Benchmarking comprehension exams on the CEFR a posteriori: an exploratory experiment

Marie-Hélène Fries, Marie-Pierre Jouannaud, Marie Thevenon, Camille Biros

ASp
la revue du GERAS

74 | 2018
Diachronie et anglais de spécialité

Benchmarking comprehension exams on the CEFR *a posteriori*: an exploratory experiment

Marie-Pierre Jouannaud, Marie Thévenon, Camille Biros and Marie-Hélène Fries

OpenEdition Journals

Electronic version
URL: http://journals.openedition.org/asp/5507
ISSN: 2108-6354

Publisher
Groupe d'étude et de recherche en anglais de spécialité

Printed version
Date of publication: 1 November 2018
Number of pages: 163-172
ISSN: 1246-8185

Electronic reference
Marie-Pierre Jouannaud, Marie Thévenon, Camille Biros and Marie-Hélène Fries, « Benchmarking comprehension exams on the CEFR *a posteriori*: an exploratory experiment », *ASp* [Online], 74 | 2018, Online since 01 November 2018, connection on 16 January 2020. URL : http://journals.openedition.org/asp/5507

Benchmarking comprehension exams on the CEFR *a posteriori*: an exploratory experiment

Marie-Pierre Jouannaud, Marie Thévenon, Camille Biros and Marie-Hélène Fries

Comment s'assurer que des étudiants attestent d'un niveau B2 en langue de spécialité ? C'est à cette question que permet de répondre le compte rendu d'expérience pédagogique proposé par Marie-Pierre Jouannaud et ses collègues de l'Université Grenoble Alpes, en nous faisant part d'une démarche innovante dans la conception d'examens de compréhension écrite en anglais de spécialité, conçus pour être calibrés sur le niveau B2 du CECRL.

Si l'évaluation a pu être considérée jusque dans les années 1990 comme le parent pauvre de l'anglais de spécialité (Robinson 1991), elle a graduellement pris une place plus importante dans les préoccupations des enseignants en secteur LANSAD, faisant écho à l'intérêt croissant dont elle fait l'objet en didactique des langues en général, notamment depuis la publication du Cadre, où elle joue un rôle central. On sait qu'une éventuelle corrélation entre les descripteurs ainsi que les échelles d'évaluation généralistes du Cadre, d'une part, et la compétence en langue de spécialité d'autre part, demeure problématique (Petit 2006) : ce compte rendu prolonge de manière très concrète les réflexions déjà amorcées il y a presque dix ans dans cette même revue (Fries 2009) pour mettre en cohérence l'anglais de spécialité et le CECRL.

Les auteurs nous montrent ainsi comment une équipe pédagogique, en s'appropriant et en adaptant des outils existants (le Cadre, le Manuel, la certification IELTS, l'expérience du test SELF), a mis au point des examens de compréhension écrite en anglais de spécialité leur permettant d'identifier les étudiants ayant un niveau B2 et ceux qui ne l'ont pas atteint, tout en proposant une meilleure adéquation entre évaluation et contenus de cours en langue de spécialité. Les étapes nécessaires à l'élaboration de ce type d'examen pour en assurer la validité et la fiabilité sont détaillées par les auteurs, qui explicitent les choix réalisés et n'éludent pas les difficultés rencontrées (moyens techniques, définition des scores de césure, processus de calibrage).

Ce compte rendu propose donc une vision très concrète et pratique des processus mis en œuvre pour aboutir à un examen calibré sur l'un des niveaux du CECRL en compréhension écrite, et

offre de nombreuses réponses pour les enseignants qui s'interrogent sur la possibilité et la faisabilité d'une telle démarche, ouvrant la voie à une adaptation de cette approche à d'autres compétences. (Catherine Colin)

Introduction

1 Since the publication of the Common European Framework for Languages in 2001, followed by the Manual on Relating Language Examinations to the Common European Framework for Languages (henceforth the Manual) in 2009, all language teachers have theoretically been given the opportunity to map their students' results to the CEFR levels (A1 to C2). For English for Specific Purposes courses, however, the input of the CEFR has not been so easy to take into account, because most of the original CEFR descriptors were devised with a general, rather than specific, use of language in mind (Petit 2007). This situation has led many ESP teachers to go further than the CEFR and define their own competence criteria as a function of their students' future needs, both in their specialized fields and in international professional interactions (Braud et alii 2016). This move had led to the creation of specialized CEFR-based validation schemes targeting realistic language competences (Fries 2009; Millot 2017). It has also in some measure been reflected in the CEFR, with the addition of new descriptors in 2017, and has prompted the present study, which reports on making ESP reading comprehension exams coherent with the CEFR, through three different experiments with Master's students.

1. Context and objectives

2 In Grenoble, up to 2017, past examinations already benchmarked on the CEFR were used to assess reading skills for Master's students in science and technology and make sure they had reached the B2 level. This made English for Science and Technology course design less than coherent: although instructors were choosing their course material according to their graduate students' main field of study, final reading examinations were based on general academic English. To solve this perceived discrepancy between preparation and final exam, the English teaching team decided to create their own examinations, more relevant to their students' needs. Thanks to the experience gained during the local development of a CEFR-based placement test called SELF (Système d'Évaluation en Langues à Visée Formative/language assessment tool with formative aims, Cervini et alii 2013), we realized that the same standard setting methods used for the SELF placement test could also be implemented to benchmark specialized reading examinations after the students had taken them, using the methods advocated in the Manual. A posteriori standard setting procedures are essential for specialized examinations, as preliminary pilot runs are not easily feasible. The present report is an account of the experiments we have been carrying out since 2017 in order to check the feasibility of this idea.

2. Methods

3 A newly designed exam will rest on shaky foundations unless it is shown to be valid, i.e. to really test what it purports to test (in our case, B2 level reading ability in ESP

contexts). In order to do this, we need to link our exam to the CEFR, and justify that our cut score indeed corresponds to the frontier between B1 and B2 reading skills. This is called "standard setting." We were compelled to benchmark our examinations *a posteriori* because unlike commercial tests, university examinations cannot be pre-tested.

2.1. Standard setting principles

4 Standard setting procedures usually involve a combination of information from three sources: "intuitive, qualitative and quantitative" (see Council of Europe 2017 [2001]: 22, 207 about scale development). Test developers use their intuition (and experience) to choose texts and write items with a certain level in mind. The test is then administered to the students and the results are used to quantify item difficulty statistically. In the last, qualitative, phase, expert panelists jointly decide which items a minimally proficient candidate should be able to answer correctly, using the quantitative results, level descriptors and their experience. In many standard setting situations, there are no level descriptors to work from, and the panel members must first come to an agreement of what a "minimally proficient candidate" means in their context (Brandon 2004). In our case, however, the CEFR provides complete scaled grids with descriptors for each level and skill, plus sub-grids for more specific activities. All of our panel members are already very familiar with the CEFR descriptors, but a (re-)familiarization phase is still used at the beginning of each meeting, as per the advice of the *Manual*. Moreover, the descriptors are fairly general and still require extensive interpretation to be applied to actual item questions (Alderson 2007: 661).

2.2. Challenges

5 In our experiments, the main challenge concerned the quantitative phase. In order to give us easy access to all students' results for each item, all examinations had to be converted to a digital format. This meant having the students take the test online on a Learning Management System (the SIDES – Système Informatique Distribué d'Évaluation en Santé, i.e. distributed computer assessment system –platform, used in medical schools), or to score answer sheets automatically using the QCMP software (multiple choice questions on paper) on another LMS, the Moodle platform. In both cases, spreadsheets with individual students' results for each item are constructed automatically, downloaded from the platforms, and can then be used for further statistical analysis.

6 The other challenge was the convening of the standard setting panel. There is no shortage of expertise on site (the requirement is "thorough familiarity with the CEFR," Council of Europe 2009: 7), but it is difficult to organize a three-hour panel with a suitable number of experts present in a constraining time frame, after the exam has been administered, but before the results need to be handed in. Brandon, reviewing previous research on the topic, advises to "use at least 10" judges (2004: 68). As will be seen below, this is not easy to achieve. In order to make participation in the standard setting panel more rewarding, it was decided that the standard setting procedure should not exceed the time normally spent by instructors correcting papers (around 3 hours). The time saved by automatic correction is instead applied to a group discussion

of the items, and the validation of the cut scores. It is also important for a variety of stakeholders to take part in the standard setting process, to ensure that different interests are represented: course instructors, but also administrators and outside experts (Cizek & Bunch 2007: 225).

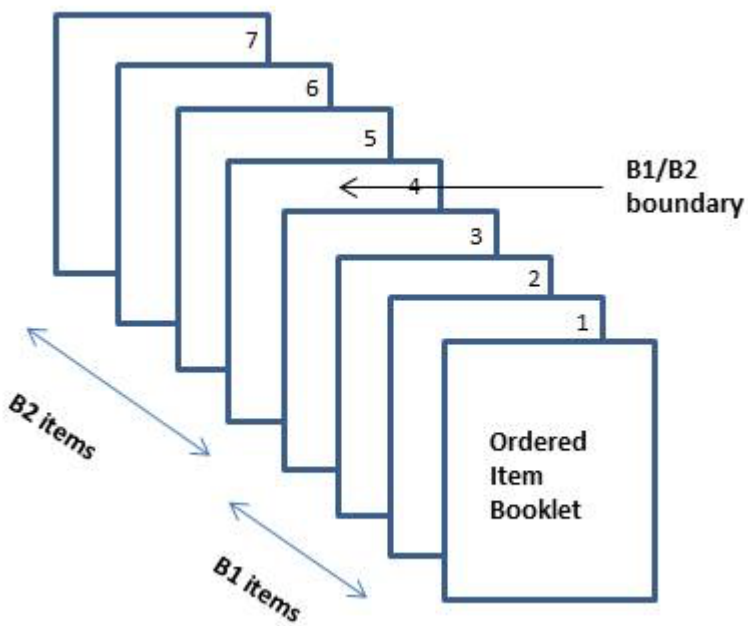
2.3. Standard setting procedures

7 We experimented with two standard setting procedures, the bookmark and the Angoff methods. In both cases, the procedure involves three rounds. In the first round, after a careful review of the test items, the panelists make individual decisions, and their results are entered into a spreadsheet so that everyone can see how much agreement or disagreement there is between judges. In the second round, panelists discuss their results in small groups to try to resolve some of their disagreements. The results are again presented to the whole group and the remaining disagreements are discussed during the third round. In case no consensus is achieved, the mean or median of cut score decisions is chosen as the final result. In between rounds, the panelists may receive additional information about the consequences of their decisions.

2.3.1. The bookmark method

8 In the bookmark method (Council of Europe 2009: 77), the empirical difficulty of each item for the students is taken into account from the start: test items are gathered in a booklet in ascending order of observed difficulty. The item with the highest percentage of right answers comes first, followed by the second easiest item, the third easiest, and so on. The panelists' task is to (metaphorically) put a bookmark between two pages in the booklet where they feel that the boundary between two levels is situated (an illustration of this principle is provided in Figure 1).

Figure 1: Illustration of the Ordered Item Booklet for the bookmark standard setting method (adapted from Mitzel et alii 2001)



9 The difficulty parameter of the boundary is then converted to an ability estimate, itself converted to a score on the test. This method is well suited to situations where several cut scores are needed, but it requires sophisticated statistics and the use of specialized software (in our case, Winsteps [Linacre 2017]) to calculate item and candidate parameters.

2.3.2. The Angoff method

10 The Angoff method (Council of Europe 2009: 63) is particularly suited to dichotomous items, i.e. multiple choice questions or similar items, where a right answer is worth one point and a wrong answer zero. The panelists are asked, for each test item, to determine the probability that a minimally proficient student will answer the question correctly (or, alternatively, to picture a group of one hundred minimally proficient students and determine how many of them would be likely to get the question right). The sum of one panelist's probabilities for all items is the provisional cut score for this panelist (for example, if a panelist believes that a minimally proficient student has a 50% chance of answering every item correctly, his/her cut score will be 10 if the exam is graded out of 20; an illustration of this principle is provided in Figure 2). The judges must be careful not to choose a probability that is lower than chance (so that a five-option MCQ question cannot get a probability estimate of less than 20%).

Figure 2: Illustration of the Angoff standard setting method: sample spreadsheet for determining the cut-off score (in our case the minimum percentage for the B2 level)

item nb	judge 1 (%)	judge 2 (%)	judge 3 (%)	judge 4 (%)	judge 5 (%)	judge 6 (%)	judge 7 (%)	judge 8 (%)	mean (%)
5	50	90	40	70	40	90	60	80	65
6	50	90	90	70	80	80	80	80	77.5
7	50	80	60	70	60	80	60	40	62.5
8	50	50	90	90	60	80	30	30	60
9	80	80	90	90	80	60	80	80	80
10	70	100	90	80	90	50	80	80	80
11	70	100	80	90	60	100	20	80	75
12	80	80	80	80	60	50	80	50	70
mean of means (cut-off percentage)									68.636

11 The quantitative information from test administration is used between rounds. Panelists are usually told what the empirical difficulty was for each item (i.e. how well the candidates did on each item), so that they can compare this information with their own estimates and perhaps tweak them accordingly.

3. Experiments and results

12 In this study, we report on three experiments carried out from December 2016 to March 2018 and involving between 100 and 500 students. All subjects were first-year

Master's students at Université Grenoble Alpes, majoring in a variety of science-related subjects (including sports or geography).

3.1. Pilot study

13 In December 2016, exploratory research was conducted to ascertain the feasibility of the standard setting procedure in our context (Biros *et alii* 2017). 155 students took a paper-based test and their results were entered into a spreadsheet by hand (a very time-consuming process which persuaded us further that automatic correction was necessary). For this pilot, the test chosen was a former IELTS paper, so that no specific validation argument was necessary. The goal was to see whether the panel members would arrive at the same cut score as the official IELTS cut score.

14 Using the bookmark method, the seven panelists arrived at the same cut scores as the official IELTS ones between B1 and B2 on the one hand (the main item of interest in our context), and B2 and C1 on the other. This encouraged us to expand our experiment to a locally designed test.

3.2. Experiment with the bookmark method

15 To write a test adapted to the specialized needs of our students, the first step is to have a clear understanding of their field of study. The students we worked with were first-year Master's students in health engineering specialising in biotechnology, medical chemistry, pharmaceutical engineering, medical physics, radiation protection, quality control, to name but a few. We needed to select texts that would reflect this pluridisciplinarity. We proceeded as for our class material and consulted the specialised teaching team to recommend scientific journals from which we ought to select and adapt texts. The themes included in this first test were the control of poliomyelitis risk in Europe, human animal hybrids and transgenic animals. The second phase was writing the questions. We were inspired by the IELTS question format (associations, true or false statements, table completions, short-answer type questions...) that seemed particularly suited to the electronic environment we were using as it only accepts selected response and short answer types. One of the main difficulties consisted in finding a good balance between getting rid of all ambiguity as to what was required in the answer, and keeping a sufficient degree of difficulty to make it a reading comprehension test. For instance, if only one of the possible answers is grammatically correct, you are not testing reading. Overall, we tried to include three to six questions for which only a very basic understanding of the text was necessary and about ten questions requiring close reading, good knowledge of the specialized vocabulary and analytical skills. Therefore, we kept the CEFR levels in mind when designing the test and tried to include a variety of levels in our different questions.

16 101 students took the test on tablets (normally, 150 is the minimum to give reliable results), with an average result of 12.6 (out of 20), and grades ranging from 4.5 to 18.5. The standard deviation was 3.06.

17 Eleven panelists took part in the bookmark standard setting procedure. After the presentation of the procedure (which was new to almost all the participants) and the familiarization phase, the individual and group rounds started. Everyone used the CEFR descriptors for reading extensively to place the bookmarks between levels, but many

felt frustrated at having to follow the order of difficulty observed during administration: it was felt that some items which the students had found relatively easy belonged to a higher CEFR level than others the students had found much more difficult. It should also be said that as in the IELTS tests, one of the difficulties is the time constraint: forty questions on three different texts need to be answered in one hour. As we soon found out, some of the questions at the end of the exam received a low rate of correct answers. This was probably due to their position in the exam rather than to their degree of difficulty. This problem has already been identified in the literature: some studies have "raised the issue of disordinality, whereby standard-setting participants disagree with the ordering of the OIB [Ordered Item Booklet]" (Karantonis & Sireci 2006: 8). There was so much discussion that only one round was completed before the three-hour mark (the second round started, but the meeting had to be adjourned due to lack of time). Even though no conclusive decision was reached, everyone agreed that the discussion had been very rich and rewarding.

18 It was thus decided to experiment with the simpler Angoff standard setting procedure next time instead, in which panelist decisions are not wedded to empirical difficulty (and because there was after all only one cut score to determine, between B1 and B2).

3.3. Experiment with the Angoff method

19 In order to have a large enough number of subjects, we designed a common examination for students majoring in the fields of biology, chemistry, geography, mechanics, physics and sports. The topics chosen were polystyrene and its impact on the environment (for biology and chemistry), physiological comfort in skiing garments (for sports), and harmonic absorbers in high-rise towers (for geography, mechanics and physics). As there were not enough tablets for this larger scale experiment (485 students took part), we decided to put the examination on the Moodle platform and use the QCMP software to scan the answer sheets, with the help of the pedagogical engineering team. We had to use a double format for this examination: an IELTS-type format for the students' papers and a QCM format for Moodle.

20 The mean was 12.7, with grades ranging from 3 to 19, and a standard deviation of 3.2. The reliability coefficient (Cronbach's alpha) for the exam was very high at .85.

21 Although eleven colleagues took part in the standard setting panel, only eight were able to stay throughout the three-hour session. After the first round, the cut score was 15, which meant that anybody scoring less than 15 would not be awarded B2 in reading. The panelists were then given what Cizek and Bunch (2007: 56) call "impact feedback," in this case the percentage of students who would "pass" B2 with such a high cut score (less than a third, representing only half as much as in previous years). After small group discussions (round 2), the cut score went down to 14, and after further discussion before the meeting had to be adjourned, the cut score was further lowered to 13. This was done by lowering the 100% probability estimates to 90%: even for very easy items, it is always possible to make a mistake, especially for speeded tests. This resulted in 57% of students passing. "Reality feedback" (Cizek & Bunch 2007: 55), i.e. observed difficulty during test administration, was not given to the participants due to lack of time. After the standard setting session, however, the correlation between mean difficulty estimates by panel participants and observed difficulty during administration

was calculated, and found to be fairly high at .6. The panelists were thus fairly adept at "guessing" how difficult each item would be for the pool of students.

22 In the future, we are planning to use the Angoff method and complement it with reality feedback in the form of students' success rate for each item.

Conclusion

23 We hope to have shown in this report that the reading descriptors of the CEFR can also be used to validate reading skills at the B2 level for ESP. This exploratory experiment can potentially be expanded to different languages and paves the way for validating receptive skills (listening and reading) in specific domains, for given levels of the CEFR. As a complement to assessment grids based on CEFR descriptors for productive skills (interacting, speaking and writing), it potentially enables LSP professionals to offer an institutional alternative to outside tests, in terms of skill and competence validation.

24 Calls for teacher empowerment are not new:

> I believe the teaching profession can make three contributions to the improvement of testing: they can write better tests themselves; they can enlighten other people who are involved in testing processes; and they can put pressure on professional testers and examining boards to improve *their* tests. (Hughes 2002: 5)

25 We hope that our experiments go some way toward fulfilling the first two of these objectives.

BIBLIOGRAPHY

ALDERSON, J. Charles. 2007. "The CEFR and the need for more research". *The Modern Language Journal* 91/4, 659–663, DOI: <10.1111/j.1540-4781.2007.00627_4.x>.

BIROS, Camille, Marie-Hélène FRIES, Marie-Pierre JOUANNAUD & Marie THÉVENON. 2017. "Benchmarking a specialised reading comprehension examination on the CEF *a posteriori*". Paper at the GERAS conference, University of Lyons 3, March 2017.

BRANDON, Paul R. 2004. "Conclusions about frequently studied modified Angoff standard-setting topics". *Applied Measurement in Education* 17/1, 59–88, DOI: <10.1207/s15324818ame1701_4>.

BRAUD, Valérie, Philippe MILLOT, Cédric SARRÉ, & Séverine WOZNIAK. 2016. "Quelles conceptions de la maîtrise de l'anglais en contexte professionnel ? Vers une définition de la 'compétence en anglais de spécialité'". *Mélanges CRAPEL* 37, 13–44, retrieved from <http://www.atilf.fr/IMG/pdf/art1.pdf> on 07/05/18.

CERVINI, Cristiana, Monica MASPERI, Marie-Pierre JOUANNAUD, & Francesca SCANU. 2013. "Defining, modeling and piloting SELF, a new formative assessment test for foreign languages". In COLPAERT J., M. SIMONS, A. AERTS & M. OBERHOFER (eds.), *Language Testing in Europe: Time for a new framework?* Antwerp: The University of Antwerp, 55–60.

CIZEK, Gregory J. & Michael B. BUNCH. 2007. Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests. Thousand Oaks, CA: Sage Publications.

Council of Europe. 2009. "Relating Language Examinations to the Common European framework for Languages: Learning, teaching, Assessment. A Manual", retrieved from <https://rm.coe.int/1680667a2d> on 07/05/18.

Council of Europe. 2017 [2001]. "Common European Framework of Reference for Languages: Learning, Teaching, Assessment", retrieved from <https://rm.coe.int/1680459f97> on 31/07/2017.

FRIES, Marie-Hélène. 2009. "Mise en cohérence de l'anglais de spécialité et du CECRL en France : difficultés et enjeux". *ASp* 56, 105–125, DOI: <10.4000/asp.177>.

HUGHES, Arthur. 2002. *Testing for Language Teachers*. Cambridge: Cambridge University Press.

KARANTONIS, Ana & Stephen G. SIRECI. 2006. "The bookmark standard-setting method: A literature review". *Educational Measurement: Issues and Practice* 25/1, 4–12, DOI: <10.1111/j.1745-3992.2006.00047.x>.

LINACRE, J.M. 2017. *Winsteps® Rasch Measurement Computer Program*. Beaverton, OR: Winsteps.com.

MILLOT, Philippe. 2017. "Spécialiser la compétence B2 en anglais dans le cadre d'une démarche qualité : Une proposition pour le secteur Lansad". *Recherche et pratiques pédagogiques en langues de spécialité. Cahiers de l'Apliut* 36/1, DOI: 10.4000/apliut.5593.

MITZEL, Howard C., Daniel M. LEWIS, Richard J. PATZ & Donald ROSS GREEN. 2001. "The bookmark procedure: Psychological perspectives". In CIZEK, G. J., *Setting Performance Standards: Concepts, methods, and perspectives*, 249–281. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

PETIT, Michel. 2006. "Les descripteurs du cadre : quelle conception de la langue de spécialité?". In Haramboure, F. *et alii* (dir), *Travaux des journées 2006 de l'EA2025*. Bordeaux: Université Victor Segalen Bordeaux 2, 14–29.

PETIT, Michel. 2007. "La correction linguistique dans le Cadre européen commun : quelle conception, quels critères ?". *Recherche et pratiques pédagogiques en langues de spécialité. Cahiers de l'Apliut* 26/2, 62–80, DOI: 10.4000/apliut.2003.

ROBINSON, Pauline 1991. *ESP Today: A Practitioner's Guide*. Hemel Hamstead: Prentice Hall International English Language Teaching.

INDEX

Mots-clés: CECRL, examens de compréhension, compte rendu d'expérience
Keywords: CEFR, comprehension exams, teaching practices

AUTHORS

MARIE-PIERRE JOUANNAUD

Marie-Pierre Jouannaud teaches courses in English grammar, phonetics and teaching methods at Université Grenoble Alpes. She is interested in foreign language acquisition and currently participates in two research projects, one on the development of an online placement test (Idefi

Innovalangues), and the other on the development of a listening comprehension game for elementary school (e-FRAN Fluence). <marie-pierre.jouannaud@univ-grenoble-alpes.fr>

MARIE THÉVENON

Marie Thévenon teaches courses in English for science and technology at Université Grenoble Alpes. She is currently working on discourse analysis and the use of science fiction as a teaching tool in ESP classes for science students. <Marie.Thevenon@univ-grenoble-alpes.fr>

CAMILLE BIROS

Camille Biros teaches English for Biotechnology and Health at Université Grenoble Alpes. Her research focuses on environmental discourse and scientific communication using the tools of discourse analysis and corpus linguistics. <camille.biros@univ-grenoble-alpes.fr>

MARIE-HÉLÈNE FRIES

Marie-Hélène Fries teaches English for science and technology at Université Grenoble Alpes. She is interested in the analysis of scientific discourse, with a theoretical focus on metaphors in science and technology and a practical concern for the specific language competences required for students in ESP. She is co-author of a textbook on oral presentation skills. <Marie-Helene.Fries@univ-grenoble-alpes.fr>