ALTE International Conference 2014

Getting to know the Minimally Competent Person

Beate Zeidler
10th April 2014

The context: Standard Setting

... 50th method (after Kaftandjieva‘s 34)?

This is not about a new method – only about *one component* in Standard Setting studies that has not received a lot of attention – the

- Minimally Competent Person (MCP)
- Minimally Qualified Candidate (MQP)
- Just Qualified Candidate (JQC)
- Borderline Candidate
- …

…whom we need to conceptualise in a standard setting.

I shall describe a Standard Setting study with special attention to the MCP:

- Rationale
- Method and outcome
- Discussion

Is there a way aroud the MCP?

- The aim of standard setting is to define the pass score
- In order to do this, most methods require the conceptualisation of the borderline candidate
- A recently proposed method (Prototype Group Method, Thomas Eckes (2012)) focuses on „prototype candidates“ (typical, „middle of the band“ candidates) and uses a mathematical model to define the borderline, but requires large samples of test takers
- So, in most contexts, we still have to work with a model of the borderline candidate

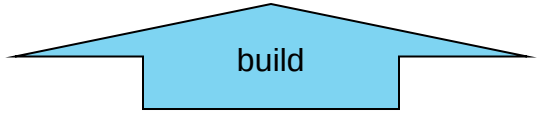
The challenge for all standard-setting methodologies is to effectively translate a participant's mental model of the target examinee (e.g., barely proficient student) into judgments that communicate the participant's recommendation of a value that characterizes the point of separation between one or more categories.

Buckendahl (2005), 219

„Point of separation“ = Cut score on a test =

Expected MCP test performance in new test

“Mental model of the target examinee” (= MCP)

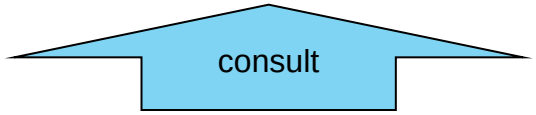


„Point of separation“ = Cut score on a test =

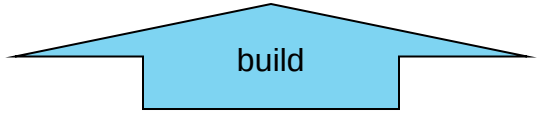
Expected MCP test performance in new test

Level descriptors =

Expected MCP live performance



“Mental model of the target examinee” (= MCP)



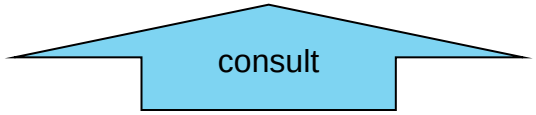
„Point of separation“ = Cut score on a test =

Expected MCP test performance in new test

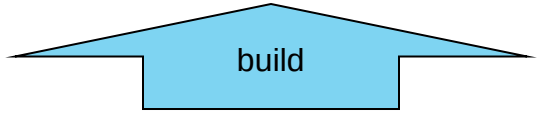
Expert knowledge

Level descriptors =

Expected MCP live performance



“Mental model of the target examinee” (= MCP)

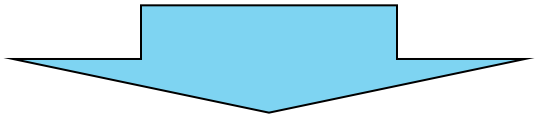


„Point of separation“ = Cut score on a test =

Expected MCP test performance in new test

Can read straightforward factual texts on subjects related to his/her field and interest with a satisfactory level of comprehension.

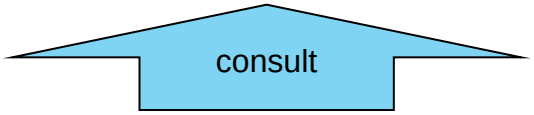
Actual MCP live performance



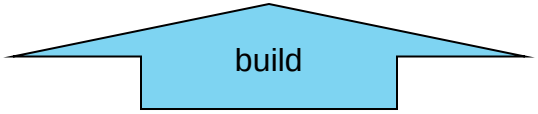
Expert knowledge

Level descriptors =

Expected MCP live performance



"Mental model of the target examinee" (= MCP)



„Point of separation" = Cut score on a test =

Expected MCP test performance in new test

Can read straightforward factual texts on subjects related to his/her field and interest with a satisfactory level of comprehension.

Can read straightforward factual texts on subjects of general interest and answer multiple choice questions targeted at salient details.

Actual MCP live performance

But we only see a subset of this …

Actual MCP test performance

Expert knowledge

Level descriptors =

Expected MCP live performance

consult

"Mental model of the target examinee" (= MCP)

build

„Point of separation" = Cut score on a test =

Expected MCP test performance in new test

Can read straightforward factual texts on subjects related to his/her field and interest with a satisfactory level of comprehension.

Can read straightforward factual texts on subjects of general interest and answer multiple choice questions targeted at salient details.

Actual MCP live performance

But we only see a subset of this …

Actual MCP test performance

Expert knowledge

Level descriptors =

Expected MCP live performance

consult

So should the mental model be informed by MCPs' test performance?

"Mental model of the target examinee" (= MCP)

build

„Point of separation" = Cut score on a test =

Expected MCP test performance in new test

Why include test-solving strategies?

We want to predict the performance of the MCP on a test.

A test, however authentic, does not capture real life: it may be ***more*** or ***less*** difficult than real life tasks.

- The proof of comprehension (in the receptive skills) is not success in an action, but the answer to an item
- Candidates may not choose texts according to their need or interest
- Candidates may not resort to „real life“ problem solving strategies, such as asking somebody else, looking something up, or giving up altogether
- Candidates have to understand an artificial context in which their communication is supposed to take place
- Context knowledge is at best restricted, at worst not available to them, hence anticipation is more difficult than in real life

It has been shown that (at least for some tests) text-item-interaction, especially vocabulary overlap between text and item, is among the best predictors for an item‘s difficulty (not only measures of text difficulty, e.g. number of subclauses, as the construct would demand).

Freedle, Roy/Kostin, Irene (1993), The Prediction of TOEFL Reading Comprehension Items Difficulty for Expository Prose Passages for Three Item Types: Main Idea, Inference, and Supporting Idea Items. ETS Report RR-93-13, TOEFL-RR-44

Kostin, Irene (2004), Exploring Item Characteristics That Are Related to the Difficulty of TOEFL Dialogue Items, ETS Report RR-04-11

Significant predictors of item difficulty:

Kostin 2004

Sentence complexity	Text complexity	Vocabulary complexity	Cognitive load	Content factors	Systemic factors
+ Density dependent clauses/total complete clauses	+ Referentials	+ Density infrequent words	+ Inference is necessary olve item	Text type	Item length/complexit
+ Density compound/sentences	+ Negatives	+ Idioms	+ Candidate has to construct situation	Topic	Text/item overlap
+ Density complex/sentences	+ Genitive constructions	+ Compound words	+ Integration of various pieces of information	Domain	
+ Density complex-compound/sentences	+ Cohesive devices	+ Modal particles	Position of relevant information (memor capacity)		
+ Density compl/comp-comp/sentences			Position of signal for relevant information (memory capacity)		
+ Fronted structu					

Freedle/Kostin 1993
(Zeidler, 2010)

Hypotheses on relevant parameters – significant:

Kostin 2004

Sentence complexity	Text complexity	Vocabulary complexity	Cognitive load	Content factors	Systemic factors
+ Density dependent clauses/total complete clauses	+ Referentials	+ Density infrequent words	+ Inference is necessary olve item	Text type	tem length/complexit
+ Density compound/sentences	+ Negatives	+ Idioms	+ Candidate has to construct situation	Topic	Text/item overlap
+ Density complex/sentences	+ Genitive constructions	+ Compound words			
+ Density complex-compound/sentences	+ Cohesive devices	+ Modal particles			
+ Density compl/comp-comp/sentences					
+ Fronted stru					

Difficulties that are only there because candidates are taking a test!

Freedle/Kostin 1993

(Zeidler, 2010)

So it may make sense to pay attention to test-specific language behaviour when constructing the MCP model.

Ways to help standard setting participants to form a mental model:

- Taking participants‘ expert knowledge for granted
- Working from level descriptors without reference to concrete candidates
- Working from level descriptors and derive a notion of the MCP from group discussion
- Working from a description of „good“ vs. „weak“ proficiency (i.e. constructing own level descriptors)
- Trying to describe the MCP him/herself (e.g. writing down MCP characteristics for reference during the standard setting)

As there are item-centered and candidate-centered methods for standard setting, there are apparently descriptor-centered and candidate-centered methods for target level definition.

A few examples …

Definition of target candidate characteristics

CEFR descriptor task

At standard setting workshop:

In preparation for the standard-setting meeting, material to familiarize the judges with the CEFR levels was prepared. Fifty-six reading, 71 listening, 17 grammar and 25 vocabulary sentence-level statements from the CEFR descriptors (see sample in Appendix 1) were presented to the judges asking them to choose the CEFR level they belong to (A1-C2). No indication of the level was presented to the judges. For faster analysis of results, the judges were asked to use numbers instead of levels in the following way: A1-1; A2-2; B1-3; B2-4; C1-5; and C2-6. The "atomization" of the descriptors into short statements, based on Kaftandjieva and Takala (2002), aimed to familiarize the judges with all constituent statements of the descriptors, which usually contain a number of sentence-level statements.

Item difficulty task

At standard setting workshop:

In order to help judges obtain a better understanding of the difficulty of test items and how this relates to the judgment task, the training material asked judges to rank a number of listening and reading MET pilot items from easiest to most difficult.

Papageorgiou (2010), Setting Cut Scores on the Common European Framework of Reference for the Michigan English Test, Testing and Certification Division, English Language Institute, University of Michigan, Ann Arbor, p. 2-3

Definition of target candidate characteristics

Target candidate task

Activity prior to standard setting:

"Prior to the study, the members on both panels were given an assignment … to review selected tables from the CEFR (the Web site to the CEFR was provided) 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. … As they completed this pre-study assignment, they were asked to consider what distinguishes a candidate with just enough skills to be considered performing at a specific CEFR level from a candidate with not enough skills to be performing at that level."

Tannenbaum/Wylie (2008), Linking English-Language Test Scores Onto the Common European Framework of Reference: An Application of Standard-Setting Methodology (RR-08-34)

Definition of target candidate characteristics

Activity at standard setting:

"During the study, time was spent developing an agreed upon definition of the minimum skills needed to be considered performing at each CEFR level. The panelists were formed into three table groups and each group was asked to define and chart the skills of the least able candidate for A2, B2, and C2 levels; this was done separately for Writing, Speaking, Listening, and Reading. Panelists referred to their pre-study assignments and to the CEFR tables for each modality. Given that the focus for the standard setting was on the candidate who has just enough skills to be at a particular level, panelists were reminded that the CEFR describes the abilities of someone who is typical of a particular level. … A whole-panel discussion of each group's charts followed, and a final agreed upon definition was established for three levels: A2, B2, and C2. Definitions of the least able candidate for A1, B1, and C1 levels were then accomplished through whole-panel discussion, using the A2, B2, and C2 descriptions as boundary markers."

Tannenbaum/Wylie (2008), Linking English-Language Test Scores Onto the Common European Framework of Reference: An Application of Standard-Setting Methodology (RR-08-34)

Definition of target candidate characteristics

Activity at standard setting – outcome:

Panel 1 Indicators of CEFR Definitions of Proficiency in Listening

Listening skills of just-qualified A1

Can understand very slow speech with familiar words and basic phrases on here and now.

Can understand short and slow speech with pauses and repetition.

Requires sympathetic speaker.

Listening skills of just-qualified A2

Can understand short, clearly, slowly, and directly articulated concrete speech on simple, everyday, familiar topics/matter.

Can understand formulaic language (basic language and expressions).

Can understand short directions, instructions, descriptions.

Can extract relevant, important information from recorded messages.

Listening skills of just-qualified B1

Can understand main points.

Can understand clear, standard speech on familiar matters and short narratives when presented relatively slowly

Will sometimes need repetition and clarification in conversation.

Can follow broadcast information carefully delivered. (Example: BBC World but not SkyNews)

Can deduce sentence meaning.

Tannenbaum/Wylie (2008), Linking English-Language Test Scores Onto the Common European Framework of Reference: An Application of Standard-Setting Methodology (RR-08-34)

Definition of target candidate characteristics – *Eng B1-B2*

Activity at standard setting:

1) CEFR scales, receptive skills, underline key words (“typical”)

2) CEFR = “typical” skills → focus on “borderline” skill

Tannenbaum/Wylie (2008) tables

3) The raters were asked to form an idea of the B1 and B2 Minimally Competent Person, using data from previous B1 and B2 exam runs. They were provided with the questions from these exams (one version each) and with p (facility) values reached by test takers who reached a result around the cut score of the respective exam, and to note down their observations.

p values (Sample: 609 candidates from B1 exam)

	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
a	40,68	13,56	13,56	0	1,69	15,25	28,81	10,17	15,25	28,81	32,2	13,56	16,95	20,34	33,9
b	0	0	1,69	0	0	66,1	66,1	52,54	20,34	20,34	16,95	45,76	54,24	27,12	59,32
c	20,34	1,69	5,08	3,39	0	18,64	5,08	37,29	62,71	50,85	50,85	38,98	28,81	52,54	6,78
d	0	1,69	3,39	0	0	0	0	0	0	0	0	0	0	0	0
e	1,69	1,69	45,76	0	0	0	0	0	0	0	0	0	0	0	0
f	0	0	0	0	1,69	0	0	0	0	0	0	0	0	0	0
g	0	0	3,39	93,22	0	0	0	0	0	0	0	0	0	0	0
h	0	0	3,39	0	1,69	0	0	0	0	0	0	0	0	0	0
i	0	0	0	0	84,75	0	0	0	0	0	0	0	0	0	0

MCPs (around cut score)

	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
a	18,32	5,45	13,37	0	0	1,98	12,38	0	8,91	23,76	28,71	40,1	11,88	7,43	21,29
b	0	0	0	0	0	91,09	84,65	6,93	5,94	10,89	22,77	31,68	82,18	1,49	75,74
c	28,71	0,5	1,49	0,99	0,5	6,93	2,97	93,07	85,15	65,35	48,51	28,22	5,94	91,09	2,97
d	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
e	0	0	62,38	0	0	0	0	0	0	0	0	0	0	0	0
f	0	0	0	0	0,99	0	0	0	0	0	0	0	0	0	0
g	0	0	0,99	97,52	0	0	0	0	0	0	0	0	0	0	0
h	0	0	0,99	0,5	0	0	0	0	0	0	0	0	0	0	0
i	0	0	0	0	96,04	0	0	0	0	0	0	0	0	0	0

Candidates grade 2/3

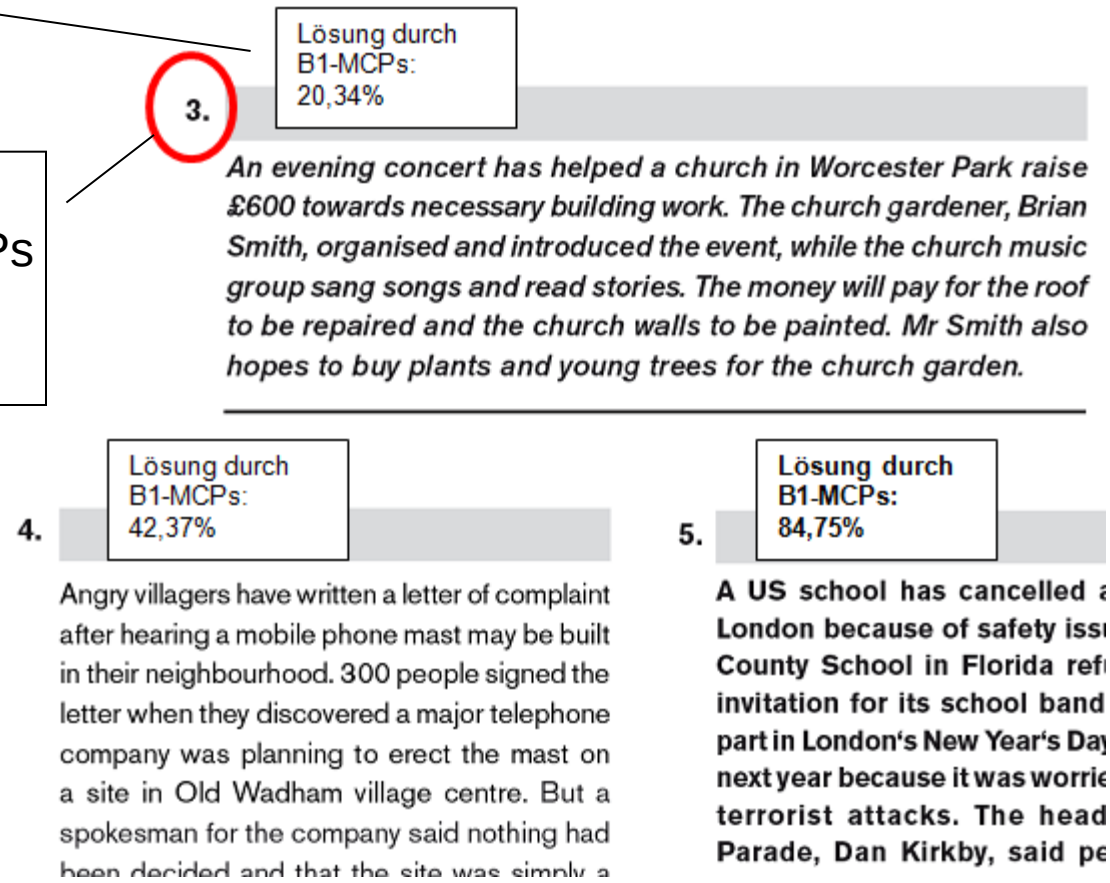
Example from Item booklet provided to Standard Setting participants

p value for MCPs

Additional information: difference between p for MCPs and p for „middle“ candidates > 0,2

= items which are more difficult for MCPs than for “middle” candidates to more than the expected extent

Instruction: look at these items especially



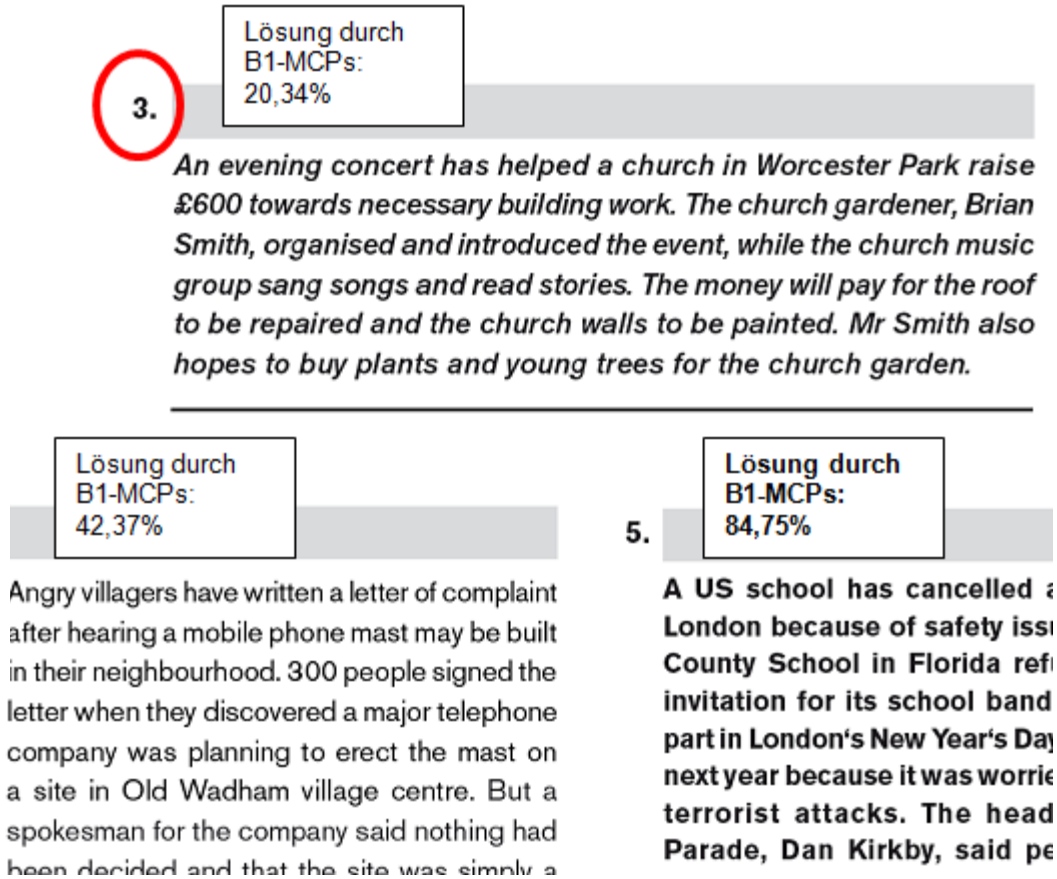
Example from Item booklet provided to Standard Setting participants

Qualitative discussion: characteristics of text/item features that are especially difficult for MCPs.

Participants were invited to write these down:

- **Abstract observations**
- **Concrete examples**

Combine own experience and evidence from data



Lösung durch B1-MCPs: 20,34%

3. *An evening concert has helped a church in Worcester Park raise £600 towards necessary building work. The church gardener, Brian Smith, organised and introduced the event, while the church music group sang songs and read stories. The money will pay for the roof to be repaired and the church walls to be painted. Mr Smith also hopes to buy plants and young trees for the church garden.*

Lösung durch B1-MCPs: 42,37%

Angry villagers have written a letter of complaint after hearing a mobile phone mast may be built in their neighbourhood. 300 people signed the letter when they discovered a major telephone company was planning to erect the mast on a site in Old Wadham village centre. But a spokesman for the company said nothing had been decided and that the site was simply a

Lösung durch B1-MCPs: 84,75%

5. **A US school has cancelled a London because of safety issu County School in Florida refu invitation for its school band part in London's New Year's Day next year because it was worrie terrorist attacks. The head Parade, Dan Kirkby, said pe**


B2 Reading
ABSTRACT
CONCRETE
CAN understand/ use
CANNOT understand/ use
3 too easy: general knowl. help + no distractor
16 difference '/.
• going for a walk
• walking
1. Distractor "e" should be changed - "too attractive"
Item 10: similar phrase structure makes item easy
understand inner text structure / logic (⇔ LC)
understand unfamiliar/ abstract topics
If you are at the lower end of a band, you fall back on safety strategies.
flexible and accurate transfer of grammatical structures
Are MCPs always MCPs? ⇒ Same strategies as B1 MCPs (non-language strategies)
Item 5: Title does not reflect content? MCP + Mid. B2
Item 6 vocab ⇒ to read
different meanings of a word e.g. "play" (Synonyms)
phrasal verbs / multi-word-verbs: "set aside"
collocations: "maintain focus" "resist the temptation"
vocab: "maintain", "resist"
B "play" misunderstood they don't know the meaning of "performance"
failure to read for details without getting stuck on unfamiliar words
cannot identify synonyms "tread - step" despite similar construction
B. Did they read and understand instructions
cannot understand "Aesthetics"
34 less frequent collocations "maintain focus"
MCPs ignore item 3 obvious clues and fixate on one word. time/reading strategies?
Item 12: x items difficult for MCPs at B2
LE, Part 2 If they start with the wrong answer, the cannot rearrange.
want to find a match, ignore X-option
cannot deal with spelling mistakes, confuses them.
item 3, 6 = lack of confidence and/or wrong time management

Raters‘ concept of B1/ B2 MCP

B1 MCPs

Reading

	Abstract	*Concrete*
Can understand/ use	Using vocabulary words “matching” but missing the concepts behind them Relies on strategies (vocab overlap etc) Recognise vocab overlap Straightforward and concrete == successful completion of the item Can understand vocab in a straightforward text	Understood the visual trick (based on “Barcelona” and “hotel”) – I14 i.e. cand. were NOT distracted by the word “Barcelona” in capital letters in one of the texts, but read the message carefully enough to solve the item The words “summer”, “waste”, “water” occur in one sentence in the text – I1 i.e. strong overlap between text and correct option “to reduce” leads to “is less” – I1 Make the connection waste -> save “I would like to” + verb

The following colour coding is applied here:
Observations relating to ...
strategies
text features
grammar
item format

telc LANGUAGE TESTS

Raters‘ concept of B1/ B2 MCP

MCPs have problems with…

B1

- Hard words especially at beginning of text
- Unusual structures (ex.: “raise … for the roof to be repaired”)
- Correct answer demands that more than 2 information items are processed

- *Gapped text*
- *Counter-intuitive items (correct answer is unexpected)*
- *If there is vocab overlap between text and wrong answer, cand. are misled into choosing the wrong answer*
- *More easily misled by the distractor being close to the the correct answer*

B2

- Idiomatic language
- Phrasal verbs
- Less frequent collocations
- Complex structures (example: “she was never offered …”)

- *Not enough time/wrong time management*
- *If there is vocab overlap between text and wrong answer, cand. are misled into choosing the wrong answer*
- *items where one option is “none of the options is correct”*
- *Cand. are misled by their hypotheses as to test construction (“this can’t be right, it is too easy”)*

→ **Input for item rating**

Modified Angoff Standard Setting task, Round 1

Eng B1-B2 items - How many of 100 MCPs can answer them correctly?

0	5	10	20	30	40	50	60	70	80	90	95	100

Rater no: .....

Task	Item	... of 100 **B1** MCPs	... of 100 **B2** MCPs	Task	Item	... of 100 **B1** MCPs	... of 100 **B2** MCPs
1	1				23		
2	2			12	24		
3	3				25		
4	4				26		
5	5				27		
	6			13	28		
6	7				29		
	8				30		
7	9				31		

Modified Angoff Standard Setting task, Round 2 (holistic)

Standard Setting
Cut score
Round 1

VERSION 1

Rater No. _____

Task No. _____

Item No. _____

B1 cut score should be: ☐ points

B2 cut score should be: ☐ points

Standard Setting
Cut score
Round 2 (FINAL)

VERSION 1

Rater No. _____

Task No. _____

Item No. _____

How confident are you of these cut scores?

B1 cut score should be: ☐ points

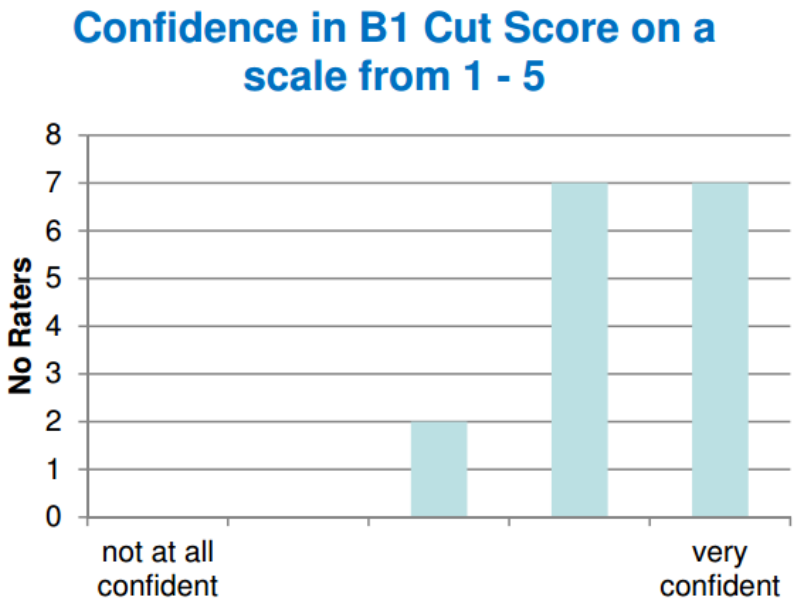
1 not at all	2	3	4	5 very
○	○	○	○	○

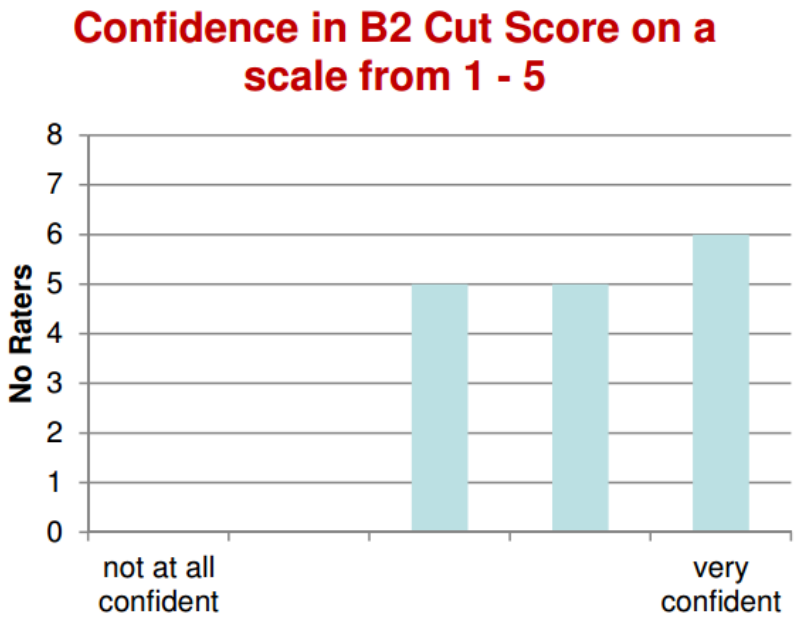
B2 cut score should be: ☐ points

1 not at all	2	3	4	5 very
○	○	○	○	○

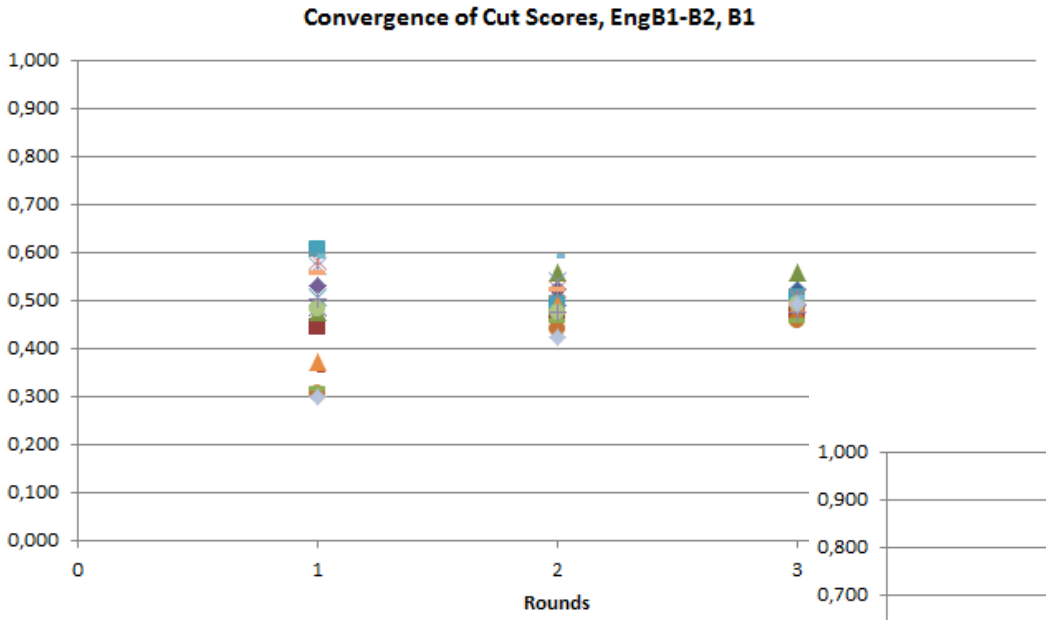
Modified Angoff Standard Setting task, Round 3 (holistic)

Modified Angoff Standard Setting task, Results

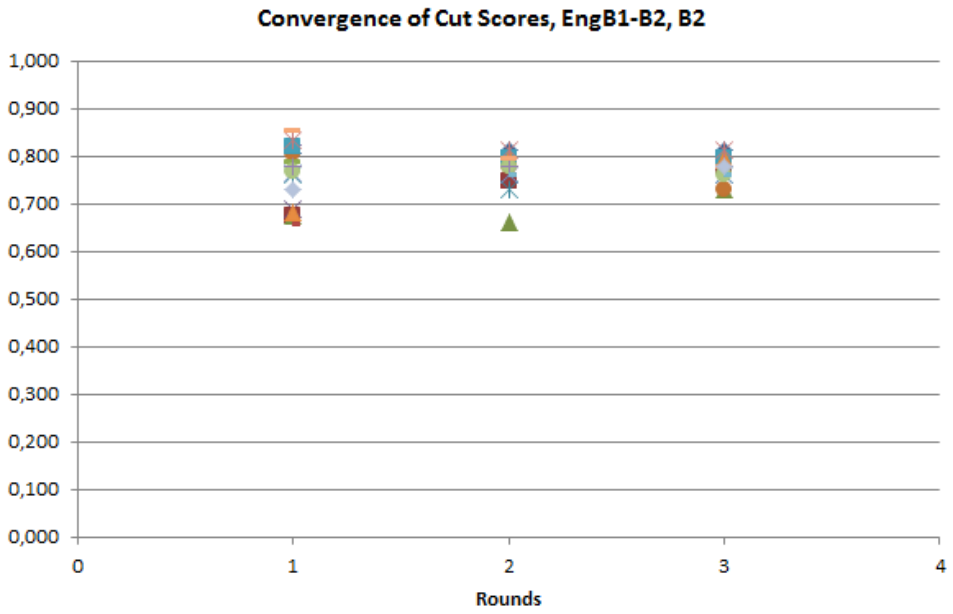




Modified Angoff Standard Setting task, Results



Cut Scores 29 (B1) and 46 (B2) of a maximum of 59 points



How can we check success?

Target candidate definition activities have two purposes:

Validity: getting closer to defining a *meaningful* cut score

Reliability: helping the group towards a more *unified* idea of a cut score

Ad 1: as there is no empirically „true“ cut score, this is an issue of plausibility. But it seems reasonable that more extensive thinking about the MCP, and integrating all features that play a role in the actual examination situation, leads to a better understanding of the threshold ability – *as it emerges in a test*

Ad 2: we can compare the range of cut scores between different standard setting workshops

Source	*Test*	*Method*	*Target candidate definition method*
Tannenbaum/ Wylie (2004)	TOEFL Structures	Angoff probabilities (0.1, 0.2, ... 0.9)	Homework: read global scale, write down key characteristics of level (not MCP) At workshop: Summarize key descriptors and produce panel-agreed version for reference during workshop
	TOEFL Reading		
	TOEFL Listening		
	TOEIC Listening		
	TOEIC Reading		
Papageorgiou (2010)	Michigan English Test, Listening	Modified Angoff (100 borderline candidates)	„Atomized" descriptors, choose right level
	Michigan English Test, Grammar+Reading		
telc	DTZ Version 1 – Version 5	Modified Angoff (100 MCPs)	Sort descriptors, complete descriptor puzzle, discuss MCP
	Deutsch Medizin B2-C1	Modified Angoff (yes/no)	Mark key characteristics in scale, discuss MCP
	English B1-B2	Modified Angoff (100 MCPs)	Mark key characteristics in scale, discuss Consider target candidate definitions from Tannenbaum/Wylie (2008) Look at MCPs' work and describe their ability

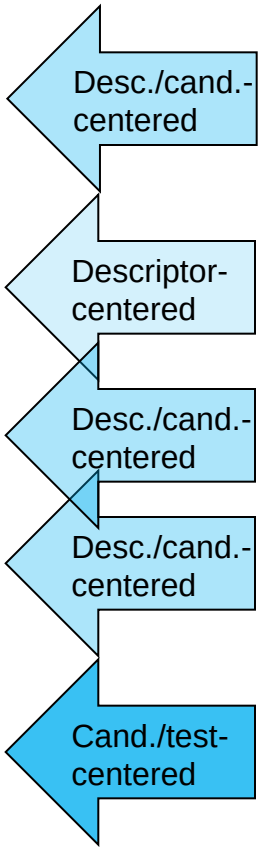
Common features:

- multi-level examinations (most studies considered 2 levels, MET: 3 levels)
- comparable number of participants (between 12 and 21)

Differing features:

- maximum number of points

Source	*Test*	*Method*	*Target candidate definition method*
Tannenbaum/ Wylie (2004)	TOEFL Structures	Angoff probabilities (0.1, 0.2, ... 0.9)	Homework: read global scale, write down key characteristics of level (not MCP) At workshop: Summarize key descriptors and produce panel-agreed version for reference during workshop
	TOEFL Reading		
	TOEFL Listening		
	TOEIC Listening		
	TOEIC Reading		
Papageorgiou (2010)	Michigan English Test, Listening	Modified Angoff (100 borderline candidates)	„Atomized" descriptors, choose right level
	Michigan English Test, Grammar+Reading		
telc	DTZ Version 1 – Version 5	Modified Angoff (100 MCPs)	Sort descriptors, complete descriptor puzzle, discuss MCP
	Deutsch Medizin B2-C1	Modified Angoff (yes/no)	Mark key characteristics in scale, discuss MCP
	English B1-B2	Modified Angoff (100 MCPs)	Mark key characteristics in scale, discuss Consider target candidate definitions from Tannenbaum/Wylie (2008) Look at MCPs' work and describe their ability



Common features:

- multi-level examinations (most studies considered 2 levels, MET: 3 levels)
- comparable number of participants (between 12 and 21)

Differing features:

- maximum number of points

Basis for comparison:

First round of judgements (reflects what participants learned from the familiarisation/target candidate definition exercise, but not the discussion afterwards)

Lowest of the levels (sometimes not enough room at the top)

Parameters: Level of disagreement (to address question 2): Range of cut scores, SE of judgements

In order to be able to compare these different studies, the cut scores were transformed into percentages of the maximum possible number of points

points

MET Listening	60 Items					
	Round 1			Round 2		
	B1	B2	C1	B1	B2	C1
R1	19,3	33,7	47,65	21,1	35,6	48,5
R2	13,6	31,4	46,8	21,5	39,1	55,7
R3	12,5	27,9	44,9	14,2	30,5	47,8
R4	26,4	40,58	51,18	31	41,4	48,8
R5	23,8	46,2	54	13	36,5	49,5
R6	11,7	35,7	52,7	12,3	33,3	51,3
R7	10,5	41,86	55,39	15,6	41,8	56,3
R8	11,5	33	51,81	12,3	34,8	49
R9	10,2	34,9	45,98	10,2	34,9	46
R10	19,1	25,75	29,1	24,1	31,2	35,7
R11	12,1	33,1	46,9	9,85	29,4	41,9
R12	17,0	43,15	54,75	24,9	43,7	54,9

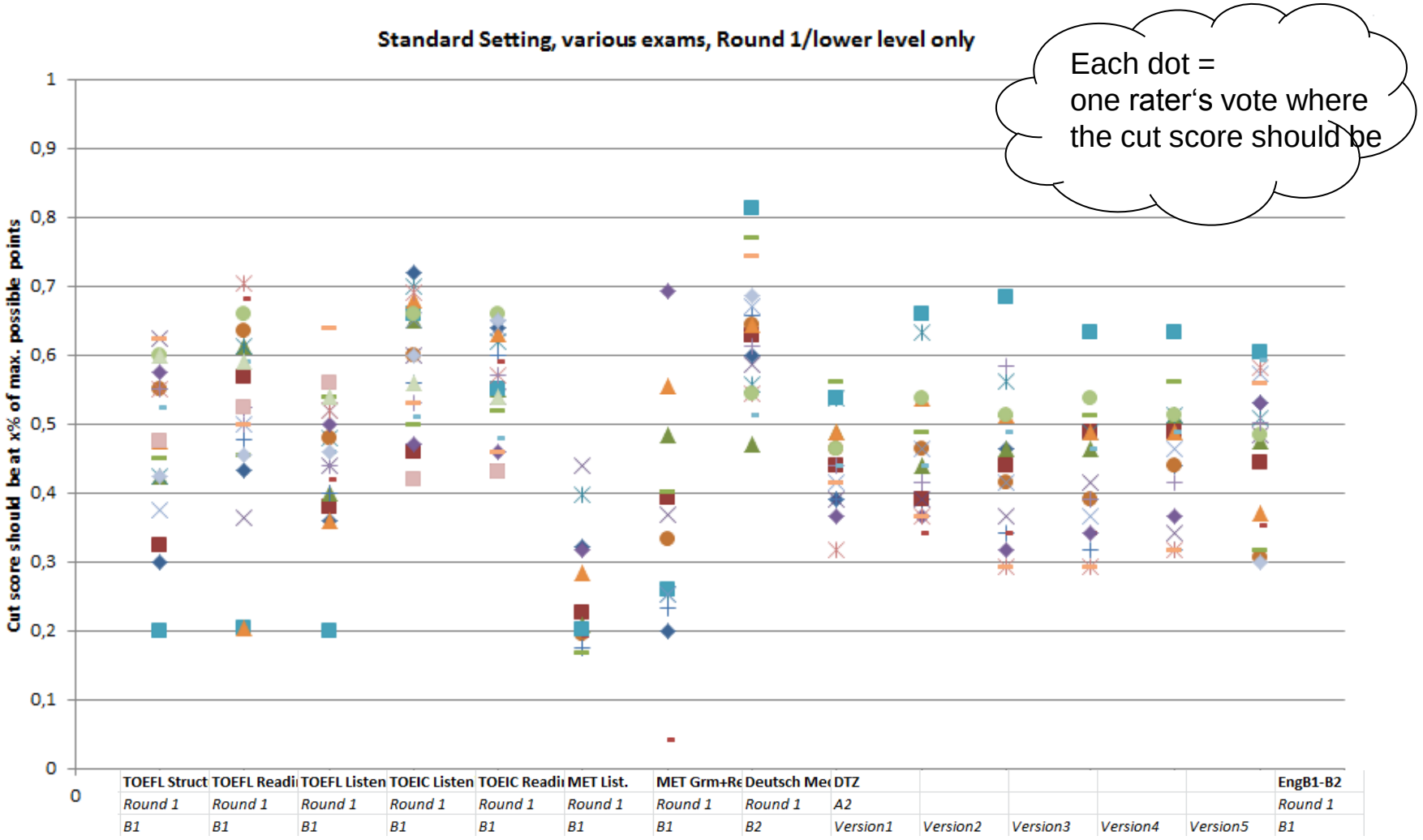
percentages

	In Prozent vom Maximalergebnis					
	Round 1			Round 2		
	B1	B2	C1	B1	B2	C1
R1	0,321	0,562	0,794	0,352	0,593	0,808
R2	0,226	0,523	0,780	0,358	0,652	0,928
R3	0,208	0,465	0,748	0,237	0,508	0,797
R4	0,440	0,676	0,853	0,517	0,690	0,813
R5	0,397	0,770	0,900	0,217	0,608	0,825
R6	0,195	0,595	0,878	0,205	0,555	0,855
R7	0,175	0,698	0,923	0,260	0,697	0,938
R8	0,192	0,550	0,864	0,205	0,580	0,817
R9	0,169	0,582	0,766	0,170	0,582	0,767
R10	0,318	0,429	0,485	0,402	0,520	0,595
R11	0,201	0,552	0,782	0,164	0,490	0,698
R12	0,283	0,719	0,913	0,415	0,728	0,915

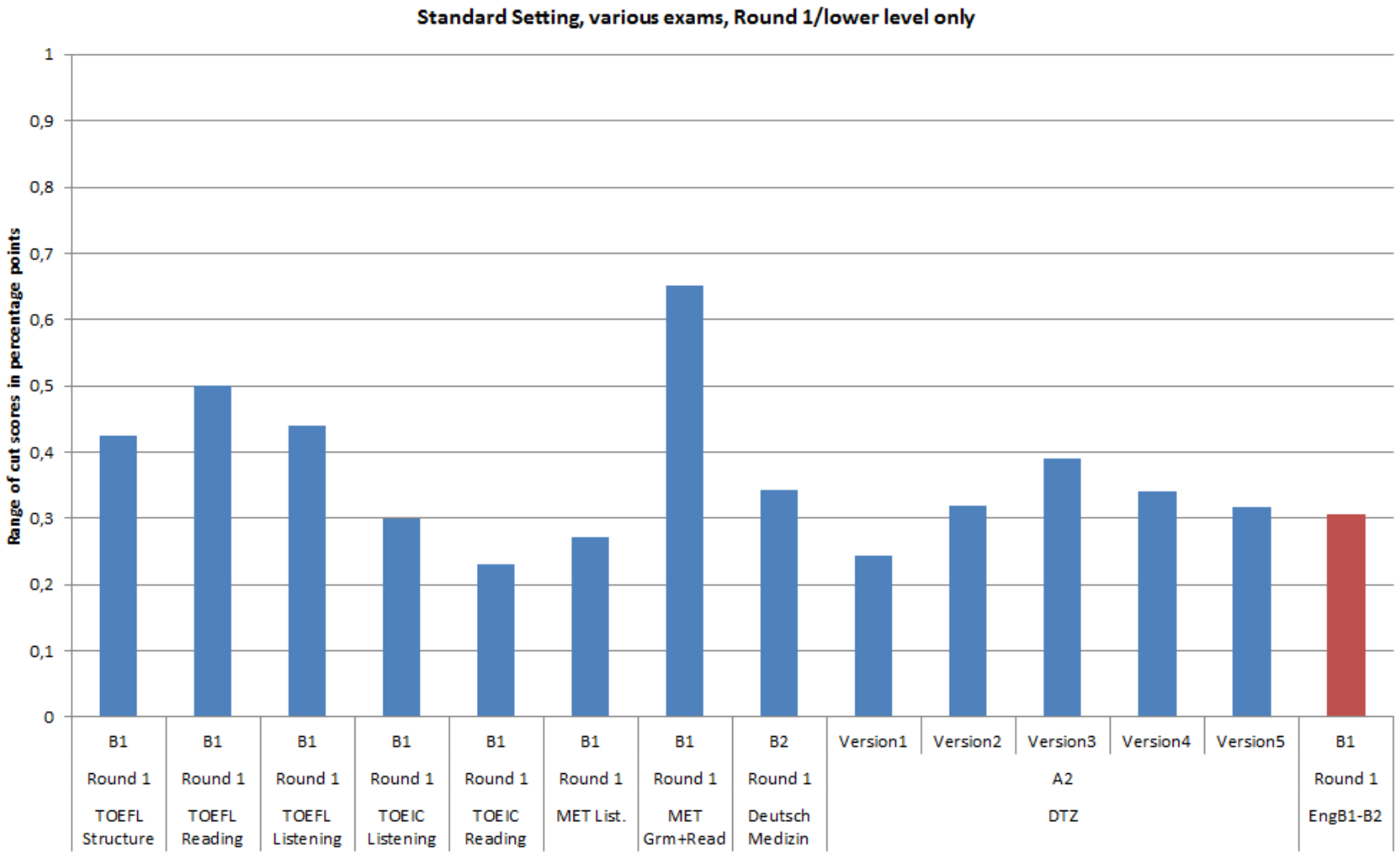
Comparison range, SEj, Round 1, lowest level

	TOEFL Structures	TOEFL Reading	TOEFL Listening	TOEIC Listening	TOEIC Reading	Michigan English Test, Listening	Michigan English Test, Grammar +Reading	DTZ Version 1	DTZ Version 2	DTZ Version 3	DTZ Version 4	DTZ Version 5	Deutsch Medizin B2-C1	English B1-B2
Range	0,425	0,500	0,440	0,300	0,230	0,271	0,652	0,244	0,318	0,390	0,341	0,317	0,343	0,306
SE j	0,0242	0,0288	0,0201	0,0185	0,0154	0,0251	0,045	0,0146	0,0209	0,0247	0,0213	0,0216	0,0213	0,0234

Comparison „range“, Round 1, lowest level



Comparison „range“, Round 1, lowest level



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

Eckes, Th. (2012), Examinee-centered standard setting for large-scale assessments: The prototype group method, Psychological Test and Assessment Modeling, Volume 54, Number 3, 257 - 283

Giraud, G., Impara, J. C., Plake, B. (2005), Teachers' Conceptions of the Target Examinee in Angoff Standard Setting, Applied Measurement in Education (18), 223–232

Figueras, N. Kaftandjieva, F., Takala, S. (2013), Relating a Reading Comprehension Test to the CEFR Levels: A Case of Standard Setting in Practice with Focus on Judges and Items, The Canadian Modern Language Review / La revue canadienne des langues vivantes, Volume 69, Number 4, 359-385

Freedle, R., Kostin, I. (1993), The Prediction of TOEFL Reading Comprehension Item Difficulty of Expository Prose Passages for Three Item Types: Main Idea, Inference, and Supporting Idea Items, ETS-RR-93-13, Educational Testing Service, Princeton, NJ

Kaftandjieva, F. (2004), Reference Supplement to the Preliminary Pilot version of the Manual for Relating Language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment.Section B: Standard Setting, Council of Europe, Strasbourg

Kostin, I. (2004), Exploring Item Characteristics That Are Related to the Difficulty of TOEFL Dialogue Items, ETS Research Reports RR-04-11, Educational Testing Service, Princeton, NJ

Papageorgiou, S. (2010), Setting Cut Scores on the Common European Framework of Reference for the Michigan English Test, Testing and Certification Division, English Language Institute, University of Michigan, Ann Arbor

Tannenbaum, R. J., Wylie, E. C. (2004), Mapping Test Scores onto the Common European Framework: Setting Standards of Language Proficiency on the Test of English as a Foreign Language (TOEFL), the Test of Spoken English (TSE), the Test of Written English (TWE), and the Test of English for International Communication (TOEIC), Educational Testing Service, Princeton, NJ

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 (RR-08-34), Educational Testing Service, Princeton, NJ

Wylie, E. C., Tannenbaum, R. J. (2006), TOEFL Academic Speaking Test: Setting a Cut Score for International Teaching Assistants (RM-06-01), Educational Testing Service, Princeton, NJ

Thank you!

Beate Zeidler
b.zeidler@telc.net