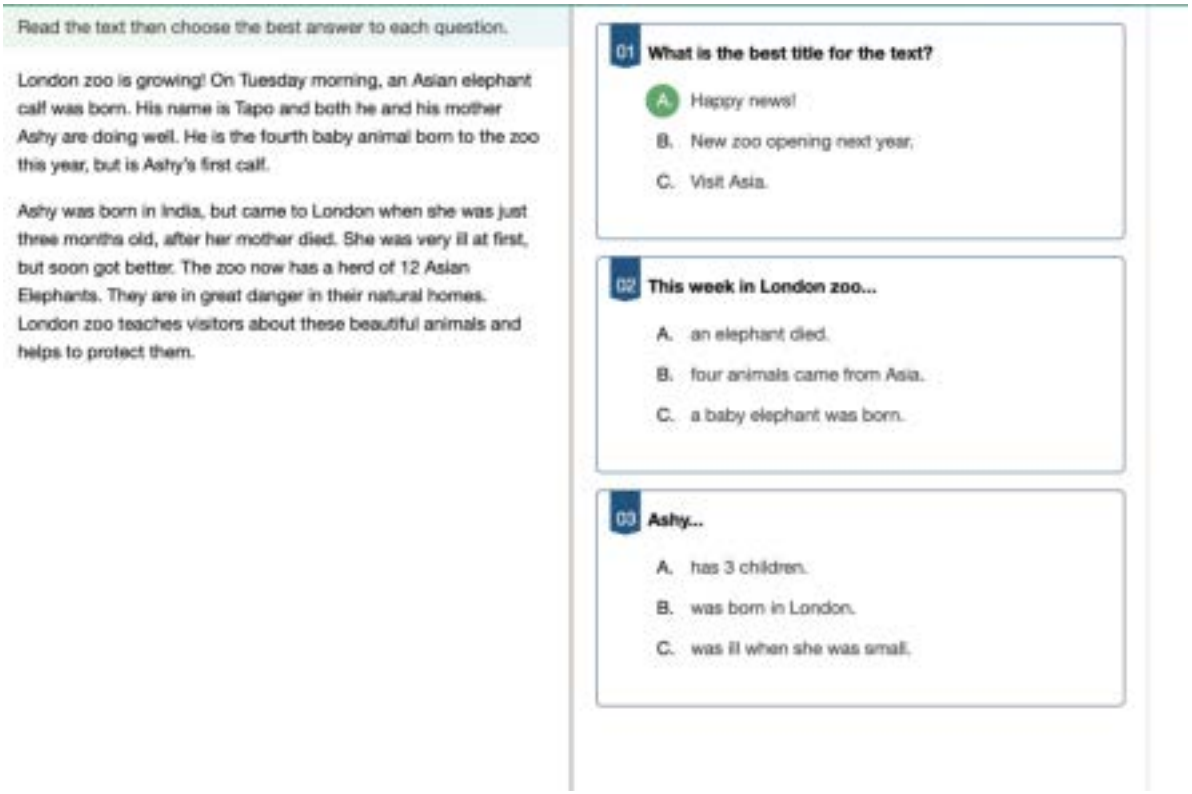
- Former times
- "Theories about how they did it" difficult to place
- Dragons, ruins: vocabulary
- B1: vocab too abstract.
- Difficult paraphrase – struggle to make connection
- 'It was very difficult for me …': Have to understand everything to be able to connect it to social skills.
- You have to understand most, if not all of the text, to be able to solve it

A1/A2 Listening Comprehension (Track)

Judges' comments

- Option: good distractor at the beginning
- Nice pictures
- Difficult for A2 to decide between Taxi and Bus in Item 2 and 3

B1 Item Reading Comprehension (Track)



Read the text then choose the best answer to each question.

London zoo is growing! On Tuesday morning, an Asian elephant calf was born. His name is Tapo and both he and his mother Ashy are doing well. He is the fourth baby animal born to the zoo this year, but is Ashy's first calf.

Ashy was born in India, but came to London when she was just three months old, after her mother died. She was very ill at first, but soon got better. The zoo now has a herd of 12 Asian Elephants. They are in great danger in their natural homes. London zoo teaches visitors about these beautiful animals and helps to protect them.

01 **What is the best title for the text?**

A. Happy news!
B. New zoo opening next year.
C. Visit Asia.

02 **This week in London zoo...**

A. an elephant died.
B. four animals came from Asia.
C. a baby elephant was born.

03 **Ashy...**

A. has 3 children.
B. was born in London.
C. was ill when she was small.

Judges' comments

- Word spot – Zoo (Wrong answer)
- Why is it happy news?
- Q1 Option "C" not Useful
- Test level appropriate
- "calf" difficult for B1
- A2: (class) rural// ill = unknown (irrelevant)

I Language Elements (Vocabulary and Grammar) items - Difficulty parameters

Code	*Difficulty parameter*
1	Vocabulary, including idiomatic language use and chunks
2	Grammar (especially: use of passive voice, tenses other than present)
3	Width of context needed to find the correct answer (word, phrase, sentence, paragraph, whole text)
4	Position of relevant information in a Listening text (beginning, middle, end, multiple places)
5	Amount of (additional/irrelevant) information
6	Counter-intuitiveness
7	Parallel processing (logical thinking, interpretation)
8	Information is put in a different way in text and item (paraphrase)
9	Word overlap (text with correct answer / with distractors / no overlap)
10	sound/script matching

A2-ordered

Item no.	*A2 MCP*	*B1 MCP*	*Codes*									
	percent correct answers		1	2	3	4	5	6	7	8	9	10
32	3,57%	63,20%	x					x		x		
30	7,14%	73,10%	x								dist/answer	
26	14,29%	79,44%	x						x		dist/answer	
24	17,86%	58,38%	x								dist/answer	
4	21,43%	79,95%	x			end	x	x		x		
17	28,57%	74,62%	x				x		x	x		x
18	28,57%	70,81%	x	tense	whole text		x			x	dist	
44	28,57%	81,98%	x		sentence							
15	50,00%	91,37%		passive		multiple	x	x	x		dist/answ	x

B1-ordered

Item no.	*A2 MCP*	*B1 MCP*	*Codes*									
		percent correct answe	1	2	3	4	5	6	7	8	9	10
24	17,86%	58,38%	x								dist/answer	
32	3,57%	63,20%	x					x		x		
18	28,57%	70,81%	x	tense	whole text		x			x	dist	
30	7,14%	73,10%	x								dist/answer	
17	28,57%	74,62%	x				x		x	x		x
26	14,29%	79,44%	x						x		dist/answer	
4	21,43%	79,95%	x			end	x	x		x		
44	28,57%	81,98%	x		sentence							
15	50,00%	91,37%		passive		multiple	x	x	x		dist/answ	x

– figure 10 –

J MCP conceptualization: Judges' views on the ETS MCP receptive skills descriptors

In the three columns to the right, the number of ticks for 'useful', the number of indications for 'not useful', and other comments are counted.

9 answers (three judges looked at the B1 descriptors only)

Listening skills of just-qualified A2 (=A2 MCP)

	Useful	*Not useful*	*Other remarks*
Can understand short, clearly, slowly, and directly articulated concrete speech on simple, everyday, familiar topics/matter.	6		
Can understand formulaic language (basic language and expressions).	5	1	sometimes
Can understand short directions, instructions, descriptions.	5		
Can extract relevant, important information from recorded messages.	6		?
As long as speech production is short, simple, slow, and clear: Can understand simple phrases and expressions that are related to the most immediate needs.	7		
Can generally catch the main point while listening to native speakers.	1	2	too general
Can understand simple directions, instructions, and everyday conversations/exchanges related to field of interest.	7		
Can understand slow, carefully articulated speech when given time to assimilate standard language/familiar variety on concrete topics.	3		
Can derive meaning if accompanied by extra-linguistic/paralinguistic clues.	3	1	

Reading skills of just-qualified A2 (=A2 MCP)

	Useful	*Not useful*	*Other remarks*
Can find specific information in simple, everyday material (e.g., advertising, brochures, menus, notices, directions, instructions, timetables, newspapers).	7		
Can understand simple and predictable material (e.g., job-related or private written communication).	4		A2
Can understand short, simple texts containing most commonly used vocabulary.	5		A2
Grasps the main point in text with predictable information or contexts, and/or texts with high-frequency vocabulary.	4		
Can infer at the vocabulary level.	0	1	?, !
As long as it is short, simply written in common, everyday language on concrete/personal topics or related to field of interest: Can find specific, predictable information in lists, signs, notices, instructions, menus.	5		
Can read and understand short personal letters.	4		
Can extract key information; can derive probable meaning of unknown words.	1		2x hopefully
Can follow specific, predictable information in simple, everyday material (e.g., tickets, calendar).	5		
Can identify main topic; unfamiliar text (especially when accompanied by visual support, logical structure).	5		?
Derives probable meaning of unknown words.	1		sometimes debatable
Needs to reread.	5	1	

Listening skills of just-qualified B1 (=B1 MCP)

	Useful	*Not useful*	*Other remarks*
Can understand main points.	5		
Can understand clear, standard speech on familiar matters and short narratives when presented relatively slowly.	8		
Will sometimes need repetition and clarification in conversation.	7		
Can follow broadcast information carefully delivered. (Example: BBC World but not SkyNews)	3		'?'
Can deduce sentence meaning.	5		sometimes
Understands main points in standard speech on familiar, regularly encountered, straightforward topics, simple technical information.	8		
Can understand speech that is articulated relatively slowly or delivered at a relatively normal pace and with clarity.	4		
May require some repetition.	6		not clearly defined
Can guess some unknown words from context.	6		not clearly defined

Reading skills of just-qualified B1 (=B1 MCP)

	Useful	*Not useful*	*Other remarks*
Reads straightforward, factual text in field of interest.	5		
Reads personal letters.	3	1	
Reads material containing some degree of abstraction.	3		
Finds relevant information in everyday material.	10		
Can infer at sentence level.	2		? not beyond
Can read straightforward, factual texts/instructions on familiar topics/field of interest.	5		
Can find and understand information in everyday material (letters, brochures, and short official documents).	6		
Can recognize significant points, events, feelings, and wishes in personal or everyday texts that are clearly structured and signposted.	7		A2
Can deduce/extrapolate meaning of occasional unknown words in familiar context.	5		

K Listening/Reading MCP characteristics, from CEFR and Tannenbaum/Wylie 2008, judges' choice

Criterion for selection of Tannenbaum/Wylie descriptors chosen by >60% of the judges (A2: 6 or more, B1: 8 or more)

	A1 (CEFR "Overall" Scale)	A2 MCP (Selection from Tannenbaum/Wylie 2008, 46-54)	A2 (CEFR "Overall" Scale)	B1 MCP (Selection from Tannenbaum/Wylie 2008, 46-54)	B1 (CEFR "Overall" Scale)
Listening	Can follow speech which is very slow and carefully articulated, with long pauses for him/her to assimilate meaning.	As long as speech production is short, simple, slow, and clear: Can understand simple phrases and expressions that are related to the most immediate needs. Can understand simple directions, instructions, and everyday conversations/ exchanges related to field of interest. Can understand short, clearly, slowly, and directly articulated concrete speech on simple, everyday, familiar topics/matter. Can extract relevant, important information from recorded messages.	Can understand phrases and expressions related to areas of most immediate priority (e.g. very basic personal and family information, shopping, local geography, employment) provided speech is clearly and slowly articulated.	Can understand clear, standard speech on familiar matters and short narratives when presented relatively slowly. Understands main points in standard speech on familiar, regularly encountered, straightforward topics, simple technical information.	Can understand the main points of clear standard speech on familiar matters regularly encountered in work, school, leisure etc., including short narratives.
Reading	Can understand very short, simple texts a single phrase at a time, picking up familiar names, words and basic phrases and rereading as required.	Can find specific information in simple, everyday material (e.g., advertising, brochures, menus, notices, directions, instructions, timetables, newspapers).	Can understand short, simple texts containing the highest frequency vocabulary, including a proportion of shared international vocabulary items.	Finds relevant information in everyday material.	

L Standard setting, judges deviations and mean

Some A2 items

rater	3149855	3158511	3158514	3158516	3158519	3158520	3158521	3158522	3158524	3158526
task										
0e5c	3			2	3	3	1	3	3	3
3629	7		5	5	5	5	5	5	6	5
3f6d	5			3	2	4	3	4	4	3
4426	6	5	5	4	4	5	4	5	6	5
468a	5	3	3	4	5	5	3	5	4	4
6360	3	5	3	4	4	3	3	3	3	3
817f	5	7	3	3	5	3	4	4	5	4
9879	3	5	3	4	3	3	3	4	3	2
a73c	5	4	4	4	5	5	3	5	4	4
bc32	6		5	6	3	5	4	5	5	5
f8b3	3	5	3	4	4	3	3	4	3	3
mean	4.6	4.9	3.8	3.9	3.9	4.0	3.3	4.3	4.2	3.7
std.dev.	1.43	1.21	0.97	1.04	1.04	1.00	1.01	0.79	1.17	1.01

Some B1 items

rater	3149855	3158511	3158514	3158516	3158519	3158520	3158521	3158522	3158524	3158526
task										
0b21	4				3	2	4	3	5	
2b1d	5	4	5		4	7	7	6	6	6
65a6	5		5		4	4	5	5	4	4
e24b	6	5	5		5	5	5	6	5	6
mean	5.0	4.5	5.0		4.0	4.5	5.3	5.0	5.0	5.3
std.dev.	0.82	0.71	0.00		0.82	2.08	1.26	1.41	0.82	1.15

Some B2 items

rater	3149855	3158511	3158514	3158516	3158519	3158520	3158521	3158522	3158524	3158526
task										
479a	6	5	5		5	5	6	6	5	6
e8c0	6	5	5		5	5	4	7	6	5
e965	7	6	7	6	6	7	6	6	5	7
fc09	6	5	5	6	6	7	5	6	5	6
mean	6.3	5.3	5.5	6.0	5.5	6.0	5.3	6.3	5.3	6.0
std.dev.	0.50	0.50	1.00	0.00	0.58	1.15	0.96	0.50	0.50	0.82

M Judges' comments on tasks

Comments are colour-coded as follows:

Ambiguous items/functioning of item (18)
Difficulty (17)
Suggestions for wording or content (13)
General comments (9)
Technicalities / Design (8)
Usability/Rubrics (6)
Typographical errors (2)
WC Word placement
SR Sentence reconstruction
TO Text organization
LC Listening comprehension
RC Reading comprehension

Task	*Comments*
WC	Too difficult for the first item. Testing map-reading (for GPS generation a bit difficult, turn left looks like going down)
SR	"Thanks for the tips …" --> double use of 'time' is clumsy! "Anyway, get in touch if you want me to …" --> long-winded style, not typical of an email (perhaps "if you want my help …")
LC	Task ("Choose the word or phrase to complete the gaps") not visible on screen / picture not important Misleading illustration style!! unnatural "Chris: This Saturday?" --> there we need an answer affirming 'yes' this Saturday! – "You see" does not fit!
RC	Pics / icons too "irregular" (choose similar types) – Don't fit on 16:9 screen to be visible at one time --> pics smaller?

	It's confusing to use the same images for two different parts of the item
SR	Nice task. Very authentic.
WP	Going shopping is outside
SP	--> Choose the best summary "you just saw" Rubric is not at all clear! What is meant by "steps"? Strong AE accent The word 'clothes' is never used –> difficult The word 'polite' is never used ('good manners') "Time for work": not clear what this means Technical query: if you get one wrong does this mean the order is wrong for all? Not a good item because only the first summary clearly deals with 5 points! It's not testing their English so much.
TO	Nice task and design! Item 2074: Is an official certificate a qualification?? Item 1961: "Health" does not fit the series
RC	Item 2242: Tricky! Intelligence test You have to think it through too much Item 2249: Too much extra info. This ought to be a boy for a school test, not a man. I imagined a 16 yr old being worried about how much it cost, not an adult

N References

Buckendahl, C W (2005) Qualitative Inquiries of Participants' Experiences with Standard Setting, *Applied Measurement in Education* 18 (3), 219–221.

Council of Europe (2017), Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with new descriptors, provisional edition.

Jones, N. et al. (2012), First European Survey on Language Competences. Technical Report, European Commission

Tannenbaum, R. J., Wylie, E. C. (2008), Linking English-Language Test Scores Onto the Common European Framework of Reference: An Application of Standard-Setting Methodology, TOEFL iBT Research Report 06 (RR-08-34)

Van der Schoot, F. (2009) Cito variation on the bookmark method, Section I in the Reference Supplement to the Manual for Relating language examinations to the Common European Framework of Reference for Languages. Language Policy Division, Strasbourg.