

Setting standards: the role of assessment in implementing the CEFR in education reform

**EALTA WEBINAR
SERIES**

9 May 2018

*Dr Jamie Dunlea
Assessment Research Group
British Council*

Common themes of the presentation

- *Adapt don't adopt*
- *The framework is a starting point, an outline, but needs to be developed, to change, to evolve to be relevant to each context*

The CEFR can be a springboard to task and test development



Common themes of the presentation

The descriptor scales are thus reference tools. They are not intended to be used as assessment instruments, though they can be a source for the development of such instruments.(page 41)

The CEFR can be a springboard to task and test development



Common themes of the presentation

- **Setting appropriate goals**
- **Ensuring quality in assessment**
 - Clear frameworks and standards to ensure test quality
 - Some key features of a test system / program
 - Specifications
 - Data analysis and evidence of technical performance quality
 - Transparency and ongoing reporting and validation
 - Standard setting and benchmarking
- **Looking beyond the test:**
 - System: assessment supporting teaching and learning across the educational context.
 - Building expertise through networks and information exchange

The CEFR: background



- ❖ Published by the Council of Europe in 2001
- ❖ “Formal origins of the CEFR date back to 1991” (Morrow,2004)
- ❖ 40 years of research in language education in Europe (Morrow,2004; Trim, 2010)
 - *Waystage, Threshold, Vantage*
- ❖ Main scaling studies carried out in Switzerland in 1994-1995

The CEFR and assessment

“At the heart of the CEF are the Common Reference levels.”
(Morrow, 2004)

3 goals from the CEFR

CEFR LEVEL	
C2	<i>[Provide] a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc. across Europe.</i>
C1	<i>Define] levels of proficiency which allow learners' progress to be measured at each stage of learning and on a life-long basis.</i>
B2	
B1	<i>[Facilitate] the mutual recognition of qualifications gained in different learning contexts</i>
A2	
A1	

Is it useful: Goals of the CEFR

CEFR LEVEL	IELTS	Cambridge	TOEFL iBT	GEPT (Taiwan)	EIKEN (Japan)
C2	8.5	CPE			
C1	7	CAE	95	Advanced	Grade 1
B2	5.5	FCE	72	High Intermediate	Grade Pre-1
B1	4	PET	42	Intermediate	Grade 2
A2		KET			Grade Pre-2
A1					3, 4, 5

Is it useful: Goals of the CEFR

IELTS	Cambridge	TOEFL iBT	GEPT (Taiwan)	EIKEN (Japan)
8.5	CPE			
6.5	CAE	95	Advanced	Grade 1
5.5	FCE	72	High Intermediate	Grade Pre-1
4	PET	42	Intermediate	Grade 2
	KET			Grade Pre-2
				3, 4, 5


Is it useful: Goals of the CEFR

CEFR LEVEL	IELTS	Cambridge	TOEFL iBT	GEPT (Taiwan)	EIKEN (Japan)
C2	8.5	CPE			
C1	6.5	CAE	95	Advanced	Grade 1
B2	5.5	FCE	72	High Intermediate	Grade Pre-1
B1	4	PET	42	Intermediate	Grade 2
A2		KET			Grade Pre-2
A1					3, 4, 5

What do the levels mean?

CEFR LEVEL
C2
C1
B2
B1
A2
A1

Provides a principled basis for evaluating the claims of test developers (and for test developers to evaluate their own claims) from both quantitative and qualitative perspectives



Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.

Cautions, criticisms...

- ❖ Morrow (2004): notes ambiguity in terminology: “what are main points?”; “How many is most?”
- ❖ Alderson et al (2006): problems for designing tests: *Inconsistencies; Terminology problems; Lack of definition; Gaps.*
- ❖ O’Sullivan & Weir (2011): “lacks the theoretical rigor, coverage and explicitness necessary...to develop tests”
- ❖ Davidson & Fulcher (2007): “does not detail particular contexts in which it is to be used, and so lacks the necessary detail on which to build test specifications.”

(North, Martyniuk, & Panthier, 2010)

- ❖ *The CEFR is language neutral – it needs to be applied with regard to each specific language.*
- ❖ *The CEFR is context neutral – it needs to be applied and interpreted with regard to each specific educational context in accordance with the needs and priorities of that context.*
- ❖ *The CEFR attempts to be comprehensive. It cannot, of course, claim to be exhaustive. Further elaboration and developments are welcomed.*

(North, Martyniuk, & Panthier, 2010)

- ❖ *The CEFR is language neutral – it needs to be applied with regard to each specific language.*
- ❖ *The CEFR is context neutral – it needs to be applied and interpreted with regard to each specific educational context in accordance with the needs and priorities of that context.*
- ❖ *The CEFR attempts to be comprehensive. It cannot, of course, claim to be exhaustive. Further elaboration and developments are welcomed.*

The CEFR was always meant to be the beginning of the sentence, not the full stop

The CEFR and assessment

Jones and Saville (2009): Implementation often perceived as focusing on assessment

Coste (2007): *“In various settings and on various levels of discourse . . . people who talk about the Framework are actually referring only to its scales of proficiency and their descriptors.”*

Looking beyond the scales

*In a school learning context, one could imagine a separate list of 'pedagogic tasks', including **ludic aspects** of language – especially in primary schools.*
(CEFR, p. 31)

*The use of language **for playful purposes** often plays an important part in language learning and development, but is not confined to the educational domain. Examples of ludic activities include:*
(CEFR, p. 55)

Looking beyond English

CEFR, page 4: the plurilingual approach emphasises:

- as an individual person's experience of language in its cultural contexts expands, from the language of the home to that of society at large and then to the languages of other peoples ... he or she does not keep these languages and cultures in strictly separated mental compartments,
- but rather builds up a communicative competence to which all knowledge and experience of language contributes
- and in which languages interrelate and interact.

Looking beyond English

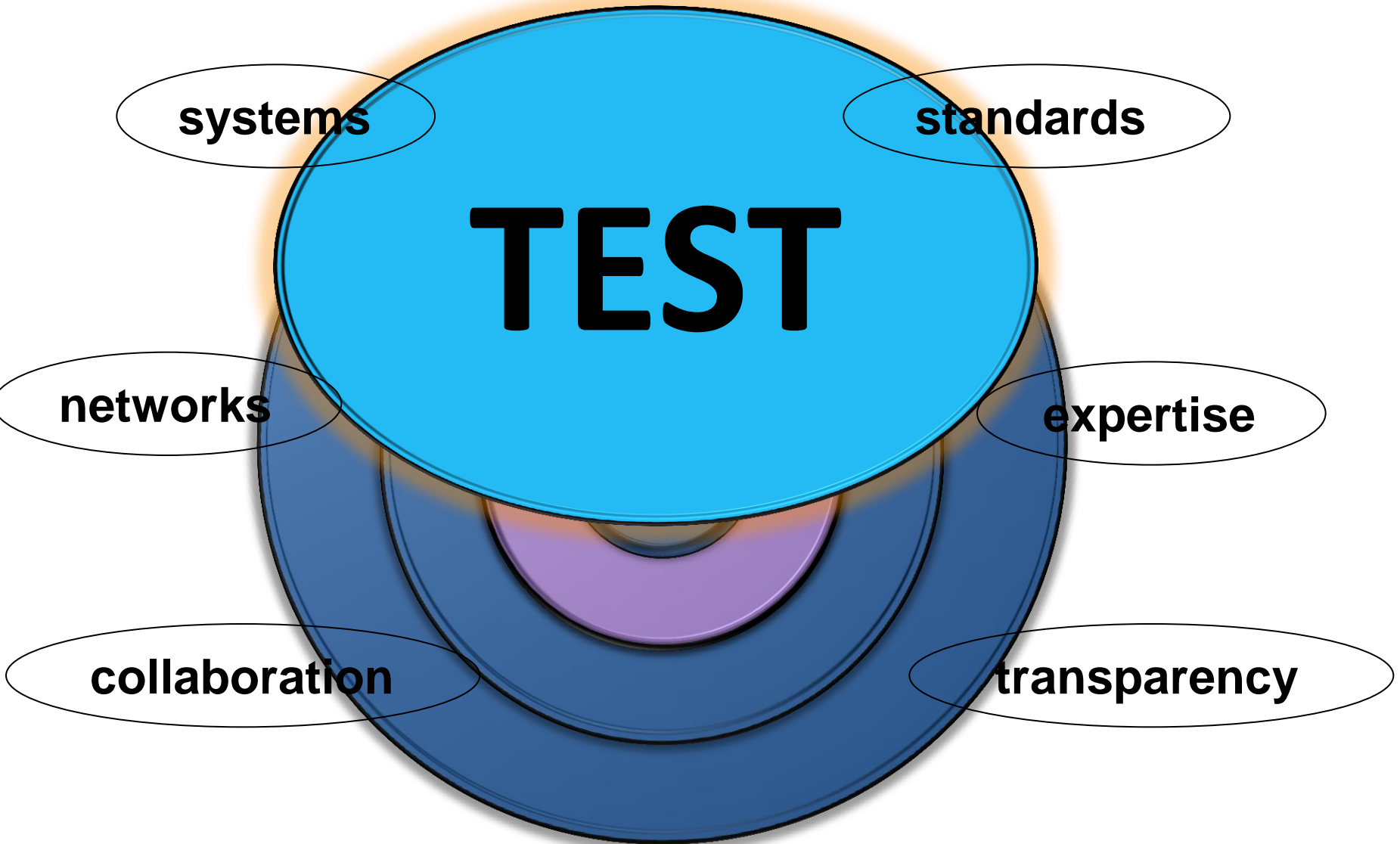
CEFR, page 4: the plurilingual approach emphasises:

- as an individual person's experience of language in its cultural contexts expands, from the language of the home to that of society at large and then to the languages of other peoples ... he or she does not keep these languages and cultures in strictly separated mental compartments,
- but rather builds up a communicative competence to which all knowledge and experience of language contributes
- and in which languages interrelate and interact.

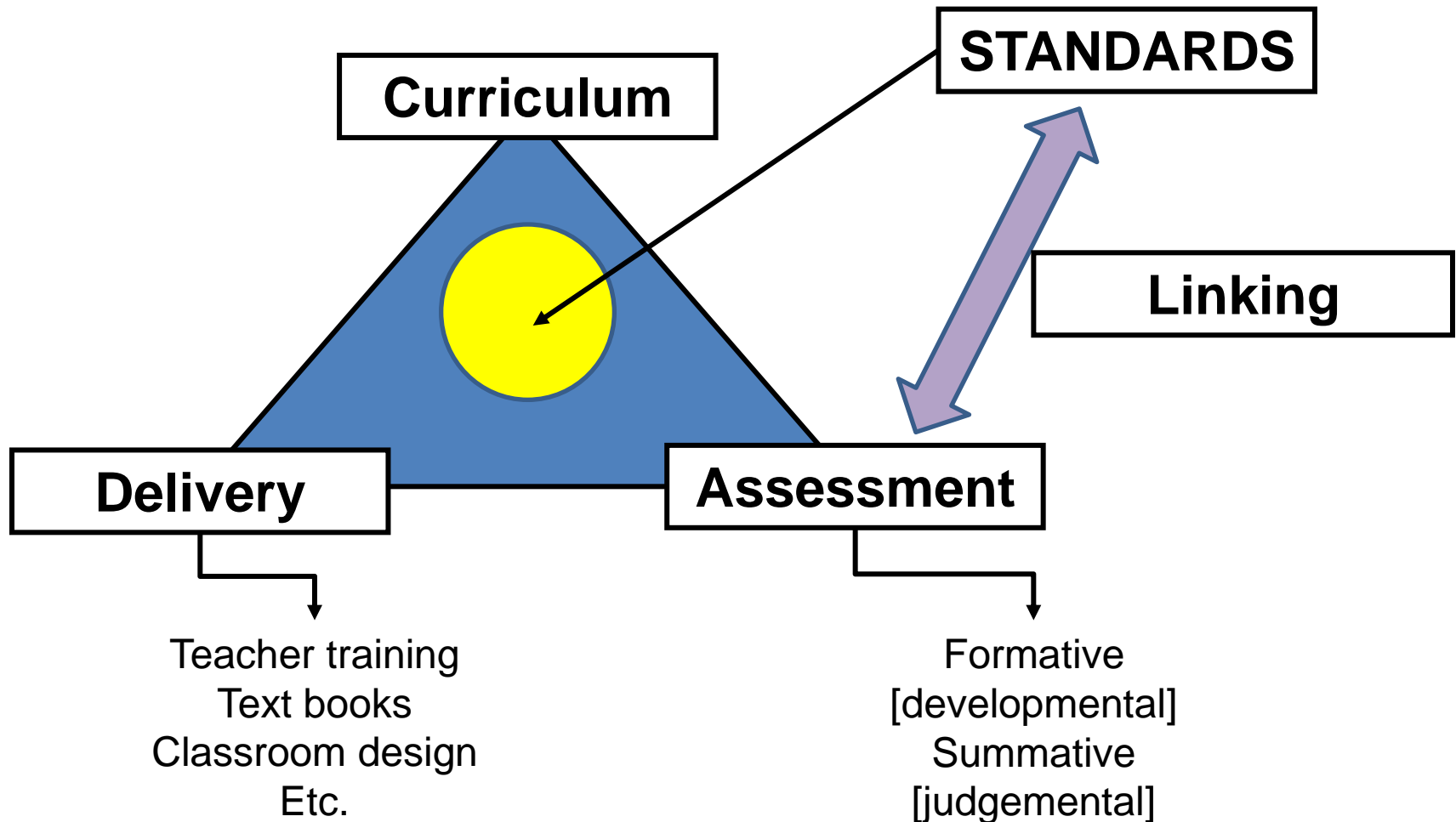
Validation and validity

- **Messick, 1986, p. 13 (also republished in Wainer & Braun (Eds), 2015)**
 - *One recommendation is to contrast the potential social consequences of the proposed testing with those of alternative procedures and even of procedures antagonistic to testing, such as not testing at all*
 - *(Ebel, 1964) .*

Assessment as part of wider systems

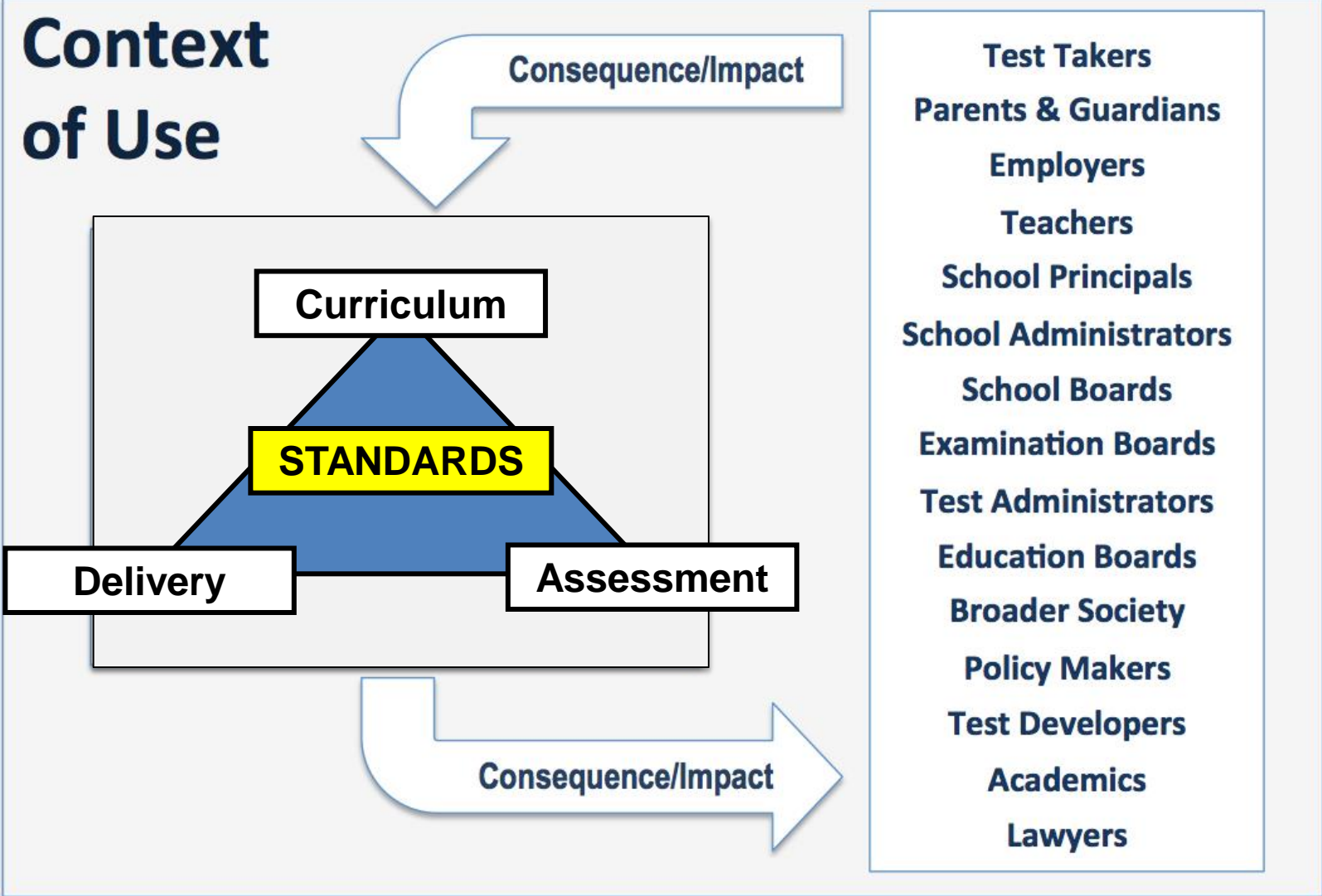


Tests as part of wider systems



(O'Sullivan, 2017)

A Socio-Cognitive CSE



(O'Sullivan, 2017a, 2017b)

C2

Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.

C1

Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices.

B2

Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.

B1

Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.

A2

Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.

A1

Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.

Look beyond the Global Scale

“The Global Scale is “just the tip of the iceberg” (Morrow, 2004)

How many *Illustrative scales* are there?

❖ Descriptors grouped in **54 scales**

- Communicative activities
- Strategies
- Communicative language competences

Look beyond the Global Scale

What aspects of the CEFR are being targeted by a particular test?

Test A test of speaking at B2

Test B test of reading at B2

Test taker A achieves a B2-level proficiency on Test A

Test taker B achieves a B2 level of proficiency on Test B

Does the “B2” mean the same thing?

Scales for speaking in the CEFR

Spoken Production

Overall spoken production

- Sustained monologue: describing experience
- Sustained monologue: putting a case (e.g. debate)
- Public announcements
- Addressing audiences

Spoken Interaction

Overall spoken Conversation

- Understanding a native speaker interlocutor
- Conversation
- Informal discussion
- Formal discussion (Meetings)
- Goal-oriented co-operation
- Obtaining goods and services
- Information exchange
- Interviewing & being interviewed

Is it useful: Goals of the CEFR

CEFR LEVEL	IELTS	Cambridge	TOEFL iBT	GEPT (Taiwan)	EIKEN (Japan)
C2	8.5	CPE			
C1	7	C1			Grade 1
B2	5.5	B2		High Intermediate	Grade Pre-1
B1	4	B1	42	Intermediate	Grade 2
A2	3.5	KET			Grade Pre-2
A1	2	A1			3, 4, 5

THE TESTS ARE NOT THE SAME EVEN IF THEY TARGET THE SAME BROAD LEVEL

The CEFR outside of its “home”

Gaurdian, 8 November, 2011

Vietnam demands English language teaching 'miracle'

As part of the strategy, which includes teaching maths in English, officials have adopted the Common European Framework of Reference (CEFR) to measure language competency. Teachers will need to achieve level B2 in English with school leavers expected to reach B1, a level below. But the initiative is worrying many teachers.....

Asian Correspondent, 25, April, 2015

Thai schools adopt European framework to boost English language proficiency

*When Thailand's new school year begins in May, teachers and schools across the country will begin the process of aligning their English language teaching with the Common European Framework of Reference for languages (CEFR). This alignment with internationally recognised language standards is a **positive step towards raising the standards** of English in Thailand, but it is going to take strategic planning and hard work to realise these goals.*

The reality check

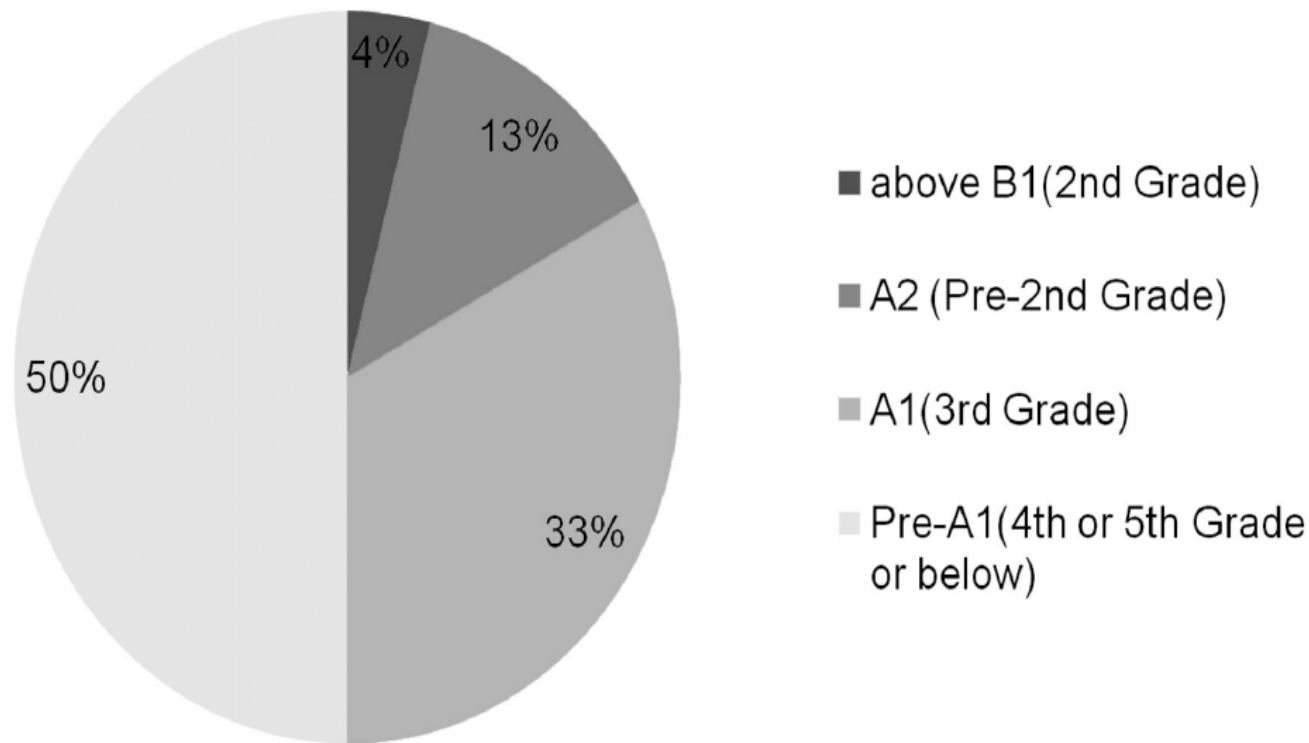
EIKEN Grade	CEFR Comparison	Example of recognition/use
1	C1	International admissions to graduate and undergraduate programs; MEXT benchmark for English instructors (Pre-1)
Pre-1	B2	
2	B1	MEXT benchmarks for high school graduates
Pre-2	A2	
3	A1	MEXT benchmark for junior high school graduates
4		
5		

The reality check

- *Eight years after the initial implementation of macro-level policies by MOET, concerns have been raised that the **reforms are not producing the expected outcomes** consistently across the country (Tran, T., Kettle, Mt, May, L, & Klenowski, V, 2016)]*
- *On November 16, 2016, Mr. Phung Xuan Nha - Minister of Education and Training - admitted that the National Foreign Language scheme for the 2008-2020 periods had been failed. However, there is no debate on why the project could not be completed **within the defined period**. (Nguyen, T., 2017)*

The reality check

Reported by Negishi (2012). Results for upper secondary students in one prefecture in Japan



http://www.tufs.ac.jp/common/fs/ilr/EU_kaken/_userdata/negishi2.pdf

The reality check

- *Eight years after the initial implementation of macro-level policies by MOET, concerns have been raised that the **reforms are not producing the expected outcomes** consistently across the country (Tran, T., Kettle, Mt, May, L, & Klenowski, V, 2016)]*
- *On November 16, 2016, Mr. Phung Xuan Nha - Minister of Education and Training - admitted that the National Foreign Language scheme for the 2008-2020 periods had been failed. However, there is no debate on why the project could not be completed **within the defined period**. (Nguyen, T., 2017)*

Issues in implementation

**Findings from a survey on implementation in several countries
(Which countries do you think these findings apply to?)**

Key conclusion 2: Major challenges in the implementation concern firstly, the lack of empirical evidence to establish links between learning outcomes and the CEFR levels and secondly, the ability of MFL teachers to use the CEFR in their lessons as intended.

Key conclusion 4: A majority of the selected countries implement the CEFR in tests or examinations; however the links between MFL learning outcomes to CEFR levels lack in general empirical evidence.

Key conclusion 7: Whether teachers know about the CEFR depends on the emphasis placed on the CEFR in curriculum and in teacher training within the country.

Directorate-General for Internal Policies (2013)
*The Implementation of the Common European Framework for
Languages in European Education Systems.* European Parliament.
<http://www.europarl.europa.eu/studies>

Issues in implementation

But a generally positive outlook...

Key conclusion 10: For learners, private providers and language assessment institutes, the CEFR provides transparency and creates possibilities to make comparisons of the courses offered. The reason for individuals to obtain a formal certificate is mostly to increase chance on the labour market.

But what about the target levels?

Key conclusion 3: There is general agreement concerning the CEFR indication of learning outcomes of MFL in upper secondary education. The stated learning outcomes across the six countries are generally similar. The level of learning outcomes related to the first MFL is usually set at level B2, for the second MFL in general the related level is B1..

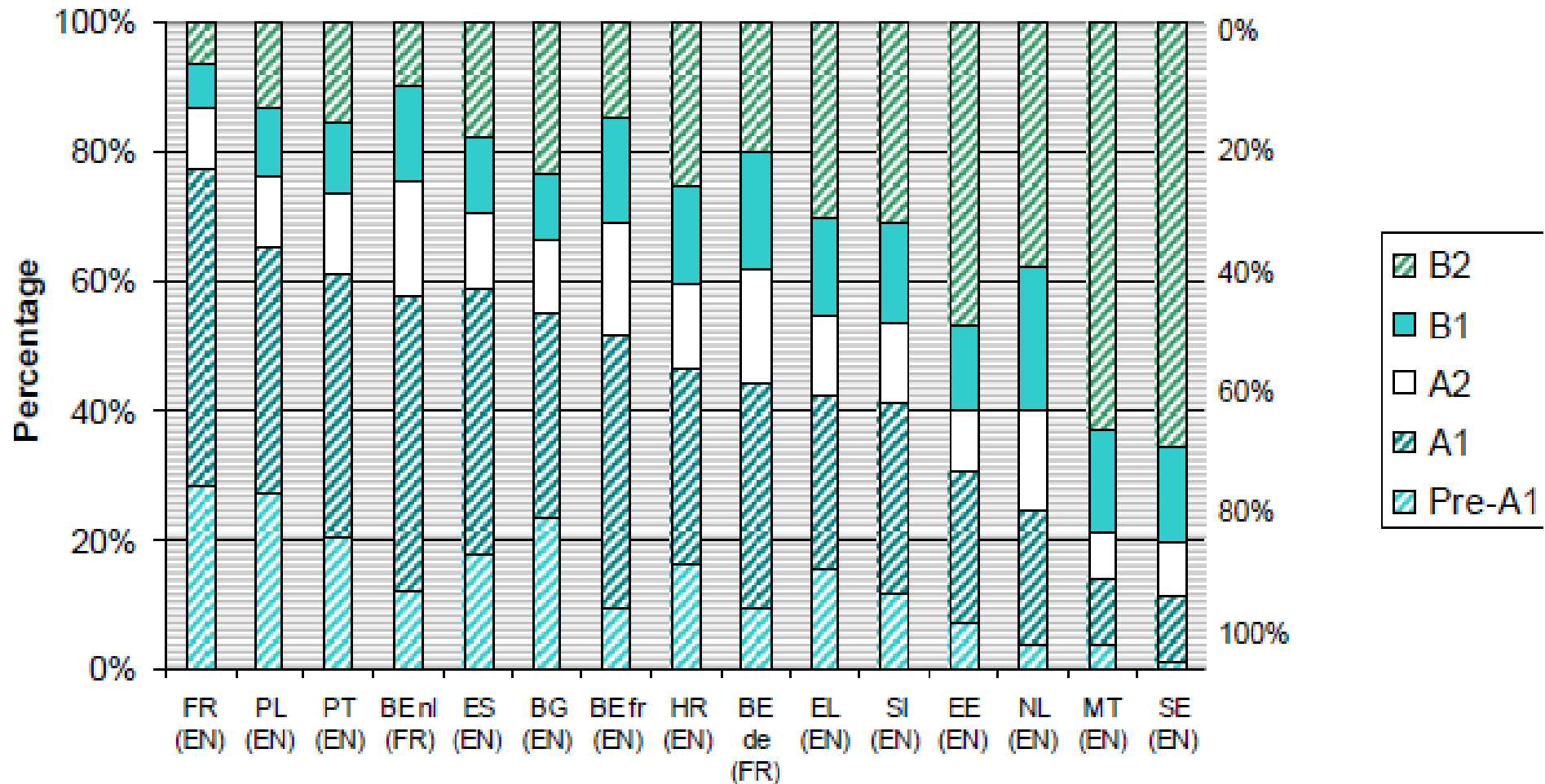
Gap between targets and reality

First European Survey on Language Competence

- Projected commissioned by the European Commission
- Large-scale evaluation of 54,000 students in last year of lower-secondary or 2nd year of upper secondary
- Focused on 5 major European languages
- Tested students in first and second foreign languages taught
- Carried out in 18 education systems
 - *In 15 of the 18 systems, English was the first foreign language*

Gap between targets and reality

CEFR levels First language Reading



Distilling (some of) the opportunities and challenges

Blanket “one-size-fits-all” targets:

- Setting blanket targets for whole populations often leads to failure to reach those targets..
- **Opportunity:** use the rich detail in the CEFR to help identify targets relevant to different sectors.
- **Opportunity:** use the adaption of the CEFR to find out what other targets and information might be relevant to groups of learners and individuals, and for employers, higher education sector, etc.

Distilling (some of) the opportunities and challenges

Setting unrealistic targets:

- Setting targets which are not realistic for the context or time frames, or for the actual needs of the learners, creates impression of failure, and leads to frustration.
- **Opportunity:** use an understanding of the different levels of proficiency in the CEFR to engage with the learners, teachers, and future employers help identify levels relevant to different sectors.
- **Opportunity:** build a strong local research capability by helping teachers and teacher educators to carry out real needs analysis.

Distilling (some of) the opportunities and challenges

Focusing on assessment as driver:

- Accountability and evaluating progress is important, but only using aspects of proficiency that can be easily measured limits real proficiency
- **Opportunity:** use the CEFR and good test development methodology to develop useful measures of proficiency
- **Opportunity:** But also look beyond the things that are “easily” measured and quantified to identify other ways of evaluating progress

Distilling (some of) the opportunities and challenges

Focusing on assessment as driver:

- 30 years of research on “impact” and “washback” tell us that assessment reform alone does not lead to change in the wider system
- **Opportunity:** Build into teacher training better understanding of how the CEFR can be used in curriculum development and in the classroom
- **Opportunity:** Build into teacher training better informed “consumers” of tests through assessment literacy. This helps teachers teach to the construct, not just the test

Distilling (some of) the opportunities and challenges

Fitting a square peg into a round hole:

- The CEFR was always meant as a starting point, and a framework to allow educators to discuss similarities and differences in contexts
- **Opportunity:** Build local research capability to investigate what aspects of the CEFR work for the local context, and what might be different.
- **Opportunity:** Adapt and introduce change into the framework in a principled way
- **Opportunity:** disseminate findings locally, regionally and internationally: become a part of the discussion

Linking to the CEFR

- ❖ The proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more states or degrees of performance. (Cizek, 1993)
- ❖ The results “are seldom, if ever, purely statistical, psychometric, impartial, apolitical, or ideologically neutral activities.” (Cizek & Bunch, 2007)

Standard-setting studies in Europe

- ❖ Aptis (O'Sullivan, 2015)
- ❖ City & Guilds Communicator IESOL Examination (O'Sullivan, 2008)
- ❖ Dutch state foreign language examinations (Berger, Kuiper, & Maris, 2009; Noijons & Kuipers, 2010)
- ❖ TestDAF (Kecker & Eckes, 2010)
- ❖ Trinity College Examinations (Papageorgio, 2007; Papageorgio, 2009)

Studies outside Europe

- ❖ TOEFL PBT (Tannenbaum & Wylie, 2005)
- ❖ TOEFL iBT (Tannenbaum & Wylie, 2008)
- ❖ GEPT, Taiwan (Wu & Wu, 2010)
- ❖ GEPT, Taiwan (Brunfaut & Harding, 2014)
- ❖ EIKEN, Japan (Dunlea & Figueras, 2012)
- ❖ EIKEN, Japan (Dunlea, 2016)
- ❖ *VSTEP, Vietnam (Tran, Nguyen, Dang, Nguyen, Nguyen, Huynh, Do, Nguyen, Davidson)*

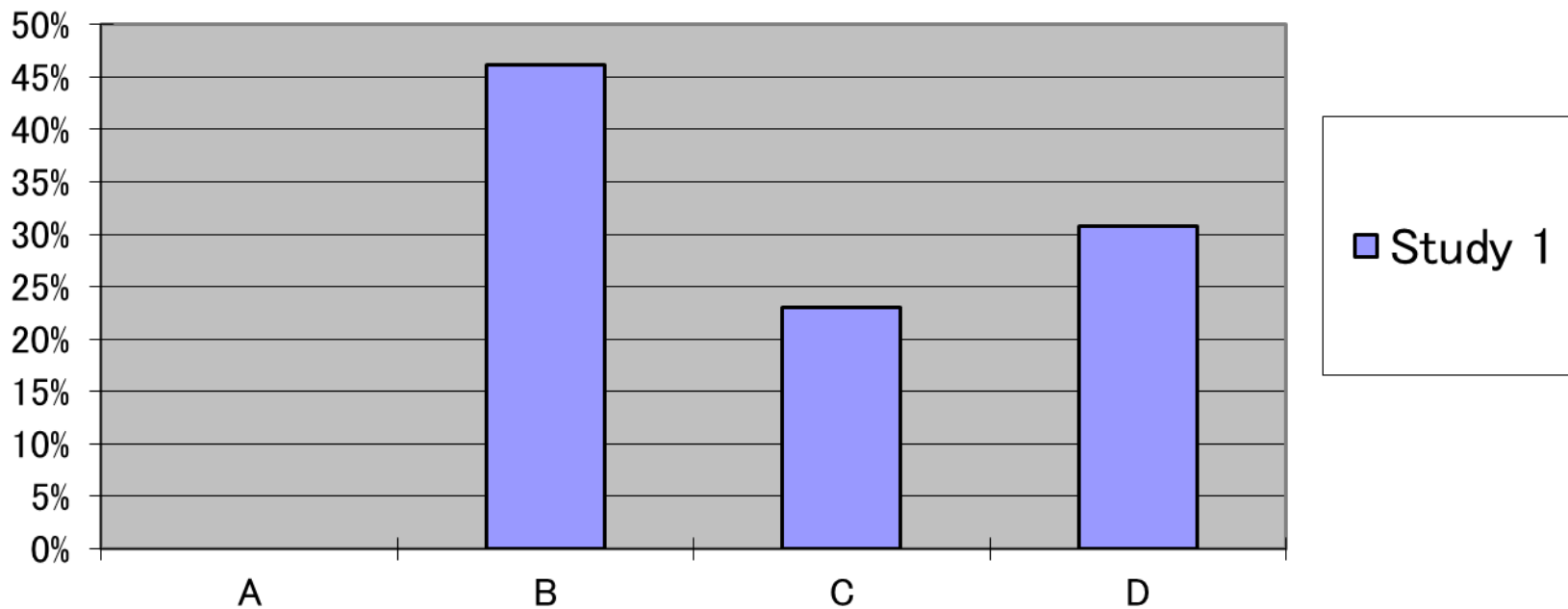
Case study: EIKEN in Japan

EIKEN Grade	CEFR	Example of recognition/use	Examinees in 2008
1	C1	International admissions to graduate and undergraduate programs; MEXT benchmark for English instructors (Pre-1)	22,055
Pre-1	B2		71,533
2	B1	MEXT benchmarks for high school graduates	312,034
Pre-2	A2		503,638
3	A1	MEXT benchmark for junior high school graduates	661,798
4			464,819
5			306,745

Standard setting in the EIKEN project

Study	Grade	Skills	Test centered	Examinee centered
Study 1	G1 / GP1	R, L, S, W	Basket Modified Angoff	Paper selection (Writing)
Study 2	G2-G5	R, L, S, W (W = indirect)	Basket Modified Angoff	
Study 3	G1-G3	Speaking	Variation of Basket method for tasks	Variation of Basket method performance
Validation study	GP1	Reading		Contrasting Groups

CEFR Linking Project in Japan (2007 – 2010)



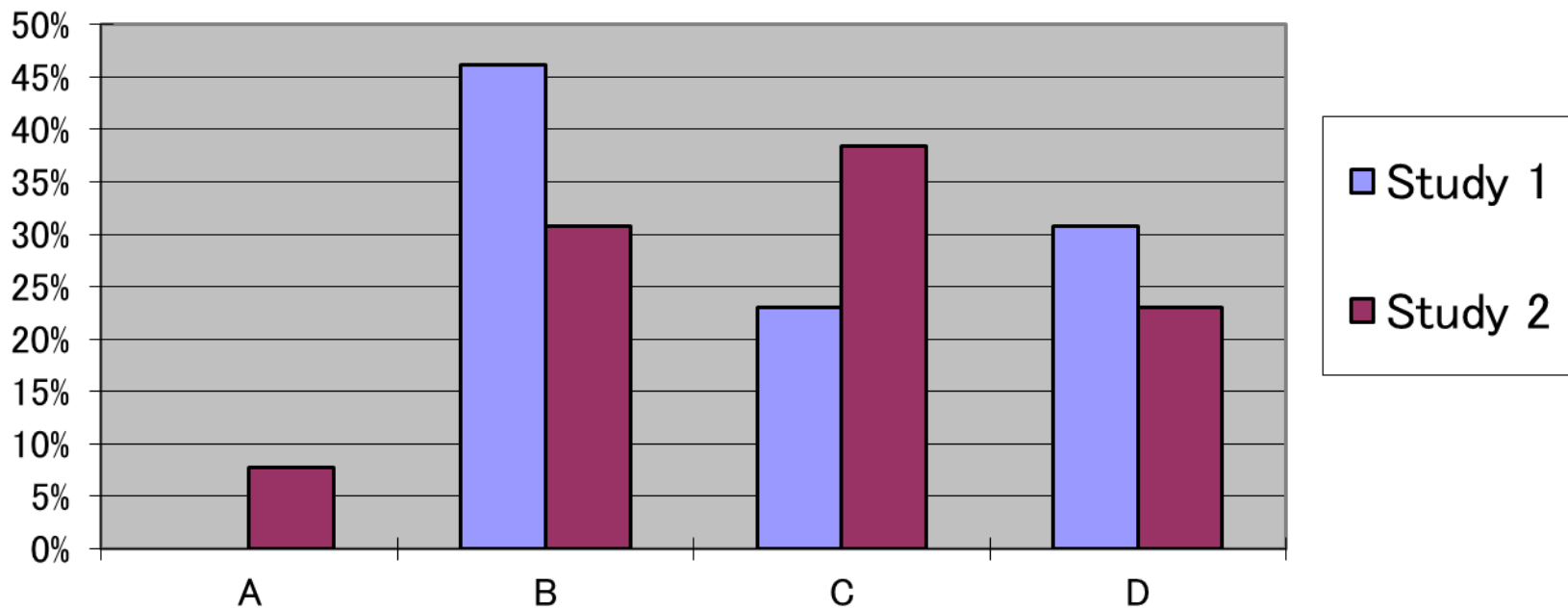
A) I had read the CEFR and was familiar with its aims and contents, including the Common Reference Levels.

B) I was familiar with the aims of the CEFR, but had not studied it in detail.

C) I had heard of the CEFR but was not familiar with its aims or contents.

D) I had not heard of the CEFR.

CEFR Linking Project in Japan (2007 – 2010)



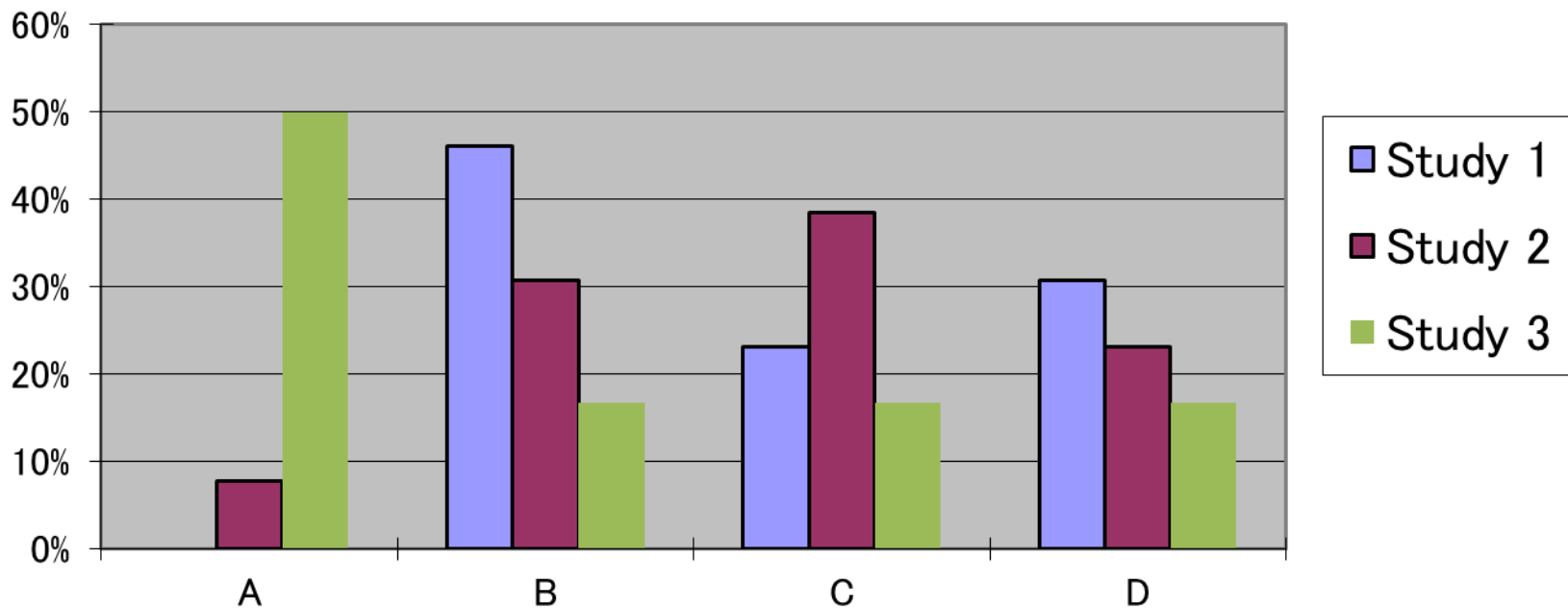
A) I had read the CEFR and was familiar with its aims and contents, including the Common Reference Levels.

B) I was familiar with the aims of the CEFR, but had not studied it in detail.

C) I had heard of the CEFR but was not familiar with its aims or contents.

D) I had not heard of the CEFR.

CEFR Linking Project in Japan (2007 – 2010)



A) I had read the CEFR and was familiar with its aims and contents, including the Common Reference Levels.

B) I was familiar with the aims of the CEFR, but had not studied it in detail.

C) I had heard of the CEFR but was not familiar with its aims or contents.

D) I had not heard of the CEFR.

Standard setting in the EIKEN project

	Qualifications	Number
Study 1	<ul style="list-style-type: none">■ At least 3 years teaching at university level in Japan■ Knowledge and experience of EIKEN tests	13
Study 2	<ul style="list-style-type: none">■ Experienced high school and junior high school teachers (all had at least 5 years in 1 sector, most had worked in both)■ Knowledge and experience of EIKEN tests	13
Study 3	<ul style="list-style-type: none">■ At least 3 years teaching at university level in Japan■ Knowledge and experience of EIKEN tests	12

Typical problems: local solution

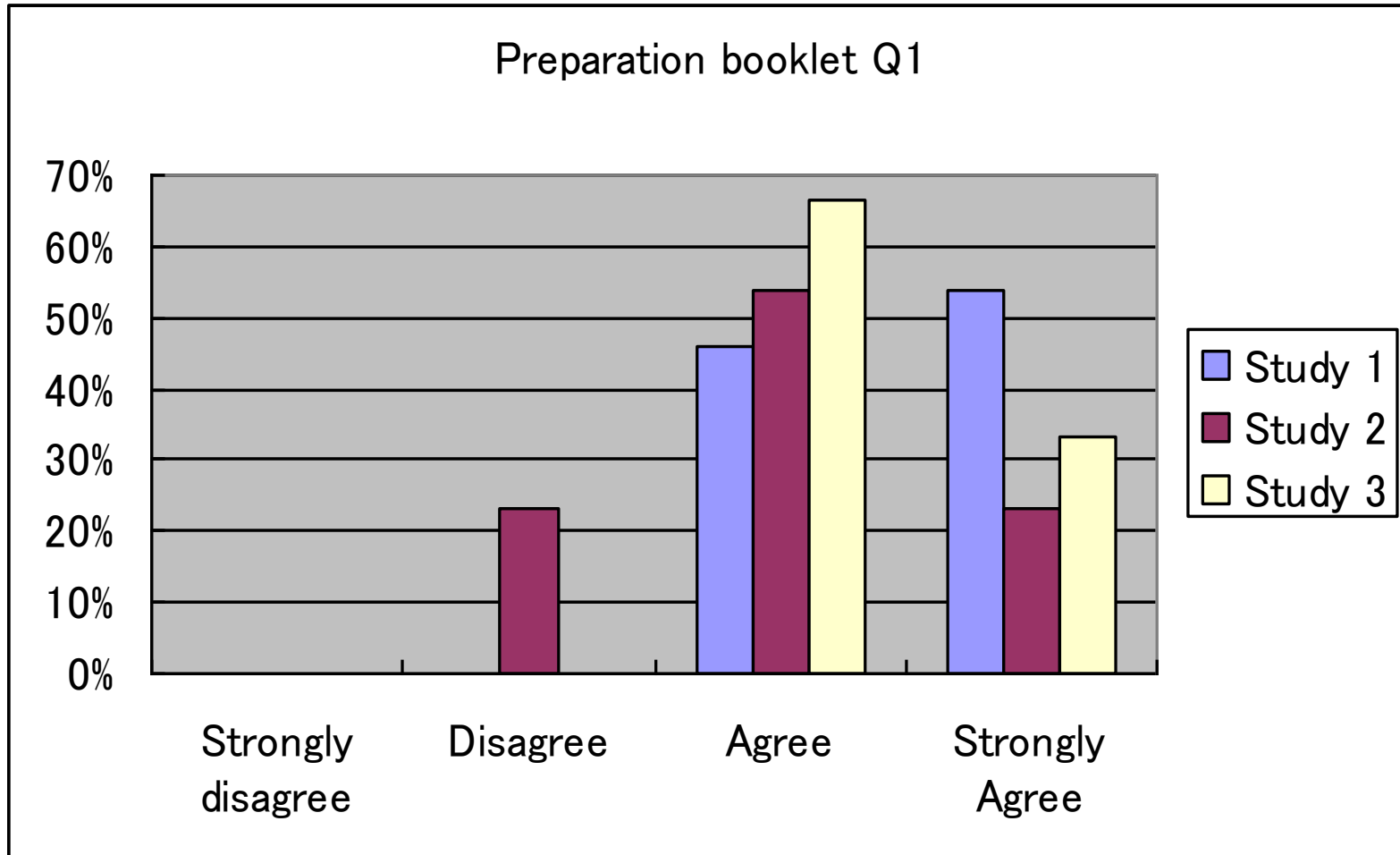
Workshop Day	Day 1	Day 2	Day 3	Day 4
Judges able to attend	13	13	12	10

- ❖ Anticipated that judges would have low familiarity with CEFR
- ❖ Use self-study preparation booklet for judges to do CEFR familiarization tasks, adapted for self-study by project team, before workshop begins

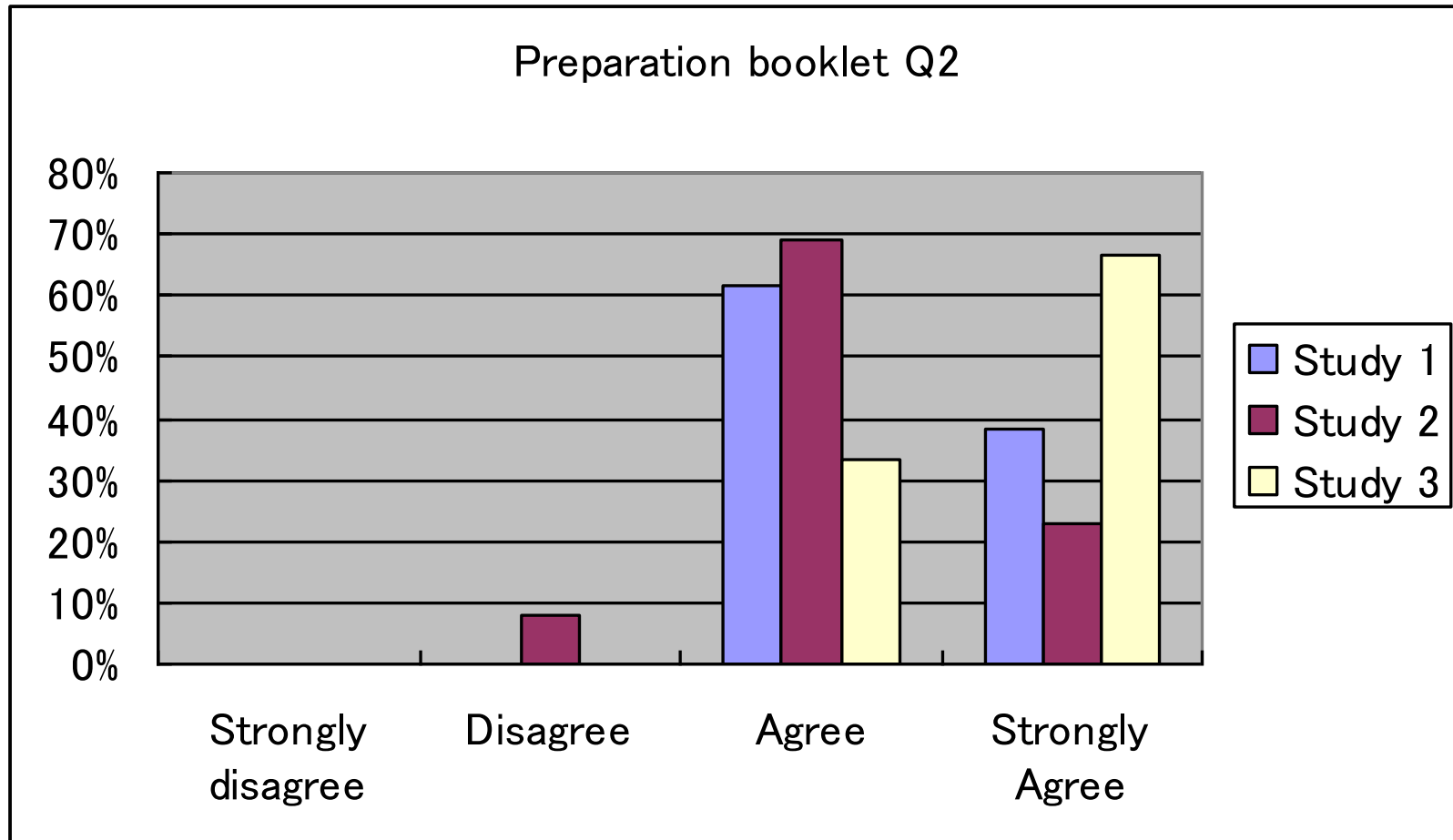
Self-study preparation booklet

Tasks	Focus	Description of activity
Tasks 1 & 2	Global Scale	Reflection, using scale to consider level of own learners, summarizing significant level features for B1, B2, C1
Tasks 3 & 4	Self-assessment grid	Rating own level, reviewing (and if appropriate revising) level descriptions made in Task 2
Tasks 5 & 6	Illustrative scales for listening	Re-ordering of jumbled descriptors within each scale, raters put descriptors in level they think appropriate
Task 7	Overall Reading scale	Reordering jumbled descriptors from Overall Reading Scale
Task 8	Overall Listening and Overall Reading	Comparing Overall Listening and Reading scales, noting any significant differences between key words and definitions in the two scales

The preparation booklet gave me a clear understanding of the purpose of the project



The explanations and tasks in the preparation booklet helped me understand the structure of the CEFR and the Common Reference Levels.



Procedural validity: judges' feedback

	Questions	Mean
Q1	The preparation booklet gave me a clear understanding of the purpose of the project.	3.5
Q2	The explanations and tasks in the preparation booklet helped me understand the structure of the CEFR and the Common Reference Levels.	3.4
Q3	The group discussion of the CEFR at the start of the workshop aided my understanding of the CEFR and the Common Reference Levels.	3.3
Q4	The time provided for the discussion was adequate.	3.1
Q5	There was an equal opportunity for everyone to contribute his/her ideas during the discussion.	3.3

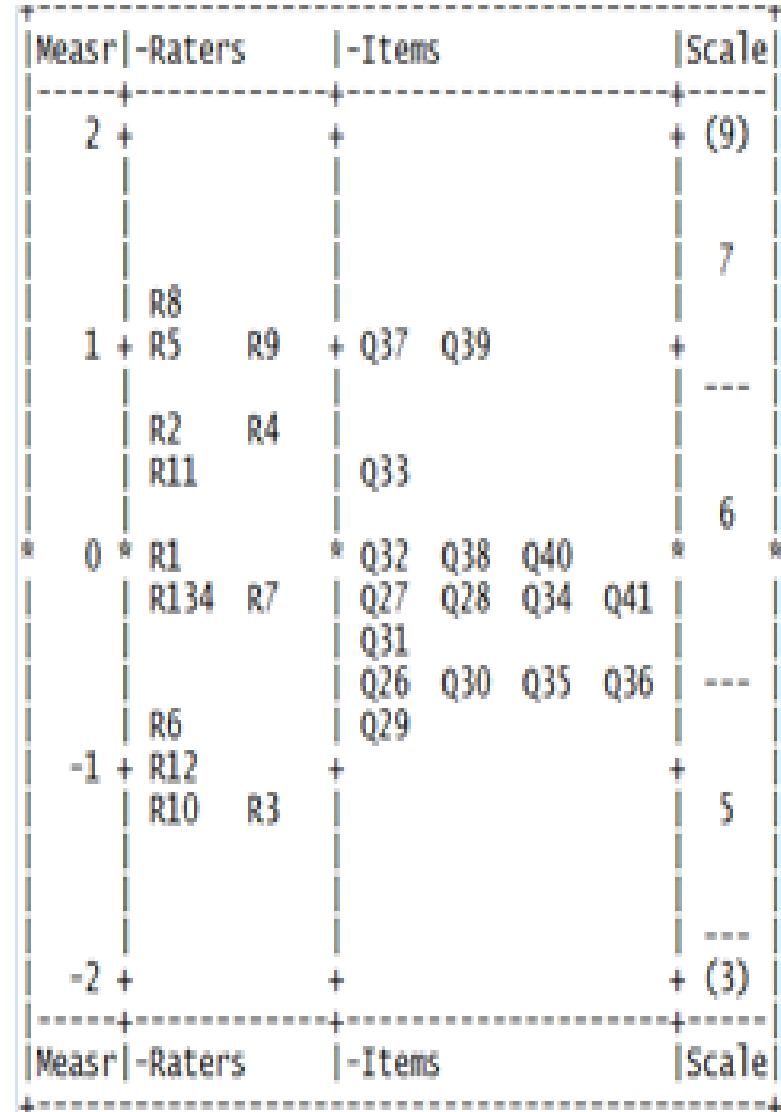
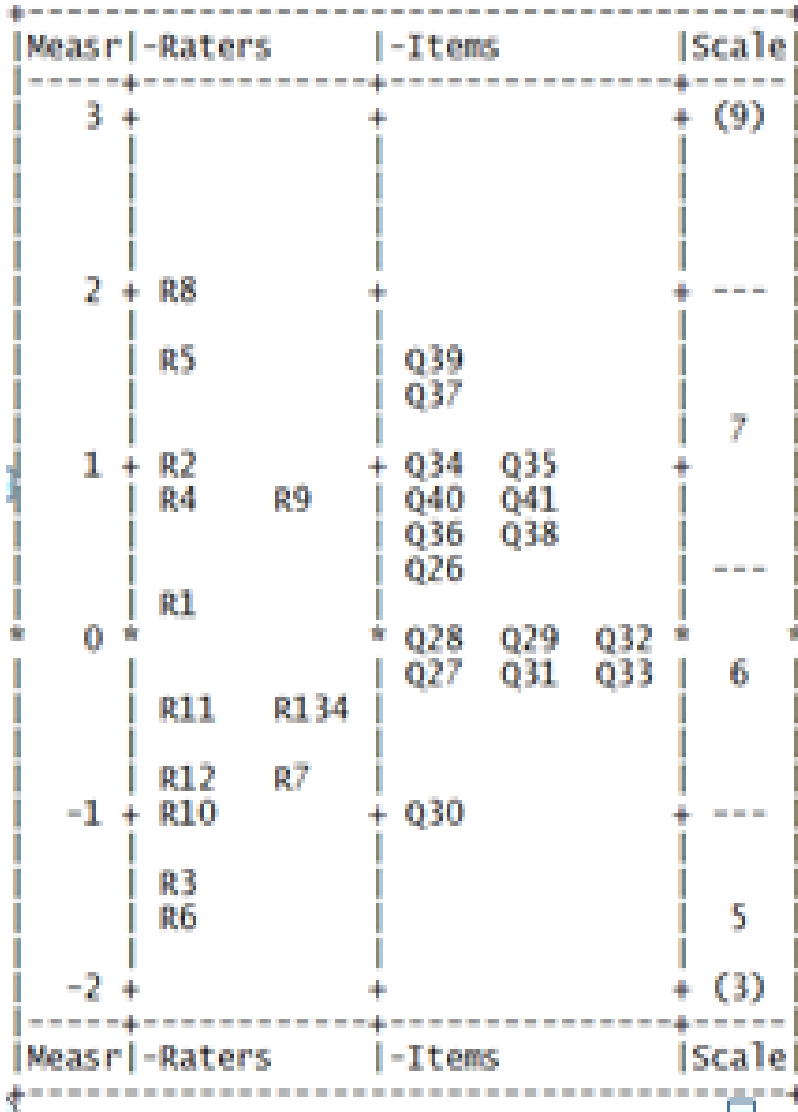
Procedural validity: judges' feedback

	Questions	Mean
Q6	The training tasks with the items supplied by the Council of Europe were useful.	3.4
Q7	The time provided for training with the Council of Europe items was adequate.	3
Q8	The explanation of the Basket Method was adequate and I felt able to undertake the rating tasks for the listening, reading, and vocabulary items.	3.2
Q9	The explanation of the Modified Angoff Method was adequate and I felt able to undertake the rating tasks for the listening, reading, and vocabulary items.	3
Q10	The time provided for rating the EIKEN listening, reading, and vocabulary items was adequate.	3.4
Q11	The feedback on item difficulty of the EIKEN listening, reading, and vocabulary items was useful.	3.2

Multiple methods: the rationale

- Basket method was chosen to act as a “primer,” helping raters to form an initial impression of items in terms of the CEFR before using the more conceptually difficult Angoff
- By forming an impression of the difficulty of each item first (with the Basket method), raters would find it easier to make probability judgments when conceptualizing the “100 minimally competent test takers” used in the Angoff procedure

G1 & Pre-1 facet maps (Reading)



G1 Rater Measurement Report

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq
104	16	6.50	6.51	-.85	.30	3.57
84	16	5.25	5.28	.83	.29	1.41
74	16	4.63	4.65	1.63	.28	1.33
99	16	6.19	6.19	-.42	.30	1.11
110	16	6.88	6.89	-1.40	.31	1.07
69	16	4.31	4.31	2.03	.29	.79
105	16	6.56	6.57	-.94	.30	.66
103	16	6.44	6.44	-.76	.30	.59
84	16	5.25	5.28	.83	.29	.55
113	16	7.06	7.08	-1.69	.32	.52
83	16	5.19	5.22	.91	.29	.31
91	16	5.69	5.70	.26	.29	.26
99	16	6.19	6.19	-.42	.30	.24
93.7	16.0	5.86	5.87	.00	.29	.95
13.4	.0	.84	.83	1.12	.01	.84
13.9	.0	.87	.87	1.17	.01	.88

Grade Pre-1 Reading: Facet map

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq
81	16	5.06	5.06	1.17	.29	1.16
83	16	5.19	5.18	1.01	.28	1.36
84	16	5.25	5.25	.93	.28	1.18
87	16	5.44	5.44	.70	.28	1.07
89	16	5.56	5.57	.55	.28	1.15
92	16	5.75	5.76	.32	.28	.48
97	16	6.06	6.07	-.07	.28	.71
98	16	6.13	6.13	-.15	.28	.66
99	16	6.19	6.19	-.23	.28	.51
105	16	6.56	6.56	-.71	.29	.49
109	16	6.81	6.80	-1.04	.29	1.61
111	16	6.94	6.92	-1.20	.29	.55
112	16	7.00	6.98	-1.29	.29	1.58
95.9	16.0	6.00	5.99	.00	.29	.96
10.5	.0	.66	.65	.83	.00	.40
10.9	.0	.68	.68	.86	.00	.42

Multiple methods: the rationale

- Do educators in the context of Europe who are experienced at using the CEFR for teaching and assessment demonstrate a similar estimation of the EIKEN tests in relation to the CEFR as was derived through Standard Setting Panels 1 and 2?
- The external validation study was undertaken in collaboration with a researcher in Spain who was prepared to collaborate in the recruitment of participants and the administration of EIKEN tests to those participants in that local context.

Multiple methods: the rationale

- Kane (2001b, p. 75) recommends replicating standard, suggesting that using different methods and participants “would provide an especially demanding empirical check on the appropriateness of the cutscore.”
- Kane (2001b, p. 75) : “the Angoff method were used in the original study, the new study might involve an examinee-centered method

Multiple methods: the rationale

- Contrasting Groups method was chosen
 - *Participants, who are unaware of examinees' actual test scores, make judgments about each examinee as to their mastery/nonmastery status. . . . Participants' category judgments are used to form distributions of total test scores for each of the two groups. . . . The two distributions are then plotted and analyzed to arrive at a cut score that distinguishes group membership. Cizek and Bunch (2007, p. 107)*

Validation study in Europe

Number of classes	Number of teachers	Number of students
10	6	170

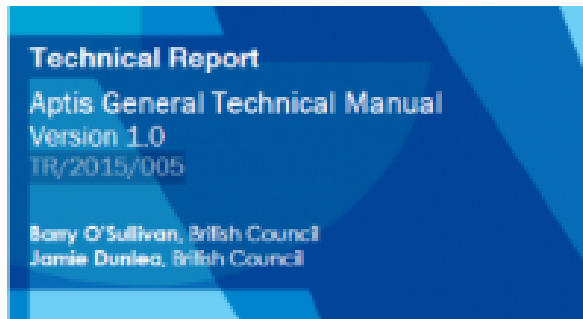
Method	Raw score	Percent	
Mean of means	30.0	73.2	
Midpoint of medians	30.0	73.2	
Overlap of distribution plots	28.0	68.3	
Logistic regression	30.3	73.9	

Test specifications

The chief tool of language test development is a test specification, which a test is a generative blueprint from which test items or tasks can be produced. A well-written test specification (or “spec”) can generate many equivalent test tasks.

(Lynch & Davidson, 2002)

Case studies: where to find detailed specs



[Aptis General Technical Manual Version 1.0](#)

TR/2015/005 This manual describes the content and technical properties of Aptis General, the standard English language assessment product offered within the Aptis test system.



Test of English for Academic Purposes

<http://www.eiken.or.jp/teap/group/report.html>

<https://www.beds.ac.uk/crella/projects/teap>