Standard setting of CEFR levels on Check S2 and Check S3 English and French

Cito

September 2016

Table of contents

1. Introduction

Introduction

In this document the standard setting of CEFR-levels on items from Check S2 and Check S3 for English and French will be reported. This standard setting took place on June 23 and June 24 in Arnhem, the Netherlands. For both language groups there was an international panel, (almost) the same experts per language were involved in setting the standards for reading and listening. Panel members play an important role in a standard setting procedure: a heterogeneous panel composition increases the validity of the standard setting results as different viewpoints of different shareholders are addressed while determining the standards. The second reason is that the standard setting outcomes are more likely to be accepted by all the stakeholders with many stakeholders involved in the standard setting procedures. Both panels consisted of Swiss and Dutch panel members.

For each language and both skills a selection of items to be included in the standard setting was made based on CEFR (the process of specification). At the start of the standard setting there were sessions concerning the familiarisation of all panel members; do they have more or less the same concept of a certain CEFR level for a specific skill? Presentations and exercises were instruments that helped in this process. Then the standard setting was performed on four different constructs: English reading (day 1), English listening (day 2), French reading (day 1), and French listening (day 2).

This report starts with a description of the expert panels. Then the implementation of the standard setting procedure (i.e., the 3DC method) will be described. Finally the advised standards, results, and analysis for each skill are presented. Rater scores are included in Appendix A of the report and advice that was given by experts regarding future item construction for French is included in Appendix B.

Wobbe Zijlstra

Anneke de Graaf

Janny Harmsen

Expert panel

The expert panel for English consisted of 15 experts. The experts were invited on the basis of their expertise on English and the CEFR levels. From this expert panel, 4 experts were English test specialist from Cito, The Netherlands. Six experts were English test specialists and/or CEFR specialist from varying affiliations in The Netherlands. Five experts came from varying affiliations in Switzerland, with expertise on tests, CEFR levels and the Swiss educational system. On both days, two experts were unable to attend. The standards for English listening and reading are therefore based on 13 experts.

The expert panel for French consisted of 13 experts. The experts were invited on the basis of their expertise on French and the CEFR levels. From this expert panel, 1 experts was French test specialist from Cito, The Netherlands. Eight experts were French test specialists and/or CEFR specialist from varying affiliations in The Netherlands. Six experts came from varying affiliations in Switzerland, with expertise on tests, CEFR levels and the Swiss educational system.

The composition of the expert panel is presented in Table 1.

Table 1: Expert panel composition

English

Expert	Country
Evelyn Reichard (Chair)	The Netherlands (Cito)
Rick Godschalk	The Netherlands (Cito)
Margreet van Aken	The Netherlands (Cito)
Marion Boxum	The Netherlands (Cito)
Erna Gille	The Netherlands
Nienke Smit	The Netherlands
Adriënne Buschmann	The Netherlands
Yvette Kroesen	The Netherlands
Ton Koet	The Netherlands
Frederike Westera	The Netherlands
Brigitte Ruhstaller	Switzerland
Bettina Coppens	Switzerland
Reto Hugenberg	Switzerland
Beat Petermann	Switzerland
Sabine Franke-Giancola	Switzerland
Prof. Dr. Stefan Keller	Switzerland

French

Expert	Country
Alma van Til (Chair)	The Netherlands (Cito)
Jan van Thiel	The Netherlands
Carla van de Pol-Eygenraam	The Netherlands
Dorine Smulders	The Netherlands
Marléne Muderwa	The Netherlands
Sylvie Ploemmen	The Netherlands
Yoeri Courtin	The Netherlands
Trees Aler	The Netherlands
Jérôme Paul	The Netherlands
Judith Schäube	Switzerland
Matthias Frey	Switzerland
Marta Oliveira	Switzerland
Daniela Dias	Switzerland
Giuseppe Manno	Switzerland

2. Procedure

The standard setting procedure that was used is called the '3DC method'. The designation '3DC' stands for *Data-Driven Direct Consensus*. In this method, it is assumed that a test consists of multiple items that can be divided into clusters. The experts are asked to indicate the scores that students would be expected to achieve in each cluster if they were exactly on the borderline of the selected CEFR-level. In this standard setting procedure, the CEFR levels that were set were A1/A2 and A2/B1. The experts were asked to indicate the score that a student would be expected to achieve if he/she would be exactly on the border of A1-A2 and A2-B1 respectively.

The items from the English listening and reading test are each divided into five clusters with the number of items ranged from 11-13 items (see table 2). All clusters were used for both CEFR levels A1/A2 and A2/B1. For French, six clusters were used with the number of items ranging from 9-12. Clusters 1-4 were used for A1/A2 and clusters 3-6 were used for A2/B1. Clusters 3 and 4 were used for both CEFR levels.

Table 2: number of items per cluster used for 3DC

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
English reading	13	12	12	12	12	-
English listening	12	12	11	11	11	-
French reading	11	11	12	11	12	11
French listening	9	9	9	9	10	10

The 3DC method used item response theory (IRT) parameters, which came from a fitting an IRT model to the Check S2 and Check S3 data. The selected items had adequate fit to the IRT model and most suitable items were selected (by the panel leaders) for the standards to be set.

In theory, it is sufficient to only present the clusters to the experts. However, providing the experts with empirical information will give an indication on how the different clusters are related to each other empirically. The assessment of the clusters will then be more realistic and contain fewer inconsistencies. Figure 1 shows how the empirical information was presented to the experts for the English reading test. This test will be used as an example.

The lines in the figure 1 represent the score scales associated with each cluster. Each item correct corresponds to one point. The maximum score is equal to the number of items in that cluster. The scoring scale ("Score") for the full test is displayed on the horizontal axis. As shown in the figure, a student answering none of the items correctly would achieve a score of 0. A student answering all of the items correctly would achieve a score of 13 + 12 + 12 + 12 + 12 = 61. If a student's score on the first cluster is 8, the student would be expected to achieve a score of 33 on the full test. This prediction can of course also be done in reverse. If a student's score on the full test is 33, we would expect the student to have the following profile: (cluster 1) score 8, (cluster 2) score 8, (cluster 3) score 5, (cluster 4) score 6, and (cluster 5) score 5/6 (NB. Due to rounding small difference occur). This figure informs the experts how the clusters are related to the full test, which enables them to consider this information in their assessments.

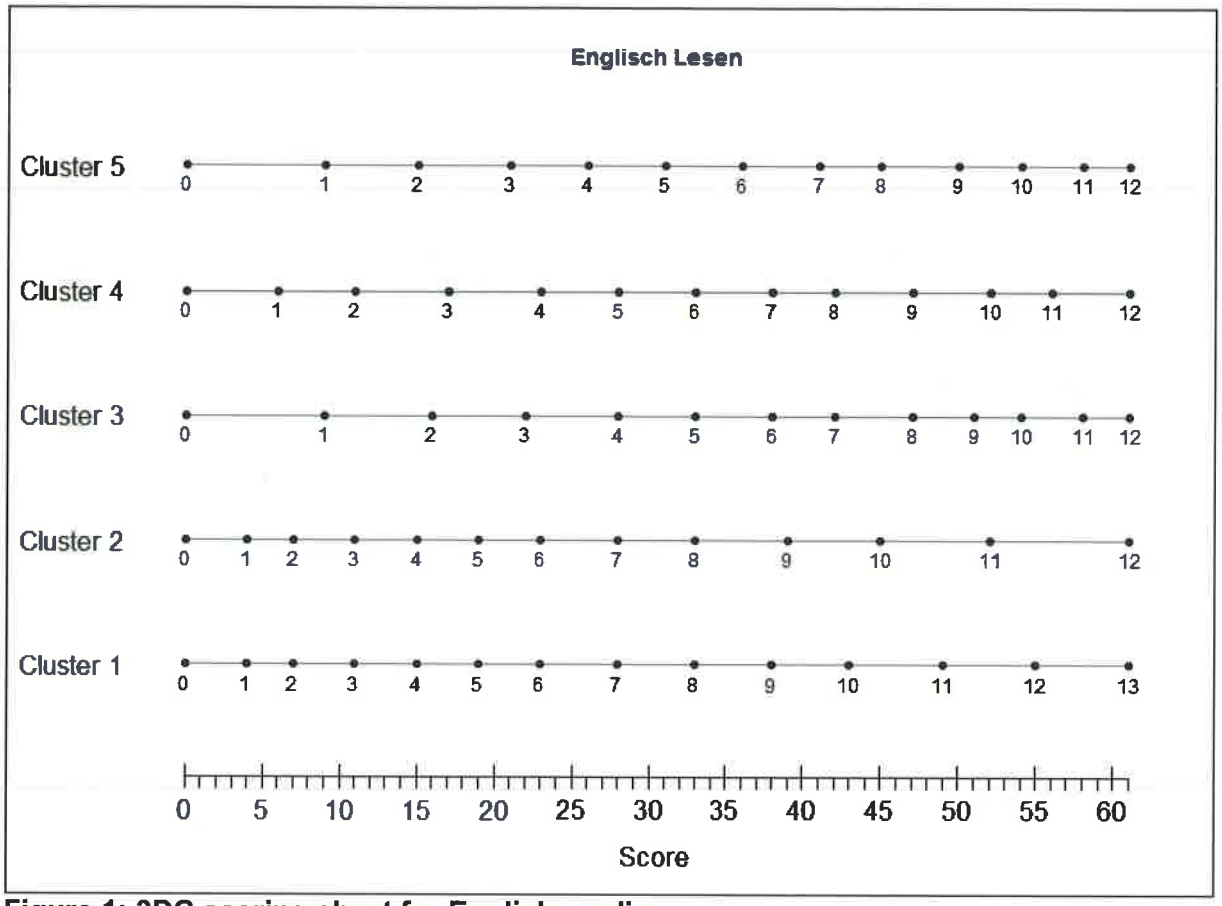

Figure 1: 3DC scoring sheet for English reading

During the standard setting, the standard was established in two assessment rounds. The same figure was used in both rounds. The following question was proposed to the experts: 'Which score would a student be expected to achieve on this cluster if his/her ability is exactly on the borderline of A1/A2 (or A2/B1)?' In the first round, the borderline candidate's score for each cluster in the figure was selected in the computer program by clicking on the score. After the first round, the results were discussed. To start the discussion, the results of the first assessment round were projected. In the second assessment round, the experts were again asked to mark the cut score for each cluster, enabling them to change their initial cut score after the discussion. After the second round, the expected percentage of students in the target population that would pass the resulting standard was presented. The average scores of the second round are presented to the Institute für Bildungsevaluation (see Appendix A for all rater scores). The advised standard for the CEFR-levels is the average cut score rounded up to the first integer.

The inter-rater agreement was calculated for both CEFR-levels. The inter-rater agreement is expressed by the Gower's coefficient, which is a measure to express absolute differences between judgments. This coefficient can take on any value between 0 and 1, where 0 indicates a complete lack of agreement between experts and 1 indicates complete agreement between experts.

3. Results and analysis of the standard setting

3.1 Standard setting English reading

The standard setting procedure resulted in the following advice regarding the standard for English reading:

A1/A2: a score of 32

A2/B1: a score of 46

on the complete set of 61 items

During the procedure, 15 experts estimated how many items within a cluster will be answered correctly by a student who is exactly on the border of A1/A2 and A2/B1 (separately). Summing the average cut score (over all experts) per cluster resulted in a cut score for the full test of 31.7 for A1/A2 and 45.1 for A2/B1 (see table 3). The average cut scores was similar for the Swiss and Dutch raters for A1/A2, whereas for A2/B1, the average cut score of the Swiss raters was about 1 point higher. The range of individual expert cut scores (sum of cluster scores) and the standard deviation was larger for A1/A2 than for A2/B1, indicating there was more consensus about the average cut score of A2/B1. Both inter-rater agreement were good (>= 0.90).

Figure 2 shows the average clusters scores. The long red (left) and blue (right) line corresponds to the advised cut scores for A1/A2 and A2/B1 respectively. Based on the student population Check S2, 72% would pass A1/A2 and 33% would also pass A2/B1. Table 4 shows the percentage that would pass a level for each cut score.

Table 5 shows the effects on the average cut score when removing one rater and/or one cluster at a time. For example, in the first cell rater 1 and cluster 1 are removed (32.5). For A1/A2, the lowest average cut scores was 30.0 (rater 14, cluster 5) and the highest was 33.6 (rater 13, cluster 2). In general, removing clusters 1 or 2 results in an higher average cut score (green) and clusters 3 or 5 results in a lower cut score. For A2/B1, the lowest average cut scores was 43.7 (rater 3, cluster 3) and the highest was 46.4 (rater 15, cluster 1). Removal of a rater resulted in small changes in the average cut score.

Table 3: Summary of standard setting results English reading

	A1/A2	A2/B1
Average cut score (SD)	31.7 (4.0)	45.1 (2.3)
Swiss raters (N = 6)	31.5 (4.3)	45.8 (2.3)
Dutch raters (N = 9)	31.8 (4.0)	44.6 (2.2)
Minimum	24	41
Maximum	39	49
% passing*	72%	33%
Inter-rater agreement	0.90	0.92

* based on Check S2 population and advised cut score

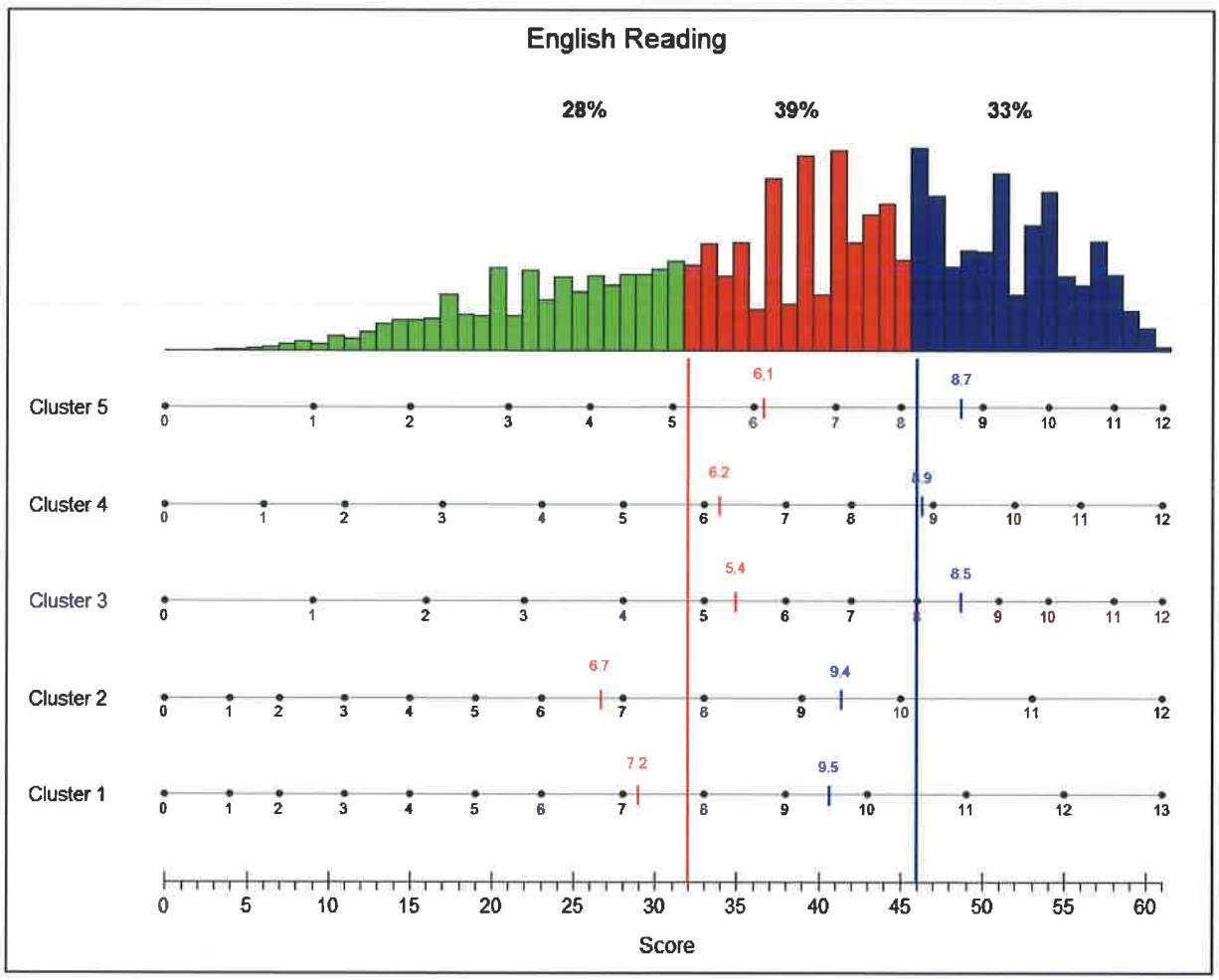

Figure 2: Average A1/A2 (red, left) and A2/B1 (blue, right) cluster scores over the raters where the line indicates the overall average cut score; and the distribution of Check S2 expected score on the 61 items English reading

Table 4: Percentage of students in the Check S2 population who will pass a certain CEFR-level English reading

Score	Percentage pass		Score	Percentage pass
0	100	**A1**	31	74
1	100	**A2**	32	72
2	100		33	70
3	100		34	68
4	100		35	65
5	100		36	64
6	100		37	61
7	100		38	59
8	100		39	55
9	99		40	53
10	99		41	49
11	99		42	47
12	98		43	42
13	98		44	41
14	97	**A2**	45	37
15	97	**B1**	46	33
16	96		47	31
17	95		48	27
18	94		49	25
19	93		50	21
20	92		51	19
21	91		52	17
22	89		53	13
23	88		54	12
24	85		55	9
25	85		56	6
26	82		57	3
27	81		58	3
28	79		59	2
29	78		60	1
30	76		61	0

Table 5: Average cut scores when removing rater and cluster English reading

A1/A2	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	No cluster removal
Rater 1 (D)	32.5	32.8	30.7	30.9	30.2	31.4
Rater 2 (S)	32.4	32.9	30.9	31.2	30.4	31.6
Rater 3 (S)	32.1	32.9	30.6	30.9	30.2	31.4
Rater 4 (S)	32.8	33.1	31.1	31.5	30.8	31.9
Rater 5 (D)	32.5	32.9	30.9	31.2	30.6	31.6
Rater 6 (D)	32.5	33.1	31.1	31.2	30.6	31.7
Rater 7 (D)	32.5	32.9	30.7	31.2	30.5	31.6
Rater 8 (D)	32.7	33.2	31.1	31.2	30.7	31.8
Rater 9 (D)	33.0	33.5	31.4	31.8	31.0	32.1
Rater 10 (S)	32.4	32.9	30.9	31.1	30.5	31.6
Rater 11 (D)	32.4	32.9	31.0	31.1	30.5	31.6
Rater 12 (S)	32.4	32.9	30.9	31.0	30.3	31.5
Rater 13 (S)	33.1	33.6	31.5	31.7	31.1	32.2
Rater 14 (D)	32.0	32.6	30.5	30.7	30.0	31.1
Rater 15 (D)	32.9	33.4	31.2	31.4	30.8	31.9
No rater removal	32.5	33.0	31.0	31.2	30.5	**31.7**

A2/B1	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	No cluster removal
Rater 1 (D)	46.1	45.8	44.0	44.7	44.1	45.0
Rater 2 (S)	46.0	45.8	44.1	44.7	44.1	45.0
Rater 3 (S)	45.8	45.5	43.7	44.5	44.0	44.8
Rater 4 (S)	46.0	45.7	43.9	44.8	44.2	45.0
Rater 5 (D)	45.9	45.6	43.8	44.6	44.0	44.9
Rater 6 (D)	46.0	45.8	44.0	44.7	44.2	45.0
Rater 7 (D)	46.4	45.9	44.0	45.0	44.4	45.2
Rater 8 (D)	46.4	46.0	44.3	44.9	44.4	45.3
Rater 9 (D)	46.1	45.8	44.1	44.8	44.2	45.1
Rater 10 (S)	46.0	45.6	43.9	44.6	44.2	44.9
Rater 11 (D)	45.9	45.7	44.0	44.8	44.2	45.0
Rater 12 (S)	46.4	46.0	44.3	44.9	44.3	45.3
Rater 13 (S)	46.2	45.9	44.0	44.8	44.2	45.1
Rater 14 (D)	46.2	46.0	44.1	44.9	44.2	45.1
Rater 15 (D)	46.4	46.1	44.3	45.0	44.5	45.4
No rater removal	46.1	45.8	44.0	44.8	44.2	**45.1**

Note: averages have been computed via the common standardized theta scale; green indicates higher average cut score and red lower average cut score; (D)utch and S(wiss).

3.2 Standard setting English listening

The standard setting procedure resulted in the following advice regarding the standard for English listening:

A1/A2: a score of 36

A2/B1: a score of 49

on the complete set of 56 items

During the procedure, 13 experts estimated how many items within a cluster will be answered correctly by a student who is exactly on the border of A1/A2 and A2/B1 (separately). Summing the average cut score (over all experts) per cluster resulted in a cut score for the full test of 35.6 for A1/A2 and 49.2 for A2/B1 (see table 6). The average cut scores was similar for the Swiss and Dutch raters for A1/A2, whereas for A2/B1, the average cut score of the Swiss raters was about 2 points higher. Both inter-rater agreement were good (>= 0.90).

Figure 3 shows the average clusters scores. The long red (left) and blue (right) line corresponds to the advised cut scores for A1/A2 and A2/B1 respectively. Based on the student population Check S2, 80% would pass A1/A2 and 32% would also pass A2/B1. Table 7 shows the percentage of students that would pass a level for each cut score.

Table 8 shows the effects on the average cut score when removing one rater and/or one cluster at a time. For A1/A2, the lowest average cut scores was 34.3 (raters 6 and 9, cluster 2) and the highest was 36.4 (rater 12, cluster 3). In general, removing clusters 1 or 3 results in an higher average cut score (green) and clusters 2 or 5 results in a lower score. For A2/B1, the lowest average cut scores was 46.4 (rater 7, cluster 2) and the highest was 49.3 (rater 12, cluster 1), and removing clusters 1 or 5 results in an higher average cut score (green) and clusters 2 or 4 results in a lower cut score. Removal of a rater resulted in small changes in the average cut score.

Table 6: Summary of standard setting results English listening

	A1/A2	A2/B1
Average cut score (SD)	35.6 (2.8)	49.2 (3.1)
Swiss raters (N = 5)	35.6 (3.3)	49.6 (2.1)
Dutch raters (N = 8)	35.6 (2.7)	47.3 (3.4)
Minimum	33	41
Maximum	41	53
% passing*	80%	32%
Inter-rater agreement	0.91	0.92

* based on Check S2 population and advised cut score

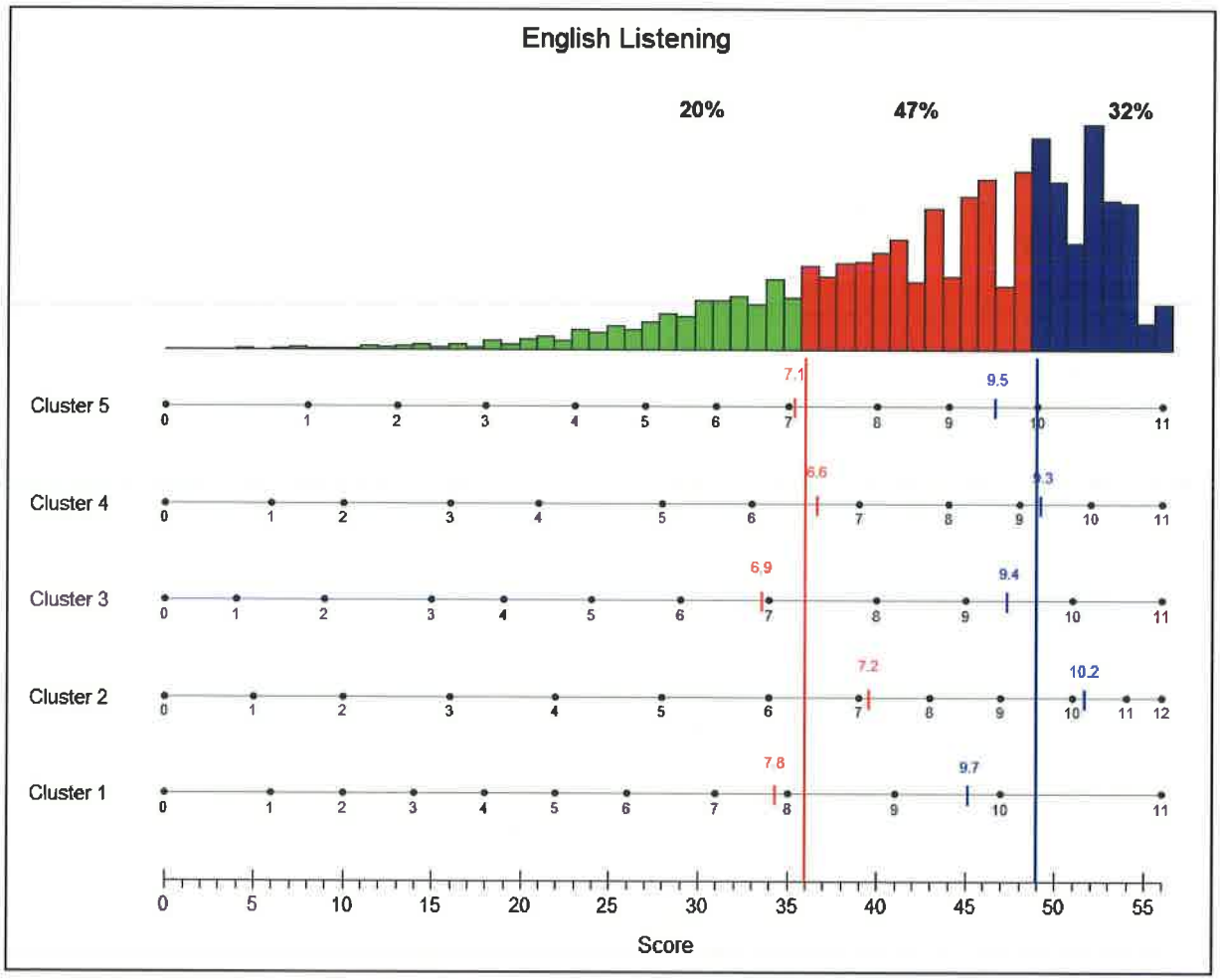

Figure 3: Average A1/A2 (red, left) and A2/B1 (blue, right) cluster scores over the raters where the line indicates the overall average cut score; and the distribution of Check S2 expected score on the 56 items English listening

Table 7: Percentage of students in the Check S2 population who will pass a certain CEFR-level English listening

Score	Percentage pass		Score	Percentage pass
0	100		29	92
1	100		30	90
2	100		31	89
3	100		32	88
4	100		33	86
5	100		34	84
6	100	A1	35	82
7	100	A2	36	80
8	100		37	77
9	100		38	74
10	100		39	72
11	100		40	68
12	100		41	66
13	100		42	62
14	99		43	58
15	99		44	56
16	99		45	51
17	99		46	48
18	99		47	43
19	99	A2	48	36
20	98	B1	49	32
21	98		50	27
22	98		51	21
23	97		52	14
24	97		53	12
25	96		54	6
26	95		55	1
27	94		56	0
28	93			

Table 8: Average cut scores when removing rater and cluster English listening

A1/A2	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	No cluster removal
Rater 1 (D)	36.0	35.0	36.2	35.6	35.9	35.8
Rater 2 (S)	36.3	34.9	36.3	35.6	36.0	35.8
Rater 3 (S)	36.2	35.0	36.3	35.7	35.9	35.8
Rater 4 (S)	36.0	34.7	36.2	35.5	35.5	35.6
Rater 5 (D)	36.2	35.0	36.3	35.6	36.0	35.8
Rater 6 (D)	35.7	34.3	35.8	35.0	35.5	35.3
Rater 7 (D)	36.2	34.9	36.1	35.3	35.8	35.7
Rater 8 (D)	36.0	34.6	36.0	35.3	35.6	35.5
Rater 9 (S)	35.6	34.3	35.8	34.9	35.2	35.2
Rater 10 (D)	35.6	34.5	35.9	35.2	35.5	35.3
Rater 11 (S)	36.1	34.8	36.2	35.4	35.8	35.7
Rater 12 (D)	36.1	34.9	36.4	35.7	36.0	35.8
Rater 13 (D)	36.3	34.9	36.2	35.4	35.9	35.8
No rater removal	36.0	34.7	36.1	35.4	35.7	**35.6**

A2/B1	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	No cluster removal
Rater 1 (D)	48.7	46.9	48.3	47.8	48.6	48.2
Rater 2 (S)	48.7	46.7	48.3	47.7	48.5	48.1
Rater 3 (S)	48.8	46.8	48.4	47.9	48.4	48.2
Rater 4 (S)	48.6	46.6	48.2	47.7	48.4	48.0
Rater 5 (D)	49.0	47.1	48.5	47.9	48.7	48.3
Rater 6 (D)	48.8	46.8	48.3	47.8	48.6	48.2
Rater 7 (D)	48.4	46.4	47.9	47.4	48.1	47.8
Rater 8 (D)	48.7	46.7	48.3	47.7	48.5	48.1
Rater 9 (S)	48.7	46.8	48.3	47.9	48.6	48.2
Rater 10 (D)	48.8	47.1	48.5	48.1	48.7	48.3
Rater 11 (S)	48.4	46.5	47.8	47.3	48.2	47.8
Rater 12 (D)	49.3	47.4	49.0	48.4	49.1	48.8
Rater 13 (D)	48.9	47.1	48.4	47.8	48.6	48.3
No rater removal	48.8	46.8	48.3	47.8	48.6	**48.2**

Note: averages have been computed via the common standardized theta scale; green indicates higher average cut score and red lower average cut score; (D)utch and S(wiss).

3.3 Standard setting French reading

The standard setting procedure resulted in the following advice regarding the standard for French reading:

A1/A2: a score of 26

on the complete set of 45 items

A2/B1: a score of 33

on the complete set of 46 items

During the procedure, 13 experts estimated how many items within a cluster will be answered correctly by a student who is exactly on the border of A1/A2 and A2/B1 (separately). Summing the average cut score (over all experts) per cluster resulted in a cut score for the full test of 26.0 for A1/A2 and 32.4 for A2/B1 (see table 9). The average cut scores was similar for the Swiss and Dutch raters for A1/A2, whereas for A2/B1, the average cut score of the Swiss raters was about 1 point higher. The range of individual expert cut scores (sum of cluster scores) and the standard deviation was larger for A1/A2 than for A2/B1, indicating there was more consensus about the average cut score of A2/B1. Both inter-rater agreement were good (>= 0.88).

Figure 4 shows the average clusters scores. The long red (right) and blue (left) line corresponds to the advised cut scores for A1/A2 and A2/B1 respectively. Based on the student population Check S2, 38% would pass A1/A2 and 5% would also pass A2/B1. Table 10 shows the percentage that would pass a level for each cut score.

Table 11 shows the effects on the average cut score when removing one rater and/or one cluster at a time. For A1/A2, the lowest average cut scores was 25.3 (raters 5, cluster 3) and the highest was 27.0 (raters 2 and 9, cluster 2). For A2/B1, the lowest average cut scores was 30.9 (rater 4, cluster 3) and the highest was 33.6 (rater 2, cluster 6). Removing clusters 2 or 6 results in an higher average cut score (green) and cluster 3 results in a lower cut score. Removal of a rater resulted in small changes in the average cut score.

Table 9: Summary of standard setting results French reading

	A1/A2	A2/B1
Average cut score (SD)	26.0 (3.1)	32.4 (2.6)
Swiss raters (N = 5)	24.2 (3.7)	32.0 (4.1)
Dutch raters (N = 8)	27.1 (2.2)	33.0 (1.4)
Minimum	20	27
Maximum	31	37
% passing*	38%	5%
Inter-rater agreement	0.88	0.90

* based on Check S2 population and advised cut score

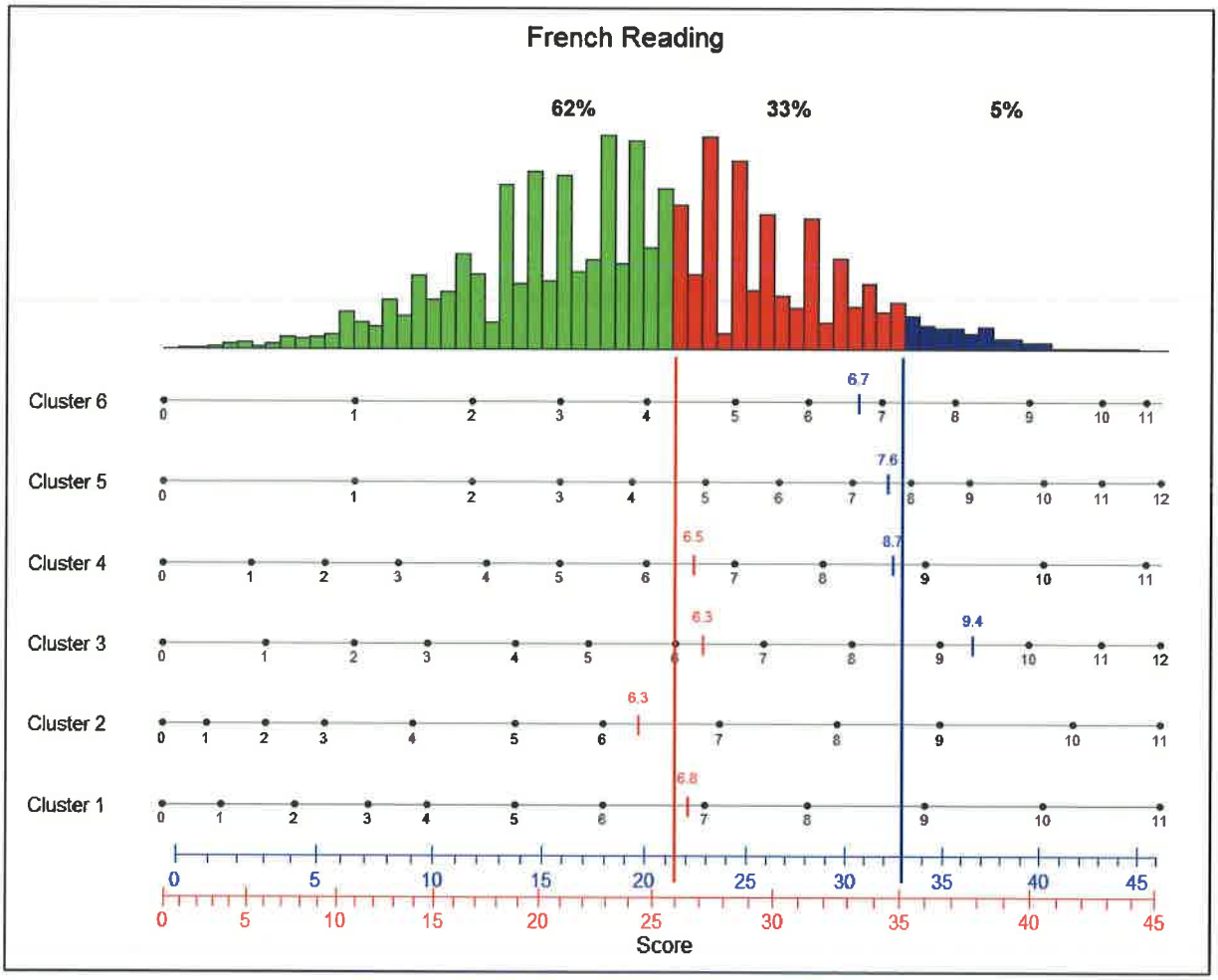

Figure 4: Average A1/A2 (red, left) and A2/B1 (blue, right) cluster scores over the raters where the line indicates the overall average cut score; and the distribution of Check S2 expected score on all 68 items. The red (lower) scale is for the 45 A1/A2 items and blue (upper) is for the 46 A2/B1 items French reading

Table 10: Percentage of students in the Check S2 population who will pass a certain CEFR-level French reading

A1/A2 Score	Percentage pass		A2/B1 Score	Percentage pass
0	100		0	100
1	100		1	100
2	100		2	100
3	100		3	99
4	100		4	99
5	100		5	99
6	99		6	98
7	99		7	96
8	99		8	96
9	99		9	93
10	98		10	90
11	97		11	88
12	96		12	84
13	95		13	81
14	93		14	78
15	90		15	72
16	88		16	70
17	85		17	63
18	81		18	57
19	78		19	54
20	72		20	47
21	70		21	42
22	63		22	37
23	57		23	32
24	53		24	29
25	46	**A1**	25	23
26	38	**A2**	26	21
27	35		27	18
28	29		28	14
29	23		29	13
30	19		30	10
31	17		31	8
32	13		**A2** 32	6
33	10		**B1** 33	5
34	7		34	4
35	5		35	3
36	3		36	2
37	2		37	1
38	2		38	1
39	1		39	0
40	0		40	0
41	0		41	0
42	0		42	0
43	0		43	0
44	0		44	0
45	0		45	0
			46	0

Table 11: Average cut scores when removing rater and cluster French reading

A1/A2	Cluster 1	Cluster 2	Cluster 3	Cluster 4	No cluster removal
Rater 1 (D)	26.1	26.5	25.8	25.9	26.1
Rater 2 (S)	26.2	27.0	26.4	26.1	26.4
Rater 3 (D)	26.0	26.5	25.8	26.0	26.1
Rater 4 (S)	26.0	26.6	25.7	25.6	26.0
Rater 5 (D)	25.9	26.4	25.6	25.8	25.9
Rater 6 (D)	26.0	26.5	25.4	25.6	25.9
Rater 7 (S)	25.7	26.0	25.7	25.5	25.8
Rater 8 (D)	25.5	26.0	25.3	25.4	25.6
Rater 9 (S)	26.5	27.0	26.2	26.2	26.5
Rater 10 (S)	26.1	26.5	25.7	26.0	26.1
Rater 11 (D)	25.5	26.1	25.6	25.8	25.8
Rater 12 (D)	25.9	26.2	25.4	25.8	25.8
Rater 13 (D)	26.1	26.5	25.7	26.0	26.1
No rater removal	25.9	26.5	25.7	25.8	**26.0**

A2/B1	Cluster 3	Cluster 4	Cluster 5	Cluster 6	No cluster removal
Rater 1 (D)	31.3	32.2	32.7	33.0	32.3
Rater 2 (S)	31.8	32.6	33.1	33.6	32.8
Rater 3 (D)	31.4	32.5	32.8	33.2	32.5
Rater 4 (S)	30.9	31.9	32.4	32.7	32.0
Rater 5 (D)	31.5	32.7	32.9	33.4	32.7
Rater 6 (D)	31.4	32.2	32.7	33.3	32.4
Rater 7 (S)	31.0	32.1	32.4	32.7	32.1
Rater 8 (D)	31.2	32.3	32.6	33.2	32.3
Rater 9 (S)	31.1	32.2	32.6	33.0	32.3
Rater 10 (S)	31.4	32.4	32.9	33.2	32.5
Rater 11 (D)	31.3	32.4	32.7	32.8	32.3
Rater 12 (D)	31.3	32.4	32.6	32.9	32.3
Rater 13 (D)	31.3	32.4	32.8	33.0	32.4
No rater removal	31.3	32.4	32.7	33.1	**32.4**

Note: averages have been computed via the common standardized theta scale; green indicates higher average cut score and red lower average cut score; (D)utch and S(wiss).

3.4 Standard setting French listening

The standard setting procedure resulted in the following advice regarding the standard for French listening:

A1/A2: a score of 26

on the complete set of 36 items

A2/B1: a score of 26

on the complete set of 38 items

During the procedure, 13 experts estimated how many items within a cluster will be answered correctly by a student who is exactly on the border of A1/A2 and A2/B1 (separately). Summing the average cut score (over all experts) per cluster resulted in a cut score for the full test of 25.4 for A1/A2 and 25.8 for A2/B1 (see table 12). The average cut scores was about 2 points higher for the Dutch raters than the Swiss for A1/A2, whereas for A2/B1, the average cut score of the Swiss raters was about 1 point higher. Both inter-rater agreement were good (>= 0.88).

Figure 5 shows the average clusters scores. The long red (left) and blue (right) line corresponds to the advised cut scores for A1/A2 and A2/B1 respectively. Based on the student population Check S2, 23% would pass A1/A2 and 4% would also pass A2/B1. Table 13 shows the percentage that would pass a level for each cut score.

Table 14 shows the effects on the average cut score when removing one rater and/or one cluster at a time. For A1/A2, the lowest average cut scores was 24.0 (raters 8 and 10, cluster 4) and the highest was 26.6 (raters 2, cluster 3). For A2/B1, the lowest average cut scores was 24.0 (rater 4, cluster 4) and the highest was 27.4 (raters 11 and 13, cluster 5). Removing clusters 3 or 5 results in an higher average cut score (green) and cluster 4 results in a lower cut score. Removal of a rater resulted in small changes in the average cut score.

Table 12: Summary of standard setting results French listening

	A1/A2	A2/B1
Average cut score (SD)	25.4 (2.91)	25.8 (2.6)
Swiss raters (N = 5)	24.0 (3.5)	26.4 (2.2)
Dutch raters (N = 8)	26.3 (2.3)	25.4 (2.1)
Minimum	21	21
Maximum	30	30
% passing*	23%	4%
Inter-rater agreement	0.88	0.90

* based on Check S2 population and advised cut score

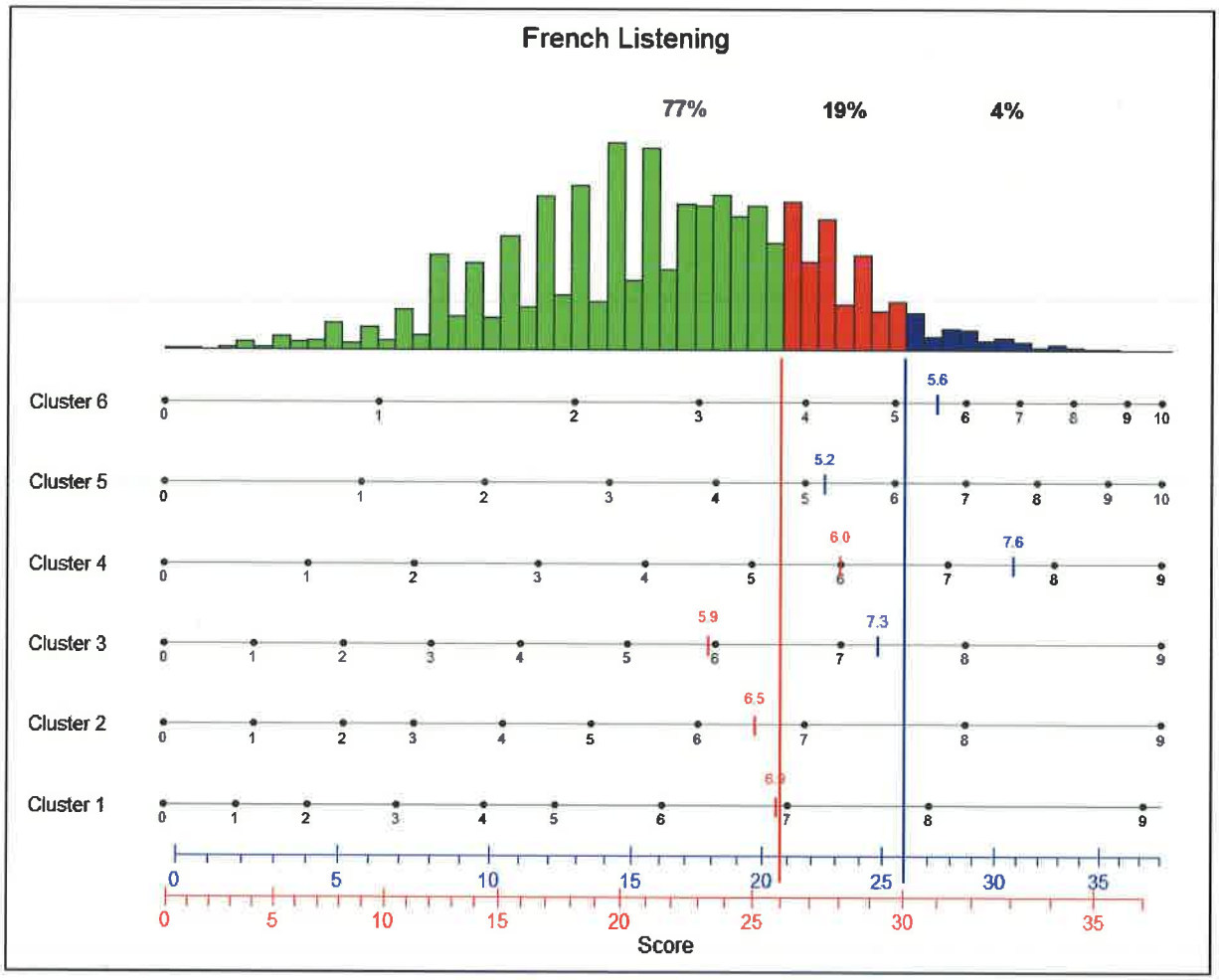

Figure 5: Average A1/A2 (red, left) and A2/B1 (blue, right) cluster scores over the raters where the line indicates the overall average cut score; and the distribution of Check S2 expected score on all 68 items. The red (lower) scale is for the 45 A1/A2 items and blue (upper) is for the 36 A2/B1 items French listening

Table 13: Percentage of students in the Check S2 population who will pass a certain CEFR-level French listening

A1/A2 Score	Percentage pass			A2/B1 Score	Percentage pass
0	100			0	100
1	100			1	100
2	100			2	100
3	100			3	99
4	100			4	99
5	99			5	98
6	99			6	97
7	99			7	95
8	98			8	93
9	97			9	90
10	97			10	88
11	95			11	83
12	93			12	77
13	92			13	70
14	88			14	66
15	86			15	60
16	83			16	52
17	77			17	46
18	72			18	40
19	69			19	32
20	62			20	27
21	56			21	20
22	50			22	16
23	42			23	11
24	36			24	8
25	28	**A1**	**A2**	25	6
26	23	**A2**	**B1**	26	4
27	16			27	3
28	11			28	2
29	7			29	1
30	4			30	1
31	3			31	1
32	1			32	0
33	1			33	0
34	0			34	0
35	0			35	0
36	0			36	0
				37	0
				38	0

Table 14: Average cut scores when removing rater and cluster French listening

A1/A2	Cluster 1	Cluster 2	Cluster 3	Cluster 4	No cluster removal
Rater 1 (D)	25.3	25.5	26.3	24.3	25.4
Rater 2 (S)	25.7	25.9	26.6	24.7	25.8
Rater 3 (D)	25.3	25.5	26.1	24.1	25.3
Rater 4 (S)	25.5	25.7	26.3	24.3	25.5
Rater 5 (D)	25.5	25.8	26.0	24.4	25.5
Rater 6 (D)	25.0	25.1	25.7	24.0	25.0
Rater 7 (S)	25.7	25.9	26.4	24.6	25.7
Rater 8 (D)	25.1	25.2	25.8	24.0	25.1
Rater 9 (S)	25.5	25.8	26.3	24.6	25.6
Rater 10 (S)	25.0	25.1	25.7	24.0	25.0
Rater 11 (D)	25.5	25.7	26.0	24.2	25.4
Rater 12 (D)	25.3	25.4	25.9	24.1	25.3
Rater 13 (D)	25.4	25.8	26.1	24.4	25.5
No rater removal	25.4	25.6	26.1	24.3	**25.4**

A2/B1	Cluster 3	Cluster 4	Cluster 5	Cluster 6	No cluster removal
Rater 1 (D)	26.1	24.5	27.2	25.1	25.8
Rater 2 (S)	26.1	24.5	27.3	25.3	25.8
Rater 3 (D)	26.0	24.5	27.3	25.1	25.8
Rater 4 (S)	25.7	24.0	27.0	24.9	25.4
Rater 5 (D)	26.1	24.6	27.3	25.2	25.8
Rater 6 (D)	26.1	24.5	27.2	25.1	25.8
Rater 7 (S)	25.9	24.4	27.2	25.1	25.7
Rater 8 (D)	25.8	24.1	27.2	25.1	25.6
Rater 9 (S)	26.1	24.5	27.3	25.3	25.8
Rater 10 (S)	26.2	24.6	27.2	25.2	25.8
Rater 11 (D)	26.5	24.8	27.4	25.8	26.2
Rater 12 (D)	26.0	24.4	27.2	25.0	25.7
Rater 13 (D)	26.1	24.7	27.4	25.3	25.9
No rater removal	26.1	24.4	27.3	25.2	**25.8**

Note: averages have been computed via the common standardized theta scale; green indicates higher average cut score and red lower average cut score; (D)utch and S(wiss).

Appendix A: Rater scores

Table A1: Rater scores for English reading

A1/A2	cluster1	cluster2	cluster3	cluster4	cluster5	sum
rater1	10	7	6	6	6	35
rater2	7	7	6	7	6	33
rater3	7	9	6	7	7	36
rater4	7	5	4	7	6	29
rater5	7	6	5	7	7	32
rater6	6	7	6	6	6	31
rater7	8	7	4	7	7	33
rater8	7	7	5	5	6	30
rater9	6	5	4	6	4	25
rater10	7	7	6	6	7	33
rater11	7	6	7	6	7	33
rater12	8	7	7	6	6	34
rater13	6	5	4	4	5	24
rater14	8	9	7	8	7	39
rater15	7	7	4	5	5	28
Average	7.2	6.7	5.4	6.2	6.1	31.7

A2/B1	cluster1	cluster2	cluster3	cluster4	cluster5	sum
rater1	10	10	9	9	8	46
rater2	9	10	10	9	8	46
rater3	10	10	9	10	10	49
rater4	9	9	8	10	10	46
rater5	10	10	9	10	9	48
rater6	9	10	9	9	9	46
rater7	11	8	6	9	9	43
rater8	10	9	8	7	8	42
rater9	9	9	9	9	9	45
rater10	10	9	9	9	10	47
rater11	8	9	9	10	10	46
rater12	10	9	9	7	7	42
rater13	10	10	8	9	8	45
rater14	9	10	8	9	8	44
rater15	9	9	8	7	8	41
Average	9.5	9.4	8.5	8.9	8.7	45.1

Table A2: Rater scores for English listening

A1/A2	cluster1	cluster2	cluster3	cluster4	cluster5	sum
rater1	6	8	6	7	7	34
rater2	8	6	6	6	7	33
rater3	7	7	6	7	6	33
rater4	8	7	8	8	5	36
rater5	7	7	6	6	7	33
rater6	9	7	8	7	9	40
rater7	9	8	6	5	7	35
rater8	9	7	7	7	7	37
rater9	9	8	9	7	8	41
rater10	7	8	8	8	8	39
rater11	8	7	7	6	7	35
rater12	6	6	7	7	7	33
rater13	9	7	6	5	7	34
Average	7.8	7.2	6.9	6.6	7.1	35.6

A2/B1	cluster1	cluster2	cluster3	cluster4	cluster5	sum
rater1	9	11	9	9	10	48
rater2	10	10	10	9	10	49
rater3	10	10	10	10	8	48
rater4	10	10	10	10	10	50
rater5	10	10	9	8	9	46
rater6	10	10	9	9	10	48
rater7	11	11	10	11	10	53
rater8	10	10	10	9	10	49
rater9	9	10	9	10	10	48
rater10	8	10	9	10	9	46
rater11	11	12	9	10	11	53
rater12	8	8	9	8	8	41
rater13	10	11	9	8	9	47
Average	9.7	10.2	9.4	9.3	9.5	48.2

Table A3: Rater scores for French reading

A1/A2	cluster1	cluster2	cluster3	cluster4	sum
rater1	7	6	6	6	25
rater2	4	6	7	4	21
rater3	6	6	6	7	25
rater4	7	8	6	5	26
rater5	7	7	6	7	27
rater6	8	8	5	6	27
rater7	8	5	9	7	29
rater8	8	7	8	8	31
rater9	6	5	5	4	20
rater10	7	6	5	7	25
rater11	6	6	8	9	29
rater12	8	6	6	8	28
rater13	7	6	5	7	25
Average	6.8	6.3	6.3	6.5	26.0

A2/B1	cluster3	cluster4	cluster5	cluster6	sum
rater1	10	8	8	7	33
rater2	9	6	6	6	27
rater3	9	9	7	6	31
rater4	10	9	10	8	37
rater5	8	9	6	6	29
rater6	10	7	7	8	32
rater7	10	10	9	7	36
rater8	9	9	7	8	33
rater9	9	9	8	8	34
rater10	9	8	8	6	31
rater11	10	10	8	5	33
rater12	10	10	7	6	33
rater13	9	9	8	6	32
Average	9.4	8.7	7.6	6.7	32.4

Table A4: Rater scores for French listening

A1/A2	cluster1	cluster2	cluster3	cluster4	sum
rater1	6	6	7	6	25
rater2	5	5	6	5	21
rater3	7	7	7	5	26
rater4	7	6	6	5	24
rater5	7	7	4	6	24
rater6	8	7	7	8	30
rater7	6	6	5	5	22
rater8	8	7	7	7	29
rater9	6	6	5	6	23
rater10	8	7	7	8	30
rater11	8	7	5	5	25
rater12	8	7	6	6	27
rater13	6	7	5	6	24
Average	6.9	6.5	5.9	6.0	25.4

A2/B1	cluster3	cluster4	cluster5	cluster6	sum
rater1	8	8	5	5	26
rater2	7	7	5	6	25
rater3	7	8	6	5	26
rater4	8	8	7	7	30
rater5	7	8	5	5	25
rater6	8	8	5	5	26
rater7	7	8	6	6	27
rater8	7	7	7	7	28
rater9	7	7	5	6	25
rater10	8	8	4	5	25
rater11	7	6	2	6	21
rater12	8	8	6	5	27
rater13	6	8	5	5	24
Average	7.3	7.6	5.2	5.6	25.8

Appendix B: Evaluation

Evaluation de l´épreuve de compréhension écrite

1 Les textes sont présentés sans une mise en contexte (s'agit-il d'un article, d'une brochure, d'une lettre ?), ce qui complique inutilement la tâche pour les élèves.

2 La chronologie des questions ne respecte pas toujours la chronologie du texte, ce qui fait perdre du temps aux élèves.

3 Le vocabulaire des questions est parfois trop difficile pour un niveau A2 ou B1.

4 Si une question à choix multiples contient plusieurs bonnes réponses, il vaut mieux indiquer le nombre de bonnes réponses que l'élève doit identifier (2 sur 6 ou 3 sur 6, par exemple).

5 Certaines questions évaluent seulement la capacité des élèves de reconnaître des mots et ne vérifient pas s'ils ont compris le texte. Pour répondre à la question, il suffit d'identifier les mots du texte qui sont identiques à ceux de la bonne réponse. L'élève peut alors obtenir des points sans avoir compris ce qui est écrit. Or dans les descripteurs du CECR, il s'agit principalement de la compréhension de textes écrits.

6 La réponse à certaines questions peut être donnée sans avoir lu le texte. Il suffit parfois d'avoir une bonne culture générale pour pouvoir y répondre.

Evaluation de l´épreuve de compréhension orale

1 Les fragments sont présentés sans une mise en contexte (s'agit-il d'une émission à la radio, d'une discussion entre deux personnes, ou d'une conférence ?), ce qui complique inutilement la tâche pour les élèves.

2 Certains dialogues sont lus par une seule personne, ce qui prête à confusion. On ne sait plus quels propos attribuer à quelle personne.

3 Certains sujets ne correspondent pas à l'univers des élèves.

4 La qualité sonore de certains fragments laisse à désirer.

5 Certains monologues manquent d'authenticité.

6 Dans certains fragments sonores, le débit est beaucoup trop rapide pour un niveau A2 ou même B1.

7 Le vocabulaire des questions est parfois trop difficile pour un niveau A2 ou B1.

8 Certaines questions évaluent seulement la capacité des élèves de reconnaître des mots et ne vérifient pas s'ils ont compris le fragment sonore. Pour répondre à la question, il suffit d'identifier les mots du fragment qui sont identiques à ceux de la bonne réponse. L'élève peut alors obtenir des points sans avoir compris ce qui est dit. Or dans les descripteurs du CECR, il s'agit principalement de la compréhension de textes oraux.

9 Le fait de poser plusieurs questions à propos d'un même fragment sonore, n'est pas en accord avec le principe que chaque question doit être indépendante de l'autre (une condition pour pouvoir analyser l'épreuve avec des programmes comme OPLM). En plus, l'élève doit parfois chercher et réécouter le fragment pour pouvoir trouver la phrase qui permet de donner la bonne réponse, ce qui lui fait perdre du temps.

10 Les questions concernent parfois des petits détails sans importance particulière et obligent l'élève à réécouter une partie des fragments à plusieurs reprises. Avec pour conséquence que l'élève n'a pas assez de temps pour finir l'épreuve.

11 Certaines images utilisées dans les questions prêtent à confusion.

12 Pour certaines questions, même les locuteurs natifs étaient incapables de trouver la bonne réponse.