COUNCIL OF EUROPE CONSEIL DE L'EUROPE

Janvier 2009

Relier les examens de langues au Cadre européen commun de référence pour les langues : Apprendre, enseigner, évaluer (CECRL)

Un manuel

Division des Politiques Linguistiques, Strasbourg
www.coe.int/lang/fr

TRADUCTION :

Gilles BRETON, Sébastien GEORGES et Christine TAGLIANTE

SOMMAIRE

Liste des schémas iii

Liste des figures iii

Liste des Tableaux v

Liste des Fiches vii

Préface ix

Chapitre 1 Le CECRL et le manuel 1

Chapitre 2 Le processus de mise en relation 7

Chapitre 3 Familiarisation 19

Chapitre 4 Spécification 27

Chapitre 5 Formation à la standardisation et au calibrage 37

Chapitre 6 Procédures de définition des points de césure 60

Chapitre 7 Validation 90

Annexes 136

Annexe A Fiches et échelles pour la description et la spécification (ch. 1 et 4) 138

A1 : Caractéristiques principales des niveaux du CECRL (chapitre 1) 138
A2 : Fiches pour la description des examens (chapitre 4) 144
A3 : Spécification : activités langagières communicatives (chapitre 4) 150
A4 : Spécification : compétence langagière communicative (chapitre 4) 160
A5 : Spécifications : résultat des analyses (chapitre 4) 173

Annexe B Grilles d'analyse de contenu Chapitre 4 175

B1 : Grille d'analyse de contenu pour la réception orale et la réception écrite 175
B2 : Grille d'analyse de contenu pour les tâches de production orale et de production écrite 181

Annexe C Fiches et échelles pour la standardisation et le calibrage (Ch.5) 209

Références supplémentaires :

Section A: Résumé du processus de mise en relation

Section B: Définition des points de césure

Section C: Théorie classique des tests

Section D: Méthodes d'analyse qualitative

Section E: Théorie de la généralisabilité

Section F: Analyse factorielle

Section G: Théorie de la réponse à l'item

Section H: Mise en équivalence des tests

Liste des schémas

Schéma 2.1	Preuve de la validité de la mise en relation de l'examen ou du test avec le CECRL	8
Schéma 2.2	Représentation graphique des procédures permettant de relier les examens au CECRL	17

Liste des figures

Figure 6.1	Distribution de fréquence pour les scores aux tests dans les deux groupes contrastés	73
Figure 6.2	Régression logistique	80
Figure 6.3	Formulaire d'enregistrement des jugements des panélistes dans le cadre de la méthode du marque-page	86
Figure 6.4	Items de discrimination différente	90
Figure 6.5	Cartographie d'items, indiquant la difficulté et la discrimination	91
Figure 7.1	Courbe caractéristique empirique de l'item pour un item problématique	113
Figure 7.2	Une courbe caractéristique de test	117
Figure 7.3	Table de décision pour neuf niveaux	125
Figure 7.4	Table de décision pour cinq niveaux	126
Figure 7.5	Cartographie d'items avec des descripteurs de compétence	130

Liste des tableaux

Tableau 3.1	Gestion du temps pour les activités de Familiarisation	25
Tableau 3.2	Documents à préparer pour les activités de Familiarisation	25
Tableau 4.1	Fiches et échelles du CECRL pour les activités langagières communicatives	33
Tableau 4.2	Echelles du CECRL pour les aspects de la compétence langagière communicative	34
Tableau 5.1	Gestion du temps pour l'évaluation des échantillons de performance orale	48
Tableau 5.2	Gestion du temps pour l'évaluation des échantillons de performance écrite	50
Tableau 5.3	Documents et matériel à préparer pour l'évaluation de la production écrite	50
Tableau 5.4	Sources de références dans le CECRL	52
Tableau 5.5	Formation à la standardisation et au calibrage : récapitulatif	59
Tableau 6.1	Vue d'ensemble des méthodes discutées	65
Tableau 6.2	Données de base dans la méthode de Tucker-Angoff	66
Tableau 6.3	Calcul du score attendu pour les 100 candidats limites	71
Tableau 6.4	Distribution de fréquence correspondant à la figure 6.1.	73
Tableau 6.5	Tables de décision pour cinq scores de césure	74
Tableau 6.6	Résumé du tour de précision de l'étendue	77
Tableau 6.7	Résultats du réajustement	79
Tableau 6.8	Exemple de réponses dans la méthode de l'appariement au descripteur de l'item (formulaire abrégé)	82
Tableau 6.9	Marque-page et niveaux de réussite	88
Tableau 6.10	Estimation de la valeur Theta	89
Tableau 7.1	Dispositif de blocs incomplets équilibrés avec trois blocs	102
Tableau 7.2	Dispositif de blocs incomplets équilibrés avec sept blocs	103
Tableau 7.3	Exemple de consistance forte et de désaccord complet	109
Tableau 7.4	Tableau de fréquence pour quatre niveaux et deux panélistes	110
Tableau 7.5	Fréquence d'attribution des niveaux du CECRL pour un item	112
Tableau 7.6	Résumé des désaccords par item	113
Tableau 7.7	Résultat d'une procédure de Tucker-Angoff	114
Tableau 7.8	Décomposition de la variance	114
Tableau 7.9	Exactitude de la décision	119

Tableau 7.10	Consistance de la décision	120
Tableau 7.11	Distributions marginales sur les niveaux (occurrences)	123
Tableau 7.12	Distributions marginales sur les niveaux (pourcentages)	124
Tableau 7.13	Dispositif de procédure de détermination des scores de césure pour une paire	128
Tableau A1	Caractéristiques principales de l'interaction et de la production.	139
Tableau A2	Caractéristiques principales de la réception	140
Tableau A3	Eléments qualitatifs pertinents pour la réception	162
Tableau A4	Eléments qualitatifs pertinents pour l'interaction orale	169
Tableau A5	Eléments qualitatifs pertinents pour la production	170
Tableau C1	Echelle globale d'évaluation de la production orale	212
Tableau C2	Grille des critères d'évaluation de l'oral	213
Tableau C3	Grille des critères supplémentaires : « niveaux plus »	214
Tableau C4	Grille des critères d'évaluation de la production écrite	215

Liste des Fiches

Fiche A1	Description générale de l'examen	146
Fiche A2	Elaboration de l'examen	147
Fiche A3	Correction	148
Fiche A4	Notation	148
Fiche A5	Communication des résultats	148
Fiche A6	Analyse et révision de l'examen	149
Fiche A7	Justification des décisions	149
Fiche A8	Impression initiale du niveau global	149
Fiche A9	Réception orale	151
Fiche A10	Réception écrite	152
Fiche A11	Interaction orale	153
Fiche A12	Interaction écrite	155
Fiche A13	Production orale	156
Fiche A14	Production écrite	157
Fiche A15	Combinaisons de capacités intégrées	158
Fiche A16	Capacités intégrées	158
Fiche A17	Médiation orale	159
Fiche A18	Médiation écrite	160
Fiche A19	Aspects de la compétence langagière pour la réception	164
Fiche A20	Aspects de la compétence langagière en interaction	165
Fiche A21	Aspects de la compétence langagière en production	167
Fiche A22	Aspects de la compétence langagière en médiation	172
Fiche A23	Représentation graphique de la relation de l'examen aux niveaux du CECR	173
Fiche A24 :	Confirmation de l'estimation du niveau global de l'examen	173

Fiche C1	Fiche de rapport de formation	209
Fiche C2	Fiche analytique d'évaluation	210
Fiche C3	Fiche d'évaluation globale (DIALANG)	210
Fiche C4	Fiche de synthèse de l'évaluation globale (DIALANG)	211
Fiche C5	Fiche d'évaluation des items (DIALANG)	211

Ces fiches sont également disponibles sur le site Internet www.coe.int/lang/fr

Préface

Le Conseil de l'Europe tient à exprimer sa reconnaissance à tous ceux qui ont contribué à l'élaboration de ce manuel et notamment :

- Les autorités finlandaises qui ont organisé le séminaire d'Helsinki où le projet fut lancé en juillet 2002.

- Les consultants qui ont expérimenté l'édition pilote (Pr. Charles Alderson, Dr Gergely A. David, Dr John De Jong, Dr Felianka Kaftandjieva, Dr Michael Makosch, Dr Michael Milanovic, Dr Günther Nold, Professor Mats Oscarson, Prof. Günther Schneider, Dr Claude Springer et aussi M Josef Biro, Melle Erna van Hest, M Peter Lenz, Melle Jana Pernicová, Dr Vladimir Kondrat Shleg, Mme Christine Tagliante et Dr John Trim), pour leur retour d'information détaillé au tout début du projet.

- Le Groupe d'auteurs, sous la direction du Dr. Brian North :

• Dr Neus Figueras	- Departament d'Ensenyament, Généralité de Catalogne, Espagne
• Dr. Brian North	- Fondation Eurocentres, Suisse
• Prof. Sauli Takala	- Université de Jyväskyla, Finlande
• Dr. Piet Van Avermaet	- Centre pour la diversité et l'apprentissage, Université catholique de Louvain, Belgique - Association des centres d'évaluation en langue en Europe (ALTE)
• Dr.Norman Verhelst	- CITO, Pays Bas

- Dr Jay Banerjee (Université de Lancaster) et Dr Felianka Kaftandjieva (Université de Sofia) pour leur contribution à l'élaboration du *Supplément de référence* du Manuel.

- Les institutions qui ont mis à disposition des exemples illustrés de performances ainsi que des exemples d'items sur DVD/CD-ROM et disponibles sur le site du Conseil de l'Europe comme appui à la formation à la standardisation (en particulier : Eurocentres ; Cambridge ESOL ; le Centre international d'études pédagogiques –CIEP-, l'université pour Etrangers de Pérugia ; l'Institut-Goethe ; les autorités finlandaises ; DIALANG ; la Generalitat de Catalunya et CAPLE).

- ALTE (en particulier Nick Saville) et les membres du « groupe du projet CECRL néerlandais » (Charles Alderson, Neus Figueras, Dr Günther Nold, Henk Kuijper, Sauli Takala, Claire Tardieu) pour leur contribution à la « boîte à outils » qui concerne directement ce Manuel sous la forme de grilles d'analyse de contenu élaborées pour la production orale et la production écrite d'une part, la réception orale et la réception écrite d'autre part.

- Les nombreuses personnes et institutions qui ont fourni un retour d'informations détaillées sur la version expérimentale, en particulier : les membres de ALTE ; ASSET languages (Cambridge ESOL) ; l'école de commerce de Budapest ; le CITO ; Claudia Harsch ; l'institut Goethe ; le ministère de l'Education polonais ; le ministère de l'Education taïwanais ; le TestDaf ; le Trinity college de Londres et l'Université pour étrangers de Pérouse.

Division des Politiques linguistiques
Direction de l'Education et des langues (DG IV)
F – 67075 STRASBOURG

www.coe.int/lang/fr
www.coe.int/lang-CECR/fr

Chapitre 1 : Le CECRL et le Manuel

1.1. Les objectifs du manuel

1.2. Le contexte du manuel

1.1. Les objectifs du Manuel

Ce Manuel a pour principal objectif d'aider les concepteurs d'examens à élaborer des procédures transparentes et concrètes pour situer leurs examens par rapport au CECRL, à les appliquer et à en rendre compte dans un processus cumulatif de perfectionnement continu. Le manuel n'est pas le seul guide permettant de relier un test au CECRL et aucune institution n'est obligée d'entreprendre ce travail d'harmonisation. Cependant, les institutions qui affirment que leurs examens sont reliés aux niveaux du CECRL trouveront les procédures proposées très utiles pour démontrer la validité de leur affirmation.

L'approche développée dans le manuel propose des conseils aux utilisateurs pour :

- décrire ce que recouvre l'examen ainsi que les procédures de passation et d'analyse ;
- mettre en relation les résultats de l'examen et les Niveaux Communs de Référence du CECRL ;
- apporter des preuves rendant compte des procédures suivies.

Toutefois, suivant en cela les meilleures traditions de l'action du Conseil de l'Europe pour le développement de l'enseignement des langues, le Manuel vise plus largement à encourager fortement et à faciliter la coopération entre les organismes concernés et les spécialistes des pays membres. Le Manuel a pour objectif de :

- contribuer à l'élaboration d'une compétence dans le domaine de la relation des examens de langues avec le CECRL ;
- encourager une plus grande transparence de la part des organismes qui produisent des examens ;
- encourager la constitution de réseaux d'organismes et d'experts, officiels ou non, tant sur le plan national qu'international.

La division des politiques linguistiques recommande aux concepteurs d'examens utilisant les procédures proposées ou d'autres procédures visant les mêmes fins, de faire un relevé d'expérience sous forme de rapport. Ces rapports devraient décrire la mise en œuvre des procédures, les points positifs et les difficultés et s'ils affirment que l'examen est relié aux niveaux du CECR, en apporter des preuves. On encourage fortement les utilisateurs à rédiger ces rapports afin :

- d'accroître la transparence du contenu (justification théorique, objectifs de l'examen, etc…) ;
- d'accroître la transparence du niveau attendu de l'examen ;
- de donner aux candidats, aux utilisateurs et aux professionnels de l'enseignement et de l'évaluation l'occasion d'analyser la qualité de l'examen et de la relation affirmée avec le CECRL ;
- de procurer un argumentaire expliquant pourquoi certaines des procédures recommandées n'ont pas été suivies ;
- de procurer à de futurs chercheurs un ensemble élargi de techniques pouvant venir en complément de celles indiquées dans ce manuel.

Il faut souligner que, si ce Manuel recouvre un large éventail d'activités, son objectif est limité :

- C'est un guide tout particulièrement axé sur les procédures à mettre en œuvre pour justifier l'affirmation selon laquelle un examen ou un test donné est relié au CECR.
- Ce n'est pas un guide général pour l'élaboration de tests ou d'examens de langue de qualité. Il existe plusieurs guides utiles pour ce faire, comme cela est mentionné au chapitre 4 et ce sont ceux-là qu'il faut consulter.
- Il *ne* prescrit *pas* une approche particulière pour élaborer des tests ou des examens de langue. Si le CECRL milite en faveur d'une approche actionnelle de l'apprentissage des langues, il admet, dans son effort d'exhaustivité, que des examens différents puissent refléter des buts différents (construits).
- Il n'exige pas que les tests soient spécialement conçus pour évaluer des **performances** en relation avec le CECRL, mais une utilisation évidente du CECRL pendant le processus de formation, de types de tâches, de rédaction d'items et d 'élaboration de grilles d'évaluation renforce l'affirmation selon laquelle le contenu est relié au CECRL.
- Il ne fournit pas de label, ni de statut de validité ou d'accréditation selon laquelle tel ou tel examen est relié au CECRL. De telles affirmations relèvent de la responsabilité des institutions. Des associations de professionnels travaillent sur les standards et les codes de bonnes pratiques (par exemple l'AERA American Educational Research Association (AERA/APA/NCME : 1999 ; EALTA www.ealta.org ; ALTE www.ALTE.org. Ces associations sont une source d'informations et de conseils pour l'évaluation des langues et les procédures de mise en relation.

Malgré tout, la version expérimentale du Manuel a été utilisée par les responsables d'examens de différentes façons :

- en appliquant les procédures à un test élaboré avant le CECRL et par conséquent sans relation évidente avec le cadre, afin de pouvoir donner des résultats en rapport avec les niveaux du CECRL ;
- pour confirmer la relation entre un test datant d'avant le CECRL et le construit du CECRL ainsi que les niveaux du CECRL ; c'est le cas de tests conçus en fonction des spécifications de contenu élaborées par le Conseil de l'Europe depuis 1970 et correspondant à présent aux niveaux du CECRL : Niveau introductif A1 ; niveau intermédiaire A2 ; Niveau seuil B1 ; Niveau autonome B2 ; Niveau indépendant C1 ; Niveau maîtrise C2 (Van Ek et Trim 2001a-c) ;
- en apportant à la révision des examens des informations qui permettent une relation plus étroite avec le concept hypothétique et les niveaux du CECRL ;
- en aidant les écoles à mettre en œuvre des procédures pour relier leurs examens au CECRL.

Même si le Manuel n'a pas été conçu comme un outil servant à relier au CECRL des cadres de référence ou des échelles en usage dans une institution, l'ensemble des procédures proposées peut malgré tout servir à cet effet. Partant d'un cadre en usage, l'étape de spécification peut servir à la mise en relation du contenu de l'examen et de ce qu'il recouvre. Les échantillons de performances calibrés sur le cadre en usage peuvent être utilisés pour un inter-calibrage après une formation sur la standardisation : il est possible d'évaluer des échantillons calibrés du CECRL avec les critères du cadre en usage et de même d'évaluer des échantillons calibrés sur le cadre en usage avec les critères du CECRL utilisés pour les performances orales et écrites fournies par le Manuel. Enfin, une étude de validation externe peut être menée sur des tests ayant pris comme référence le cadre en usage.

Pour aider les utilisateurs à savoir s'il est pertinent d'utiliser les procédures dans leur propre contexte et ce qu'implique leur utilisation, des encadrés reprenant quelques uns des points essentiels et des enjeux sont proposés à la fin de chaque chapitre sur le modèle du CECRL (les utilisateurs peuvent se demander si…).

1.2. Le contexte du manuel

Le cadre européen commun de référence pour les langues se fixe un objectif ambitieux, celui de fournir :

> « … une base commune à l'élaboration de programmes de langues vivantes, de référentiels, d'examens, de manuels, etc. partout en Europe. Il décrit aussi complètement que possible ce que les apprenants d'une langue doivent apprendre afin de l'utiliser dans le but de communiquer ; il énumère également les connaissances et les capacités langagières qu'ils doivent acquérir afin d'avoir un comportement langagier efficace. La description englobe aussi le contexte culturel d'utilisation de la langue. Le cadre définit aussi les niveaux de compétence qui permettent de mesurer les progrès de l'apprenant à chaque étape de l'apprentissage et à tout moment de la vie » (Conseil de L'Europe 2001a :1).

Mais le CECRL traite aussi de l'évaluation et des examens, et c'est à ce niveau que le manuel peut servir de référence :

> « L'un des principaux objectifs du Cadre de référence est d'aider tous les partenaires de l'enseignement et de l'apprentissage des langues à décrire les niveaux de compétence exigés par les standards et les examens existants afin de faciliter les comparaisons entre les différents systèmes de certification. C'est dans ce but qu'ont été élaborés le Schéma descriptif et les Niveaux communs de référence. Ceux-ci fournissent une grille de lecture conceptuelle que les utilisateurs peuvent utiliser pour décrire leur système » (Conseil de l'Europe 2001a :21).

L'objectif du CECRL est de faciliter la réflexion, la communication et le travail en réseau dans le domaine de l'enseignement et de l'apprentissage des langues. Au niveau local, l'objectif de toute stratégie devrait être de répondre aux besoins propres à un contexte. La clef pour concilier les deux objectifs en un système cohérent est la souplesse. Le CECRL est un outil de référence semblable à un accordéon, fournissant des catégories, des niveaux et des descripteurs que des professionnels de l'éducation peuvent regrouper ou subdiviser, détailler ou résumer – tout en gardant la structure hiérarchique commune. On encourage les utilisateurs à mettre en place des ensembles d'activités langagières, de compétences et de performances convenant à leur contexte local mais qui soient aussi en rapport avec le schéma général afin de permettre une communication plus aisée avec des collègues d'autres institutions et d'autres parties prenantes telles que les apprenants, les parents et les employeurs.

Il n'y a pas de contradiction entre d'une part un cadre commun de référence nécessaire à l'organisation de l'enseignement et facilitant les comparaisons et d'autre part des stratégies et des décisions locales nécessaires pour faciliter un apprentissage efficace et élaborer des examens convenant à tout type de contexte.

Le CECRL remplit déjà cette fonction avec souplesse dans son application avec le Portfolio Européen des Langues. Le portfolio est un nouvel outil dans le domaine de l'éducation qui a été conçu grâce à une coopération à la fois intensive et extensive. Les conditions de son application de façon suffisamment uniforme sont assez bonnes, même si le projet du portfolio a dû prendre en compte un certain nombre de contraintes.

Par contre la reconnaissance mutuelle de qualifications langagières octroyées par toutes les parties concernées est une question beaucoup plus compliquée.

En Europe, les professionnels de l'évaluation en langue ont des traditions très différentes. D'un côté, on trouve les producteurs d'examens qui opèrent selon le mode classique d'examens annuels préparés par une commission de spécialistes et notés en fonction de la connaissance intuitive du standard exigé. Il existe de nombreux cas où l'examen ou le test débouchant sur une qualification reconnue est préparé par l'enseignant ou le personnel de l'école plutôt que par une commission externe, parfois sous le contrôle d'un expert extérieur. Il y a ensuite de nombreux examens qui se concentrent sur la mise en œuvre de

spécifications de tâches, avec des critères écrits, un barème et une formation des examinateurs permettant d'assurer une cohérence ; ils incluent ou excluent selon le cas une forme de pré-test ou de validation empirique. Enfin, de l'autre côté, on trouve des systèmes extrêmement centralisés qui utilisent essentiellement des questions à réponse fermée pour mesurer des capacités de réception. Les questions sont extraites de banques d'items. On y ajoute quelquefois des tâches de production (habituellement écrites) afin de mesurer la compétence et de délivrer les certifications. Les politiques nationales, les traditions et les cultures de l'évaluation autant que les politiques, les cultures et les intérêts légitimes des organismes spécialisés dans les tests et les examens de langue sont des facteurs qui peuvent être un frein à l'intérêt qu'il y a à une reconnaissance mutuelle des qualifications. Toutefois, il y va de l'intérêt de chacun que l'on applique des procédures convenables en matière d'évaluation.

Parallèlement à la question de la tradition, se pose celle de la compétence et des ressources. Des établissements reconnus ont, ou peuvent avoir les ressources à la fois humaines et matérielles qui leur permettent de mettre en œuvre et d'appliquer des procédures traduisant de bonnes pratiques ainsi que des systèmes convenables de formation, d'assurance qualité et de contrôle. Dans d'autres cas, l'expérience de l'évaluation et les connaissances nécessaires sont moindres. Il peut n'y avoir qu'une familiarité limitée avec les techniques de travail en réseau et de formation des examinateurs à l'évaluation en fonction de standards et qui sont un préalable à toute évaluation cohérente de la performance. D'un autre côté, il peut n'y avoir que peu de familiarité avec des approches qualitatives et psychométriques, préalable nécessaire à la validation adéquate d'un examen. Mais surtout, il peut n'y avoir qu'une familiarité limitée avec les techniques de mise en relation des examens puisque, dans la plupart des cas, les groupes qui s'occupent d'évaluation ont l'habitude de travailler de manière isolée.

Il n'est donc pas étonnant qu'à la suite de la publication du CECRL, on ait souvent fait appel au Conseil de l'Europe pour qu'il joue un rôle plus actif auprès des producteurs d'examens dans leurs efforts pour valider la relation de leurs examens avec le Cadre européen commun de référence. Ce fut le thème central d'un séminaire aimablement organisé par les autorités finlandaises à Helsinki en juillet 2002 (Conseil de l'Europe 2002) qui déboucha sur la décision de la Division des Politiques linguistiques de Strasbourg de démarrer le projet d'élaboration de ce Manuel.

Ce Manuel fait suite au travail entrepris par la Division des Politiques linguistiques du Conseil de l'Europe pour concevoir des outils permettant la mise en place de projets. Ces outils fournissent un socle commun d'éléments de référence et d'objectifs constituant une structure cohérente et transparente pour un enseignement/apprentissage et une évaluation efficaces correspondant aux besoins à la fois des apprenants et de la société. Cette structure peut aussi faciliter la mobilité personnelle. C'est avec la publication du Threshold level (Van Ek 1976 ;Van Ek et Trim 200o1b) en 1970 , et l'élaboration des versions de ce niveau dans différentes langues (Niveau Seuil (D. Coste 1976) que ce travail s'est fait largement connaître. En 1990, les recherches sur le CECRL et son élaboration ont donné lieu à des expérimentations de deux versions. 2001 a été l'année de la publication de la version définitive en anglais et en français et celle de l'organisation de l'année européenne des langues (Conseil de l'Europe 2001a, 2001b), Le CECRL est publié à présent en plus de 30 langues. A l'origine, les « niveaux communs de référence » (A1-C2) constituaient l'impact principal du CECRL. A présent les concepteurs de programmes s'inspirent du CECRL pour élaborer une nouvelle génération d'objectifs plus détaillés à partir des descripteurs du CECRL (voir partie 4.3.3). Ce Manuel, avec l'accent mis sur la mise en relation réciproque des évaluations grâce à la médiation du CECRL, est un complément logique à ce travail sur les niveaux et les objectifs.

Il n'est nulle part envisagé une équivalence quelconque entre des examens différents qui auraient été reliés au CECRL en suivant les procédures proposées dans ce manuel. Le contenu et la forme des examens varient en fonction des besoins liés au contexte et des traditions de la culture pédagogique qui ont déterminé leur conception. Deux examens peuvent très bien être au « niveau B2 » tout en étant très différents. Des apprenants dans

deux contextes différents peuvent obtenir des résultats différents (a) à un examen dont la forme et le contenu leur sont familiers et (b) à un examen du même niveau conçu dans un contexte différent. Ensuite, ce n'est pas parce qu'ils ont suivi les procédures pour relier les examens préconisées par ce Manuel que plusieurs examens peuvent, par exemple, se réclamer exactement du même niveau, par exemple B2. B2, comme tout autre niveau, est à situer sur une « bande » de performance langagière qui est très large ; le point de césure pour passer d'un niveau à un autre dans ces différents examens peut être déterminé à des endroits différents dans cette bande qui ne correspondent pas tous exactement à la même ligne de démarcation entre le B1 et le B2.

Les programmes et les examens pour l'apprentissage des langues doivent être conçus et adaptés au contexte dans lequel ils seront utilisés. Les auteurs du CECRL sont très clairs à ce propos : le CECRL ne doit en aucune façon être considéré comme un projet d'harmonisation. Le CECRL n'a aucune intention de dire aux professionnels du domaine des langues quels devraient être les objectifs :

« Il n'est PAS dans notre intention de dire aux praticiens ce qu'il faut faire et comment le faire. Nous posons des questions, nous n'y répondons pas. Le CECRL n'a pas pour fonction d'imposer aux intéressés des objectifs à atteindre ou des méthodes à utiliser » (Conseil de l'Europe 2001a :Xi Note à l'utilisateur).

Il n'est pas non plus dans l'intention de ce Manuel de dire aux professionnels du domaine ce que devraient être les standards et la façon de prouver le lien établi avec eux. Le CECRL et ce Manuel ont tous les deux comme objectifs d'encourager la réflexion, de faciliter la communication (entre les professionnels du domaine et entre les parties concernées par l'éducation) et de fournir des outils de référence concernant les processus et les techniques. Les Etats membres et les institutions concernés par l'enseignement et l'apprentissage des langues travaillent et coopèrent de façon autonome ; c'est à eux et elles que reviennent le privilège et la responsabilité du choix de l'approche la plus appropriée à leur but et leur contexte.

Une version expérimentale de ce Manuel a été publiée en septembre 2003 (Conseil de l'Europe 2003) et a été présentée au séminaire de Strasbourg en avril 2004. L'existence de ce Manuel en septembre 2003, juste après la publication complète du CECRL en anglais et en français (2001), a eu un impact considérable. D'une certaine façon, on peut dire que l'importance de l'impact à la fois du CECRL et du manuel est le fruit d'un calcul heureux.

Juste au moment où les concepteurs d'examens cherchaient les moyens de rendre ces examens plus transparents et plus pertinents dans un contexte européen, le CECRL et le Manuel étaient là pour les leur proposer. En conséquence, la méthodologie de beaucoup de projets de mise relation avec le CECRL a été influencée par l'approche proposée dans le Manuel. En même temps, ces approches ont donné lieu à des remises en question et des commentaires à l'occasion des études de cas (plus de 20) menées en relation avec le projet.

Beaucoup de ces études ont été présentées à une réunion à Cambridge en décembre 2007 et au colloque du séminaire d'EALTA à Athènes en 2008. Des retours d'informations d'institutions impliquées dans l'expérimentation et d'un large éventail de professionnels intéressés que ce soit en Europe ou au delà ont largement contribué à la préparation de cette version revue, qui, pour ne pas être totalement définitive, est plus exhaustive. Les articles de la réunion de Cambridge sont publiés dans un recueil d'études de cas dans la série des « Studies in Language Testing » publiée au Cambridge University Press ; les articles de la réunion d'Athènes sont publiés dans un recueil d'études de cas par le CITO, en coopération avec le Conseil de l'Europe et EALTA. Nous espérons que ces études, ce Manuel et l'ensemble croissant d'outils accompagnant le CECRL contribuent au développement de l'expertise pour relier les examens de langues au CECRL et aux discussions sur les enjeux de ce processus.

Les utilisateurs peuvent se demander :

- si l'utilisation du CECRL est pertinente dans leur évaluation et leur contexte ;
- pourquoi et dans quels buts ils appliquent ce Manuel ;
- quelles modifications sont à introduire dans leur contexte spécifique pour l'application du Manuel ;
- quelles parties du Manuel les concernent le plus;
- comment ils pensent faire connaître leurs résultats afin de contribuer à l'amélioration de l'expertise dans le domaine de la mise en relation.

Chapitre 2 : Le processus de mise en relation

2.1. Approche adoptée

2.2. Questions liées à la qualité

2.3. Etapes du processus

2.4. Utilisation du CECRL

2.5. Utilisation du Manuel

2.1. Approche adoptée

Relier un examen ou un test au CECRL est une entreprise complexe. L'existence d'une relation entre l'examen et le CECRL n'est pas un fait directement observable, mais relève d'une affirmation pour laquelle le concepteur d'examen devra apporter des preuves tant au plan théorique qu'empirique. La procédure par laquelle on obtient ces preuves est la « validation de l'affirmation ».

Mettre en relation des examens ou des tests avec le CECRL présuppose que l'on ait défini un ou plusieurs points de césure. Ces points de césure répartissent la distribution des performances des candidats sur deux ou plus de deux niveaux du CECRL.

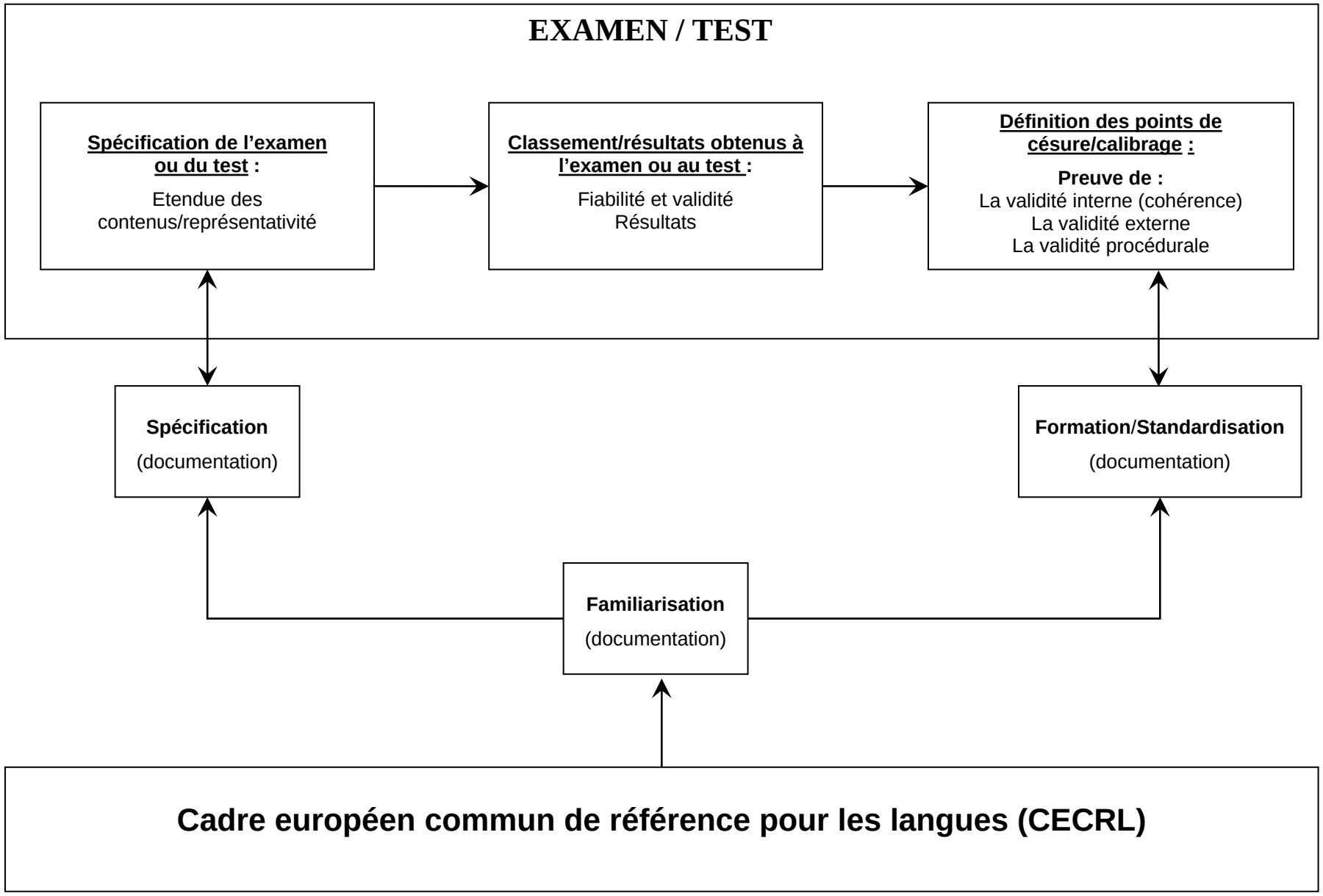
On peut garantir des normes convenables si on applique un processus approprié de définition des points de césure dès le début de la mise en relation. La définition de points de césure implique des prises de décision exigeant des données de haute qualité et un travail rigoureux. A partir du moment où ces décisions peuvent avoir des conséquences importantes, elles doivent être équitables, ouvertes, valides, efficaces et défendables. Ce sera le cas si des processus systématiques éprouvés et des critères explicites sont utilisés.

Lors de la définition de points de césure, il est fréquent de se référer à des contenus et des performances standards. Les contenus standards décrivent les contenus liés aux domaines à partir desquels l'examen peut être ou a été élaboré. Cette description renvoie très souvent aux niveaux de performance. De telles descriptions ont forcément un caractère général et sont habituellement formulées en termes qualitatifs. Dans les textes traitant de définition de points de césure, on les appelle « Descripteurs du niveau de performance » (DNP : cf. partie 6.7) et ils jouent le rôle d'un système général de référence à partir duquel des examens précis peuvent être décrits. Les normes de performance renvoient à des examens spécifiques et présentent la performance minimale pour cet examen ou ce test spécifique ; en ce sens ils sont synonymes de point de césure.

Il faut toutefois souligner un point important. Le Cadre européen commun de référence pour les langues (CECRL) fournit les contenus et les descripteurs du niveau de performance. Les DNP y sont donnés, contrairement à ce qui se passe dans les autres contextes de définition de points de césure, où les DNP doivent être définis en premier.

Cela signifie qu'on doit se référer au CECRL à chaque étape du processus de mise en relation (cf. schéma 2.1). L'approche retenue dans ce Manuel rend indispensable une connaissance approfondie du CECRL.

SCHÉMA 2.1 : PREUVE DE LA VALIDITÉ DE LA MISE EN RELATION DE L'EXAMEN OU DU TEST AVEC LE CECRL



On peut envisager la mise en relation d'un examen ou d'un test avec le CECRL comme un processus « d'élaboration d'une argumentation » basé sur un raisonnement théorique. Le concept de « validité » est au centre de ce processus. Le manuel présente cinq séries de démarches étroitement liées. On conseille aux utilisateurs de suivre ces démarches une à une, ce qui leur permet de concevoir leur plan de mise en relation comme une suite d'activités indépendantes et réalisables :

- Familiarisation
- Spécification
- Formation à la standardisation/calibrage
- Définition des points de césure
- Validation

Le projet doit démarrer par l'étape de « Familiarisation » décrite au chapitre 3. Ce n'est qu'après cette étape qu'il est possible de décrire l'examen ou le test concerné, à l'aide des procédures de « Spécification » (chapitre 4). Ces procédures débutent par des contrôles et des rapports témoignant de la qualité de l'examen (fiabilité et validité) ; la démonstration de la qualité de l'examen est un pré-requis au processus de mise en relation.

La définition des points de césure impliquant que l'on porte des jugements sur les items et les performances, les données obtenues doivent être de grande qualité. C'est pourquoi la formation des personnes concernées doit être extrêmement rigoureuse. Elle est décrite dans le chapitre 5.

Il existe un grand nombre de méthodes permettant de définir les points de césure. Celles que l'on considère comme étant les plus pertinentes dans ce contexte sont décrites au chapitre 6. La qualité de la définition des points de césure peut varier, c'est pourquoi il est important d'apporter des preuves sur le degré de justification des points de césure. Différents types de preuve de validité de la définition des points de césure, qui devront être apportés, sont présentés au chapitre 7.

Les utilisateurs du Manuel devront choisir la procédure la plus appropriée à leur contexte parmi toute la gamme proposée ici ou dans la littérature consacrée au sujet. L'approche retenue est globale. Un des objectifs du Manuel est de favoriser l'application des meilleures procédures même dans les cas où les ressources et l'expertise disponibles sont limitées. Les premiers pas peuvent être modestes mais le but est d'aider les producteurs d'examens à travailler dans un cadre structuré de sorte que le travail ultérieur puisse s'appuyer sur ce qui a été fait précédemment. La structure commune préconisée par le Manuel peut donner l'occasion à des organismes de conjuguer plus facilement leurs efforts et de chercher des synergies dans certains domaines.

Il est important d'insister sur le fait que les cinq séries de procédures (ou « étapes »), ne constituent pas uniquement des jalons isolés les uns des autres sur un processus linéaire. Il est primordial de vérifier, à l'issue de chaque étape, qu'on est sur la bonne voie : l'interprétation des niveaux correspond bien à l'interprétation courante, illustrée par des exemples représentatifs. Dans le cas de la révision ou du développement d'un examen, il est conseillé d'appliquer les procédures recommandées à chaque étape du développement ou de la révision, de façon à ce que la mise en relation avec le CECRL se fasse d'une façon organisée, cyclique, à mesure que l'équipe devient de plus en plus familiarisée avec le CECRL – et que le projet ne soit pas remis en cours de route à une autre équipe, interne ou extérieure à l'institution, avant que le projet principal ne soit achevé.

Bien qu'elles ne doivent pas être considérées comme des jalons sur un parcours linéaire, les cinq étapes s'organisent selon un ordre logique. A chaque étape on demande aux utilisateurs de commencer par les capacités de production (orale et écrite) car ces compétences peuvent être plus directement reliées aux riches descriptions du CECRL, fournissant ainsi une base claire pour la formation, les jugements et les discussions.

2.2. Questions liées à la qualité

La mise en relation d'un examen ou d'un test avec le CECRL ne peut être valide si l'examen ou le test en question ne peut démontrer une validité en lui-même. Un test qui ne convient pas à un certain contexte ne conviendra pas plus s'il est mis en relation avec le CECRL ; de même, un examen qui ne dispose pas de procédures permettant de s'assurer que les examinateurs et les correcteurs appliquent les mêmes normes de sévérité, ou que les versions d'un test administré lors de différentes sessions sont équivalentes, ne peut rendre crédible une affirmation de mise en relation avec le CECRL car il ne peut démontrer de cohérence interne dans l'opérationnalisation de ses normes.

Plusieurs ouvrages de référence proposent des conseils de bonne pratique dans le développement de tests. Ce Manuel ne les remettra pas en question, car son objectif principal est de fournir des conseils pour la définition des points de césure. Le chapitre 7 traite des problèmes liés au développement des tests, à leur expérimentation et analyse. Le «Supplément de référence » propose des informations complémentaires, notamment sur les techniques d'analyses. Le lecteur est toutefois renvoyé à la nombreuse littérature sur ce sujet : Alderson et al. (1995), Davidson & Lynch (2002), Ebel & Frisbee (1986), Downing & Haladyna (2006), Milanovic (2002), Weir (1993), ainsi que l'ensemble des publications et du matériel produits pour le projet « Into Europe » sous les auspices du British Council de Hongrie (www.examsreform.hu/Pages/Exams.html).

La préoccupation pour la qualité dans la conception de tests est également présente dans les critères de bonnes pratiques des organismes suivants :

- EALTA (European Association of Language Testing and assessment – Association européenne de l'évaluation en langues, www.ealta.eu.org). Le *Guide de bonnes pratiques dans l'évaluation en langues* de EALTA comporte une liste abordable des points les plus importants à prendre en compte, avant, pendant et après la conception du test, par tous ceux qui sont impliqués dans l'évaluation et les pratiques de test (qu'il s'agisse d'individus ou d'institutions).

- ALTE (Association of Language testing in Europe – Association des organismes européens d'évaluation en langues, www.alte.org). Le *Code de pratiques* et les *Standards minimum pour établir des profils de qualité en évaluation en langue*, proposent une série de 17 normes minimales qui permettent aux concepteurs de certifications de structurer et d'évaluer la conception de leur test ainsi que son processus d'administration.

- AERA (American Educational Research Association – Association américaine de recherche en éducation, www.aera.net). AERA (en 1999), propose une série détaillée et reconnue de normes théoriques pour les tests dans les domaines de l'éducation et de la psychologie.

- ILTA (International Language Testing Association – Association internationale de l'évaluation en langues, www.ilta.org). Dans la même ligne qu'AERA et d'autres autorités, ILTA a réuni et résumé dans son *Code de pratiques pour les évaluateurs en langues* les principes essentiels théoriques et pratiques de l'évaluation en langues.

2.3. Etapes du processus

Le processus de mise en relation d'un test avec le CECRL consiste à mettre en œuvre les différentes étapes d'une série de procédures :

La familiarisation (chapitre 3) : il s'agit d'une sélection d'activités de formation visant à ce que ceux qui participent au processus de mise en relation parviennent à une bonne connaissance du CECRL, de ses niveaux et de ses descripteurs. Cette étape de « Familiarisation » doit se faire en amont des démarches de « Spécification » et de

« Standardisation ». L'étape de familiarisation constitue également un pré-requis logique à une mise en relation efficace. Une fois cette étape achevée, le degré de réussite de la formation doit être évalué et faire l'objet d'un rapport.

La spécification (chapitre 4) : il s'agit d'un inventaire de l'étendue de ce que l'examen recouvre (contenu et types de tâches) par rapport aux catégories présentées dans le CECRL au chapitre 4 : « L'utilisation de la langue et l'apprenant/utilisateur » et au chapitre 5 : « Les compétences de l'apprenant/utilisateur ». Tout en faisant fonction de compte rendu, ces procédures servent également, dans une certaine mesure, à la prise de conscience qui pourra ultérieurement aider à l'amélioration de la qualité de l'examen en question. Les fiches A2 et A8-A20 du chapitre 4 mettent l'accent sur l'analyse des contenus et la relation qu'ils entretiennent avec le CECRL. On peut considérer la spécification comme une méthode essentiellement qualitative : elle apporte des preuves à l'aide « d'arguments fondés sur le contenu ». Des méthodes quantitatives (Kaftandjieva, 2007) peuvent également être utilisées pour la validation des contenus.

La formation à la standardisation, le calibrage (chapitre 5): les démarches proposées facilitent la mise en œuvre d'une compréhension commune des « Niveaux communs de référence », à l'aide des exemples représentatifs des performances orales et écrites. Ces procédures renforcent la familiarité avec les niveaux du CECRL, telle qu'elle a été obtenue grâce aux activités présentées au chapitre 3 (familiarisation). Elles garantissent que les évaluations des performances reflètent les construits décrits dans le CECRL. Il est logique de standardiser ainsi – par une formation suffisante- l'interprétation des niveaux, avant de passer a) au calibrage d'exemples de performances locales et de tâches/items (partie 5.7), et b) à la définition de points de césure (chapitre 6). Un calibrage réussi d'exemples locaux peut venir à l'appui d'une affirmation basée sur les résultats de la spécification. En effet, si les conclusions du processus de calibrage indiquent que les échantillons de performances du test ont été avec succès calibrés sur les niveaux pour lesquels ils avaient été conçus, cela confirme l'affirmation à laquelle on est arrivé dans la spécification.

La définition des points de césure (chapitre 6) : le point crucial dans le processus de mise en relation d'un examen avec le CECRL est l'instauration d'une règle permettant de décider si on attribue l'un des niveaux du CECRL à un candidat, à partir de la performance qu'il a réalisée lors de l'examen. On prend généralement une décision sur les scores de césure, sur les performances à la limite du niveau. Les étapes précédentes, de familiarisation, spécification et standardisation peuvent être considérées comme des activités préparatoires à des décisions valides et logiques. Le chapitre 6 décrit les procédures qui mènent aux décisions finales permettant de définir les scores de césure. Le matériel présenté s'appuie sur l'importante littérature au sujet de la définition des points de césure et les procédures présentées au chapitre 6 ont été choisies parmi les nombreuses procédures disponibles censées convenir au contexte de l'évaluation en langues. Des procédures supplémentaires basées sur l'exploitation de jugements d'enseignants et sur la théorie de réponse à l'item (TRI) pour inclure un critère extérieur (par exemple des items représentatifs du CECRL, ou des évaluations d'enseignants utilisant des descripteurs du CECRL) dans une étude de mise en relation, sont présentées dans le *Matériel supplémentaire* (Extra Material) fourni par Brian North et Neil Jones.

La validation (chapitre 7) : Bien que les étapes précédentes de familiarisation, spécification, standardisation et définition des points de césure puissent être réalisées dans un ordre chronologique, il serait imprudent d'attendre que tout soit terminé avant d'entreprendre les activités de validation, comme si elles constituaient l'ultime verdict sur la qualité du processus de mise en relation. La validation doit plutôt être considérée comme un processus continu de contrôle de la qualité, qui peut permettre de répondre à la question générale : « Avons-nous atteint notre but pour cette activité ? ». On a cité plus haut un exemple simple mais néanmoins important : il est important de donner aux participants une formation à la familiarisation et à la standardisation, mais il est tout aussi important de vérifier si les activités de formation ont atteint leur but ; c'est cela qu'on entend par validation. Certains aspects de la validité ainsi que les procédures permettant de réunir des preuves de la validité sont présentés dans ce dernier chapitre (chapitre 7).

2.4. Utilisation du CECRL

Un cadre commun de référence permet à différents examens d'être reliés entre eux indirectement sans qu'ils prétendent être exactement équivalents. L'objectif d'un examen peut varier, mais ce qu'il recouvre peut être défini en relation directe avec les catégories et les niveaux du CECRL. De même que deux étudiants de niveau B2 sont à ce niveau pour des raisons différentes, deux examens de niveau B2 auront des aspects qui ne seront pas totalement identiques.

Les parties du CECR les plus pertinentes pour effectuer la mise en relation des examens sont :

- le Chapitre 3 : « Les Niveaux communs de référence » ;
- le Chapitre 4 : « L'utilisation de la langue et l'apprenant utilisateur » ainsi que les échelles pour les *Activités langagières communicatives* et pour les *Stratégies langagières communicatives*;
- le chapitre 5 : « Les compétences de l'utilisateur/apprenant », en particulier la partie 5.2 « Compétences langagières communicatives » ainsi que les échelles qui illustrent les aspects de la compétence linguistique, pragmatique et sociolinguistique.

Les utilisateurs de ce Manuel trouveront sur le site du Conseil de l'Europe le texte complet du CECRL, les documents qui s'y rapportent, ainsi qu'un certain nombre d'outils utiles :

Documents :

- Le CECRL, en anglais et en français, incluant les annexes.
- Des liens vers les versions en d'autres langues : www.coe.int/lang et www.coe.int/portfolio
- Le Manuel, incluant les annexes.
- Les fiches et grilles de référence du Manuel.
- Le *Supplément de référence*.

Grilles d'analyse de contenus

- CECRL : grille d'analyse de contenus pour la réception orale et écrite (parfois appelée « grille du CECRL néerlandais) : annexe B1.
- CECRL : grilles d'analyse de contenus pour la production orale et écrite, développée par ALTE : annexe B2.

Descripteurs explicatifs (www.coe.int/portfolio)

- Les descripteurs du CECRL (en anglais).
- La banque de descripteurs du Portfolio européen des langues, montrant la relation entre ces descripteurs et les descripteurs originaux.
- Un recueil de descripteurs C1/C2 (en anglais), du CECRL et de projets liés au CECRL, indiquant ceux qui sont calibrés sur les niveaux du CECRL et ceux qui ne le sont pas.

Echantillons illustrés

- Documentation des DVD présentant des échantillons représentatifs de productions orales (adultes), disponibles, à ce jour, en anglais, français, italien et portugais[1].
- Documentation du DVD illustrant des productions orales d'adolescents en allemand, anglais, espagnol, français et italien, calibrés au Séminaire de calibrage inter-langues à Sèvres en juin 2008.
- Echantillons représentatifs de productions écrites, disponibles à ce jour en allemand, anglais, français, italien et portugais.
- Items représentatifs de réception orale et écrite en allemand, anglais, espagnol, français et italien.

D'autres ressources seront ajoutées à la « Boîte à outils » du CECRL. On en trouvera la liste à www.coe.int/lang et www.coe.int/portfolio dès qu'ils seront disponibles.

[1] Le DVD en allemand est publié avec sa documentation (Bolton et al., Langenscheidt, 2008).

Parties du CECRL particulièrement appropriées
L'utilisateur de ce Manuel trouvera particulièrement utiles, dans une perspective globale, les échelles et descriptions de niveaux suivants :

	Version anglaise	**Version française**
Vue d'ensemble des Niveaux communs de référence		
▪ Tableau 1 « Niveaux communs de référence », Chapitre 3	p. 24	p. 25
▪ Partie 3.6 « Cohérence de contenu dans les niveaux communs de référence »	pp. 33-36	pp. 32-34
▪ Document B5 « Cohérence dans le calibrage des descripteurs	pp.223-224	pp. 159-160
▪ « Niveaux de compétence dans le Cadre de référence de ALTE »	pp. 249-250	pp. 176-177
Vue d'ensemble des activités communicatives		
▪ Tableau 2, Grille des Portfolios pour l'auto évaluation	pp. 26-27	pp. 26-27
▪ DIALANG Document C3 « Echelles descriptives détaillées … »	pp. 238-243	pp. 170-172
▪ ALTE Document D1 « Résumé des capacités langagières »	p. 251	p. 178
▪ Réception générale de l'oral : échelle	p. 66	p. 55
▪ Réception générale de l'écrit : échelle	p. 69	p. 57
▪ Interaction orale générale : échelle	p. 74	p. 61
▪ Interaction écrite générale : échelle	p. 83	p. 68
▪ Production orale générale : échelle	p. 58	p. 49
▪ Production écrite générale : échelle	p. 61	p. 51
Vue d'ensemble des aspects de la compétence langagière communicative		
▪ Table 3 « Aspects qualitatifs de l'utilisation de la langue parlée	pp. 28-29	p. 28
▪ Etendue linguistique générale : échelle	p. 110	p. 87
▪ Correction grammaticale	p. 114	p. 90
▪ Adéquation sociolinguistique	p. 122	p. 95
▪ Aisance à l'oral	p. 129	p. 100

En ce qui concerne les examens liés au monde du travail ou à l'entrée en université, les utilisateurs trouveront, en outre, les échelles suivantes, particulièrement appropriées dans la mesure où elles traitent des demandes fonctionnelles.

Activités communicatives particulièrement pertinentes dans les domaines éducatif et professionnel		
▪ Comprendre en tant qu'auditeur	p. 67	p. 56
▪ Prendre des notes (conférences, séminaires)	p. 96	p. 77
▪ Lire pour s'orienter	p. 70	p. 58
▪ Lire pour s'informer et discuter	p. 70	p. 58
▪ Lire des instructions	p. 71	p. 59
▪ Traiter un texte	p. 96	p. 77
▪ Echange d'information	p. 81	p. 67
▪ Discussions et réunions formelles	p. 78	p. 64

▪ Comprendre une interaction entre locuteurs natifs	p. 66	p. 55
▪ Monologue suivi : argumenter	p. 59	p. 50
▪ S'adresser à un auditoire	p.60	p. 50
▪ Essais et rapports	p. 62	p. 52

Le calibrage des descripteurs du CECRL est décrite dans l'annexe A du CECRL, North (2000a), North et Schneider (1998), et Schneider et North (2000).

2.5. Utilisation du Manuel

Les chapitres suivants concernent les différentes étapes du processus de mise en relation. Pour chaque étape, l'utilisateur peut choisir parmi l'ensemble de procédures proposées celles qui correspondent le mieux à leur contexte.

Le manuel ne prétend pas être un modèle pour la conception d'un nouvel examen. En revanche il a vocation à encourager une réflexion sur les bonnes pratiques. En fait, les utilisateurs qui ont expérimenté la première version ont considéré que suivre les procédures indiquées permettait une analyse critique et une évaluation du contenu et des caractéristiques de l'examen – et qu'en fait le résultat du processus avait autant d'importance que l'affirmation de la mise en relation.

Le Manuel propose un ensemble de procédures et de techniques de principe qui étayent le processus techniquement complexe et exigeant de mise en relation. Chaque étape suppose un jugement averti. C'est à l'organisme certificateur concerné qu'incombe la responsabilité de la mise en place d'un processus cohérent et approprié. Cette responsabilité comprend :

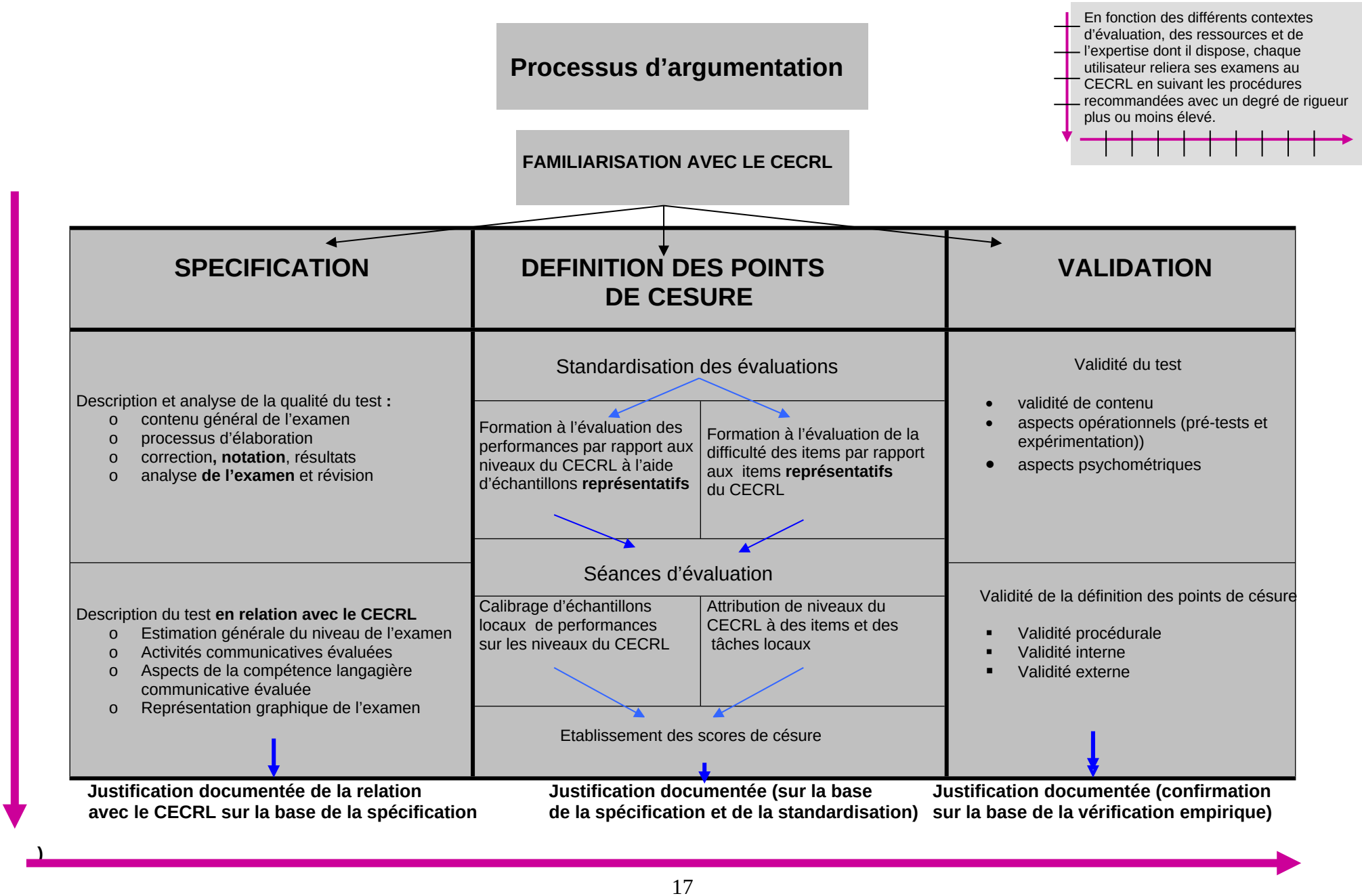
- Une réflexion sur les besoins, les ressources et les priorités dans le contexte concerné.
- Un choix des procédures adéquates parmi celles qui sont proposées ou parmi d'autres dont fait état la littérature.
- Une gestion réaliste du projet selon une approche modulaire et par étape qui en assure la qualité.
- Une collaboration et une mise en réseau avec des collègues d'autres domaines professionnels et d'autres pays.
- Une utilisation réfléchie des procédures.
- Une communication fidèle des résultats.
- Une communication précise, transparente et détaillée des conclusions.

Le schéma 2.2 est une représentation graphique des étapes du processus de mise en relation avec le CECRL. Il souligne le fait que la mise en relation d'un examen ou d'un test peut être considérée comme une suite d'arguments justifiant ses différents aspects et proposant des preuves certifiant leur validité au fur et à mesure que le processus se développe. Les organismes certificateurs peuvent considérer qu'ils ne peuvent pas tous entreprendre des études dans tous les domaines indiqués dans le Manuel. Cependant, même ceux qui disposent de peu de ressources doivent choisir un certain nombre de techniques dans tous les domaines. Une affirmation selon laquelle un examen est relié au CECRL ne peut être prise au sérieux qu'à partir du moment où une preuve existe que cette affirmation, fondée sur la spécification (contenus standards) et sur la définition des points de césure (performances standards) est confirmée par la validation.

Les utilisateurs du manuel peuvent se demander avant d'entamer le processus de mise en relation :

- Ce que l'approche proposée signifie, de manière générale, dans leur contexte.
- Ce que l'approche proposée signifie, de façon plus spécifique (temps, ressources, etc.), dans leur contexte.
- Si les différentes procédures sont praticables dans leur contexte.
- S'ils doivent se concentrer sur une ou plusieurs procédures ou bien appliquer les principes de chacun des cinq ensembles de procédures de façon limitée, en particulier si les ressources sont limitées.
- Comment ils vont justifier leur conclusion auprès du grand public et de leurs collègues.

SCHÉMA 2.2. : REPRESENTATION GRAPHIQUE DES PROCEDURES PERMETTANT DE RELIER LES EXAMENS AU CECRL



Chapitre 3 : Familiarisation

3.1. Introduction

3.2. Activités préalables au séminaire

3.3. Activités introductives pendant le séminaire

3.4. Analyse qualitative des échelles du CECRL

3.5. Préparation à l'évaluation

3.1. Introduction

Avant d'entreprendre les activités de Spécification et de Standardisation, il faut organiser des tâches de « familiarisation » pour que les personnes impliquées dans une démarche de mise en relation de leurs examens avec les niveaux du Cadre aient une excellente connaissance de ce processus. L'expérience tirée des études de cas et les séminaires de calibrage produisant des DVD ont mis en évidence que de nombreux professionnels du domaine des langues participant à un projet de mise en relation ont, en fait, un niveau de familiarisation avec le CECRL bien inférieur à celui qu'ils pensent avoir. Alors que la plupart des professionnels connaissent bien les tableaux du CECRL plus globaux (tableau 1 : échelle globale et tableau 2 : grille d'auto-évaluation du Portfolio), beaucoup n'ont pas une idée très précise des caractéristiques de la compétence de l'apprenant aux différents niveaux dans les différentes capacités langagières.

Il faut faire une différence entre la familiarisation avec le CECRL, avec les instruments d'évaluation à utiliser et avec les activités à entreprendre. Il n'y a pas de frontière nette entre la fin de la familiarisation et le début de la spécification ou de la standardisation ; à chaque fois, les premières activités de la tâche principale s'inscrivent dans le continuum du processus de familiarisation.

Il faut aussi prendre en compte ce qui est en jeu et avoir à l'esprit le public et les applications par un panel sélectionné d'experts ou l'application du CECRL par une équipe ou à l'échelle d'une institution. Il faut aussi se demander quelles activités de mise en relation pourront servir comme introduction à une session spécifique de familiarisation. Le temps que les individus vont consacrer aux activités de familiarisation dépend essentiellement du degré de familiarité qu'ils ont avec le CECRL. La durée que prendra le processus de familiarisation dans sa totalité (repris avant les activités de spécification et de standardisation) dépendra de l'objectif et de l'importance du projet de mise en relation.

Les membres du groupe de travail peuvent aussi être nettement influencés par des normes institutionnelles locales données aux niveaux du CECRL, ainsi que par leurs interprétations des descripteurs ou par les variantes locales des descripteurs du CECRL. De plus, ils ignorent souvent qu'il existe une différence entre le niveau des descripteurs du CECRL (dans toutes les sous-échelles ainsi que dans les tableaux récapitulatifs 1,2 et 3) et les « niveaux plus » du CECRL (que l'on trouve uniquement dans les sous-échelles). Il est important que ceux qui sont impliqués dans le processus de mise en relation se concentrent sur les descripteurs du CECRL – et ne se laissent pas influencer outre mesure par des descripteurs représentant une performance exceptionnelle à ce niveau (un « niveau plus »).

C'est avec ces éléments à l'esprit que ce chapitre propose des activités de familiarisation dans les quatre parties indiquées ci-dessous. Ces techniques sont expliquées plus en détail dans la suite du chapitre. Il est vivement conseillé aux utilisateurs de sélectionner des activités dans chaque partie au début des processus de Spécification et de Standardisation.

Activités préalables au séminaire

Avant un atelier de familiarisation, il faudrait que chaque membre de l'équipe responsable du projet entreprenne plusieurs activités bien ciblées qui rappellent les aspects importants des niveaux du CECRL.

a) Lire la section 3.6 du CECRL (version française pages 32 à 34) qui décrit les principales caractéristiques des niveaux, issues des descripteurs représentatifs.
b) Sélectionner les questions se trouvant dans l'encadré à la fin des parties concernées du chapitre 3 du CECRL (Niveaux communs de référence), chapitre 4 (L'utilisation de la langue et l'apprenant/utilisateur) et chapitre 5 (Les compétences de l'utilisateur/apprenant).
c) Aller sur le site CEFT (www.CEFTrain.net), qui se concentre sur les caractéristiques des niveaux et qui propose, uniquement pour l'anglais, des exemples de vidéos, des textes et des échantillons d'items d'examens pour l'enseignement dans le primaire, le secondaire et l'enseignement aux adultes.

Activités introductives pendant le séminaire

d) Choisir le texte concernant les différents niveaux dans le tableau A1 de ce manuel, qui résume les traits caractéristiques des niveaux communs de référence (CECRL 3.6).
e) Faire une auto-évaluation de son niveau de langue dans une langue étrangère – à l'aide du tableau 2 du CECRL (grille d'auto-évaluation du PEL) suivie d'une discussion en tandem.

Analyse qualitative des échelles du CECRL

f) Trier selon leur niveau ou leur rang les descripteurs d'une échelle du CECRL pour une capacité langagière. Par exemple, pour la production orale, on peut utiliser les descriptifs de l'aspect qualitatif « Aisance » ou bien deux ou trois niveaux du CECRL apparentés (par exemple Conversation, tours de parole dans la Discussion informelle).
Pour réaliser cette activité, on découpe les descripteurs qui composent l'échelle.

g) Reconstituer le tableau 2 du CECRL à partir des descripteurs de chaque case.

Préparation pour l'évaluation des capacités langagières de production orale et écrite

h) Reconstituer la grille d'évaluation du CECRL qui va être utilisée et dans laquelle certaines cases sont vides. Si le séminaire commence par la production orale, ce sera le tableau 3 du CECRL (tableau C2 du Manuel). Si le séminaire commence par la production écrite, ce sera le tableau C4 de ce Manuel (ou réciproquement).

i) Montrer des performances filmées d'apprenants sur les DVD illustrant les niveaux du CECRL dans la langue concernée.

3.2. Activités préalables au séminaire

Les organisateurs d'activités de familiarisation doivent bien faire la différence entre une présentation du CECRL et un séminaire/atelier de familiarisation. Alors que la première vise à présenter de façon générale l'importance et le contenu du CECRL à des fins diverses, la familiarisation est supposée assurer une connaissance suffisamment fine des niveaux du CECRL pour analyser et évaluer des tâches d'examens et des performances en rapport avec eux.

Le séminaire de familiarisation sera d'autant plus utile et réussi que le coordinateur aura réuni les documents nécessaires et les informations permettant aux participants de s'y préparer en leur faisant parvenir un « paquet de pré-tâches » (par envoi postal ou par courrier électronique) 2 à 3 semaines avant le séminaire. Cela donnera l'occasion aux participants qui ont déjà assisté à une présentation du CECRL de « rafraîchir » leur mémoire

et aux autres d'étudier le matériel de présentation du CECRL. Quel que soit le degré de familiarisation des participants avec le CECRL, le coordinateur doit les informer qu'une préparation à l'atelier suppose un minimum de 3 à 5 heures de travail si on prend en compte les trois activités.

Après la première information sur le CECRL, une des activités suivantes peut être choisie pour commencer le séminaire lui-même ou pour contribuer à la cohésion du groupe.

a) Lecture de la partie 3.6 du CECRL (dont le tableau A1)

On recommande cette activité aux organisateurs qui ne connaissent pas avec certitude le degré de familiarisation des participants avec les niveaux du CECRL mais elle peut aussi rafraîchir la mémoire des connaisseurs. On demande aux participants de prendre connaissance des niveaux du tableau A1 et du texte de la partie 3.6. pour pouvoir identifier les caractéristiques de chaque niveau et indiquer de façon sûre le niveau atteint par des apprenants avec lesquels ils travaillent. Le travail qui a été fait individuellement avant le séminaire peut être repris comme activité introductive ou pour « rompre la glace ».

b) Prise en compte d'une sélection des questions du CECRL de l'encadré

Cette activité convient plutôt à une majorité de professionnels qui sont supposés avoir une certaine connaissance des niveaux du CECRL (qui ont par exemple travaillé avec le CECRL ou qui connaissent les niveaux). L'objectif de l'exercice est de leur faire prendre conscience des nombreux aspects à prendre en compte lors de la conception et de l'analyse des tâches d'examens ainsi que de l'étendue de ce que le CECRL recouvre.

On peut organiser cette activité suivant différentes modalités :

- On peut photocopier une liste de contrôle telle que celle qui est présentée ci-dessous, centrée sur la production orale pour amener les participants à réfléchir aux différents aspects en jeu dans l'évaluation de la production orale.

Les personnes qui utilisent le Cadre pour analyser et évaluer les performances de production orale envisageront et expliciteront selon le cas : - *comment les conditions matérielles dans lesquelles l'apprenant sera amené à communiquer affecteront ce qu'il doit faire* ; - *comment le nombre et la nature des interlocuteurs affecteront ce que l'apprenant doit faire* ; - *avec quelles contraintes de temps l'apprenant devra effectuer sa performance* ; - *dans quelle mesure les apprenants devront s'adapter au contexte mental de leur interlocuteur* ; - *comment tenir compte de la perception du niveau de difficulté d'une tâche pour l'évaluation de sa réalisation réussie et pour l'(auto) évaluation de la compétence communicative de l'apprenant.*	Approprié ? Pourquoi ?

- Les coordinateurs peuvent sélectionner les questions des encadrés du CECRL qui leur semblent pertinentes et élaborer une nouvelle liste de contrôle en fonction des capacités langagières sur lesquelles le groupe travaillera.
- Les coordinateurs peuvent s'inspirer du travail accompli par les participants lors de cette activité quand ils discuteront des exercices consistant à trier (f-g) la partie 3.4.

c) Avoir accès au site de formation CEFTrain

Le projet CEFTrain[2] a consisté à choisir des activités visant à familiariser les professeurs avec les niveaux du CECRL. Il comprend des exercices avec les échelles du CECRL, les tâches et les performances (pour des enseignements dans le primaire, le secondaire et l'enseignement aux adultes) qui ont été analysées et mises en relation avec les niveaux du CECRL en tenant compte des avis partagés des membres du projet. Ce site est très utile car il propose aux participants un exemple de ce qui va se faire pendant le séminaire. On conseille aux participants de fixer leur attention sur les capacités langagières qui correspondent à leurs préoccupations et dont il sera question pendant le séminaire.

3.3. Activités introductives pendant le séminaire

Après avoir accueilli les participants, le coordinateur doit s'assurer qu'ils ont bien compris la finalité du séminaire et son organisation.

La première activité du séminaire consiste à présenter brièvement l'importance du CECRL pour l'évaluation ; par la suite, le coordinateur organisera une ou deux des activités présentées ci-dessous, tout en s'assurant que les participants réinvestissent le travail fait avant le séminaire.

d) Tri du texte des différents niveaux du tableau A1

C'est une activité qui permet de faire le lien avec le travail fait individuellement avant le séminaire.

- L'exercice consiste à demander aux participants de trier les traits caractéristiques du tableau A1 du manuel qui est une simplification de la partie 3.6 du CECRL. Il faut supprimer les références aux niveaux pour que les participants soient obligés de lire attentivement les descripteurs. Le coordinateur distribue une feuille avec les descripteurs dans le désordre et la tâche consiste à attribuer des niveaux A1 à C2 aux descripteurs.

- Une fois le travail achevé, le coordinateur distribue le tableau A1 avec les réponses.

- Le coordinateur demande ensuite aux participants d'échanger- en tandem ou en petits groupes - leurs points de vue sur les traits caractéristiques de chaque niveau du CECRL, selon la lecture qu'ils ont faite du tableau A1 et de la partie 3.6 (activité a) du CECRL et de l'exercice de tri qu'ils viennent d'achever. La meilleure façon de procéder est de demander aux participants de surligner les éléments clefs.

- On peut demander aux participants quel est le niveau qui leur semble le plus approprié dans leur activité professionnelle, puis leur demander de former des groupes de même niveau et distribuer une liste de contrôle des descripteurs du niveau, comme celle du prototype suisse du PEL disponible sur le site www.sprachenportfolio.ch/esp_e/esp15plus:index.htm (sélectionner dans le menu de gauche : ELP model 15+ ; Learners ; Downloads).

e) Auto- évaluation avec le tableau 2 du CECRL

Cette activité est un bon point de départ pour des groupes de participants qui connaissent déjà le Portfolio. Le tableau 2 représente une partie importante du PEL et on en parle souvent comme de « la grille du PEL ».

---

[2] Le projet CEFTrain est un projet européen Socrate dont l'Université d'Helsinki a assuré la coordination avec des partenaires de quatre pays : l'Italie, l'Autriche, l'Allemagne, et l'Espagne, et la participation de Neus Figueras, une des auteures de ce manuel.

- On demande aux participants de faire une auto-évaluation de leur compétence dans deux langues étrangères à l'aide de la grille du PEL (Tableau 2 du CECRL). Ils en débattent ensuite avec leurs voisins. L'importance de cette discussion ne doit pas être sous-estimée. La discussion doit être dirigée de telle façon que les participants prennent conscience de l'existence de profils non uniformes. Le coordinateur explique alors comment le CECRL prend en compte cette non uniformité et encourage sa reconnaissance.

- On peut avantageusement compléter cette auto-évaluation (conseillée aux utilisateurs du PEL) en consultant une liste de contrôle des descripteurs du CECRL, telle qu'on la trouve dans le prototype suisse du PEL déjà mentionné, correspondant au niveau en question.

- On peut également demander aux participants d'auto-évaluer leur niveau en termes de ***qualité*** : dans quelle mesure font-ils bien ce qu'ils disent savoir faire ? Pour ce faire, on peut utiliser :

    a) soit le Tableau 3 du CECRL (tableau C2) qui définit chaque niveau pour :
        - l'étendue linguistique,
        - la correction grammaticale,
        - l'aisance,
        - la cohérence
        - et l'interaction.

    b) soit l'échelle correspondant à l'Aisance (CECRL p.100) et l'échelle pour la Correction grammaticale (CECRL p.90)

3.4. Analyse qualitative des échelles du CECRL

Une fois les activités introductives achevées, il faut poursuivre la familiarisation par un travail d'approfondissement des niveaux du CECRL et par des discussions sur les descripteurs spécifiques de la compétence. Le coordinateur doit choisir au moins l'une des deux options suivantes.

f) Tri des descripteurs isolés d'une échelle du CECRL

L'activité de tri de descripteurs a été largement expérimentée dans le projet suisse d'élaboration des descripteurs, dans la conception du PEL dans des contextes différents et dans plusieurs projets finlandais. Cette tâche a le mérite d'obliger les participants à examiner les descripteurs indépendamment les uns des autres comme des critères autonomes.

Elle exige toutefois qu'on la prépare et qu'on en fasse une activité relativement simple.

- Le coordinateur prépare des enveloppes à l'avance pour chaque participant ou pour un tandem. Chaque enveloppe contient une ou plusieurs échelles dont les descripteurs ont été découpés en bandes. Si l'on mélange des échelles apparentées (par exemple, Conversation, tours de parole dans la Discussion informelle), on doit s'assurer que le nombre de descripteurs isolés n'excède pas 40 ! En découpant les descripteurs, il faut veiller à supprimer la ligne de séparation entre deux descripteurs consécutifs afin de ne pas donner d'indication sur la capacité ou l'incapacité du coordinateur à couper droit ! On demande aussi aux participants de ne rien écrire sur les bandes pour pouvoir les réutiliser.

- Individuellement ou par deux, les participants trient alors les descripteurs selon leur niveau. Ils peuvent commencer avec « A », « B » ou « C » qu'ils divisent ensuite ou se lancer tout de suite dans les six niveaux, s'ils le souhaitent.

- Ils en discutent ensuite avec les autres participants afin d'arriver à un consensus.

- Puis ils comparent avec la bonne réponse.

Il faut s'attendre à ce que certains descripteurs ne se retrouvent pas à leur place mais, en règle générale, si l'on a pris le temps nécessaire pour atteindre un consensus, l'ordre trouvé sera plus ou moins le même que celui des échelles du CECRL.

g) Reconstitution du Tableau 2 du CECRL

Cette activité est une variante de la précédente mais elle utilise le tableau 2 du CECRL (grille du PEL) – elle-même élaborée à partir des descripteurs du CECRL – plutôt que les échelles du CECRL. On peut utiliser l'ensemble des échelles (6 activités langagières x 6 niveaux = 36 descripteurs) ou une version plus simple (une seule colonne = 6 descripteurs). Là aussi, la meilleure façon de procéder est de mettre les cases découpées dans une enveloppe.

- On distribue la grille vierge agrandie au format A3 de la grille du PEL dont les cases ont été vidées de leur contenu. On demande aux participants de replacer les descripteurs dans les cases convenables.

- Pour éviter de faire perdre du temps aux participants, on peut affecter les descripteurs de symboles correspondant aux différentes capacités langagières. Il est en effet inutile de leur faire trouver que « Je peux utiliser des expressions et des phrases simples pour décrire l'endroit où je vis et les gens que je connais » est un descripteur de Production orale.

- Cette activité peut aussi être menée en vidant de leur contenu la moitié seulement des cases. Il est recommandé de procéder ainsi avec des grands groupes ou dans des salles dans lesquelles il n'y a que de petites tables.

On a constaté que la combinaison de cette activité de reconstitution avec l'auto-évaluation de son propre niveau de langue (c : voir ci-dessus) était particulièrement efficace si on la pratiquait comme suit :

- En petits groupes, les participants lisent attentivement chaque descripteur et en discutent pour reconstituer le tableau. Le coordinateur contrôle le travail de groupe et aide à clarifier les doutes sur l'interprétation des différents descripteurs.

- Le coordinateur distribue une copie du Tableau 2 achevé et « complet » pour que les participants vérifient leur exercice de reconstitution et pour faciliter la discussion.

- On demande aux participants de faire une auto-évaluation de leur propre connaissance des langues étrangères (d'abord individuellement) puis d'en discuter avec le groupe en se référant au Tableau 2 du CECRL « Niveaux communs de compétences – Grille pour l'auto-évaluation ».

3.5. Préparation à l'évaluation

Une fois que l'on s'est assuré que les participants se sont familiarisés avec les niveaux du CECRL, on peut entamer la dernière étape de familiarisation. Cela suppose une préparation plus poussée à l'évaluation de tâches et de performances dans les différentes capacités langagières. S'il s'agit d'évaluer des tâches de réception écrite ou orale, le coordinateur peut décider de ne pas faire l'activité (i). Par contre l'activité (h) est obligatoire pour chaque capacité langagière avant de commencer l'évaluation.

h) Reconstitution de la grille du CECRL à utiliser

Le coordinateur va préparer cette activité à partir de l'échelle qu'il va utiliser pour évaluer les tâches ou performances.

L'exercice est organisé exactement de la même façon qu'en (f) (tri des descripteurs du CECRL).

Au lieu de trier les descripteurs découpés et mis dans une enveloppe, on peut utiliser une fiche type d'une liste de contrôle avec, dans le désordre, les descripteurs de la capacité langagière. C'est ensuite aux participants de rattacher chaque descripteur au niveau correspondant (comme cela est décrit dans le (d) ci-dessus).

A l'issue des discussions sur les descripteurs et les « corrections » apportées en grand groupe, le coordinateur distribue une liste de contrôle complétée avec les réponses.

i) Exemples filmés représentatifs de performances d'étudiants.

Cette activité donne une très bonne idée correspondant à la réalité des niveaux du CECRL. Elle est tout à fait appropriée même si les participants ne vont pas travailler sur la production orale.
Le coordinateur ne peut mener à bien cette activité que s'il peut avoir accès aux échantillons de performances du CECRL (www.ciep.fr/publi_evalcert/dvd-productions-orales-cecrl/index.php). Il faut choisir avec soin les performances les plus appropriés, en termes de niveau et d'âge. La procédure à suivre peut être la suivante :

- Le coordinateur fait visionner la performance et demande aux participants d'attribuer un niveau en utilisant le tableau A1.
- On distribue ensuite aux participants, avant qu'ils ne discutent entre eux, le tableau 3 du CECRL (Tableau C2) et on leur demande de confirmer le niveau choisi individuellement.
- Le coordinateur demande ensuite aux participants de discuter, en petits groupes, du niveau qu'ils ont attribué en se référant au tableau 3 du CECRL (tableau C2).
- Le coordinateur annonce ensuite le niveau attribué à la performance et distribue les commentaires qui justifient du niveau attribué (voir le site ci-dessus), toujours en se référant aux descripteurs du tableau 3 du CECRL (tableau C2).

Tableau 3.1 : Gestion du temps pour les activités de Familiarisation

Familiarisation

- Ces activités peuvent être organisées indépendamment de toute autre activité de formation. On peut les utiliser au début des activités de Spécification et de Standardisation.
- Elles durent environ 3 heures :

- Brève présentation du CECRL par le coordinateur	(30 minutes)
- Activités introductives (d - e) et discussion	(45 minutes)
- Activités qualitatives (f - g) y compris le travail de groupe	(45 minutes)
- Préparation à l'évaluation (h-i)	(30 minutes)
- Conclusion	*(15 minutes)*

Tableau 3.2 : Documents à préparer pour les activités de familiarisation

- Ensemble de documents à envoyer par courriel ou par la poste aux participants avant la rencontre
  - Tableau 1 du CECRL
  - Partie 3.6
  - Listes de questions reprenant les encadrés du CECRL les mieux adaptés à la situation (à la fin de chaque chapitre).

- Copies des descripteurs dans le désordre des traits caractéristiques du tableau 2.1 pour tous les participants

- Copies du Tableau A1 du manuel pour tous les participants
- Copies du tableau 2 du CECRL pour tous les participants (tous contextes)
- Versions découpées du Tableau 2 du CECRL en vue du travail de groupe (tous contextes, un ensemble par enveloppe, une enveloppe pour chaque groupe de travail)
- Descripteurs découpés des échelles appropriées du CECRL pour l'évaluation choisie (pour entrer dans le détail d'une compétence particulière : une échelle découpée par enveloppe, une enveloppe pour chaque sous-groupe de travail). par exemple :
  - pour la Production orale : (1) Interaction orale générale, (2) Aisance à l'oral, (3) Etendue linguistique générale ;
  - pour la réception orale : (1) Compréhension générale de l'oral, (2) Comprendre une interaction entre locuteurs natifs, (3) Comprendre des émissions de radio et des enregistrements).
- Copies de listes de contrôle de descripteurs[3] pour un ou deux niveaux particuliers, choisies sur l'ensemble des échelles du CECRL (pour entrer dans le détail d'un niveau donné).
- Copies du tableau 3 du CECRL (tableau C2) quand nécessaire.
- Sélection de deux échantillons représentatifs de performances d'étudiants filmées en vidéo
- Documentation sur les échantillons de performance utilisés.

Les utilisateurs du manuel peuvent se demander :

- *jusqu'à quel point les participants se sont familiarisés avec les finalités et les fonctions du CECRL ;*
- *quelle est la meilleure stratégie pour renforcer la familiarisation avec le CECRL ;*
- *s'il est nécessaire de demander aux groupes de lire ou relire certains chapitres ou des parties en supplément du 3.6 du CECRL ;*
- *quelles questions de l'encadré peuvent être utiles ;*
- *s'il serait judicieux de donner une tâche préliminaire sur le CEFRL, de recueillir le travail et l'analyser ou le faire de façon informelle ;*
- *quelles seraient les échelles de niveaux le plus utiles pour effectuer les exercices de tri ;*
- *s'il faut montrer des échantillons représentatifs du DVD à cette étape ;*
- *Si un moyen tel qu'un quiz serait approprié pour savoir s'il est nécessaire de renforcer la familiarisation ;*
- *Si les résultats de cette étape de familiarisation entraînent une modification de l'organisation.*

[3] A cet effet, n'utiliser que des descripteurs validés du PEL : il faudrait pouvoir faire correspondre chaque descripteur adapté du PEL au descripteur d'origine du CECRL - comme ce qui est fait dans la base de données de descripteur de Günther Schneider et Peter Lenz dans www.coe.int/portfolio

Chapitre 4 : Spécifications

4.1 Introduction

4.2 Description générale de l'examen

4.3 Outils disponibles pour la spécification

4.3.1 Tableaux et fiches
4.3.2 Grilles d'analyse de contenus

4.3.2.1 Grille d'analyse du CECRL pour la réception orale et écrite
4.3.2.2 Grille d'analyse du CECRL pour la production orale et écrite

4.3.3 Ouvrages de référence

4.4 Procédures

4.5 Déclaration du niveau : représentation graphique de la relation de l'examen avec les niveaux du CECRL

4.1. Introduction

Ce chapitre traite de l'analyse du contenu d'un examen ou d'un test dans le but de décrire le ou les niveaux du CECRL qu'ils recouvrent. La procédure proposée peut prendre la forme d'un débat ou d'une analyse individuelle suivie d'un débat.

Au final, en se fondant sur les spécifications, l'institution disposera de descriptions détaillées lui permettant de déclarer le degré de relation de ses examens avec les catégories et les niveaux du CECRL.

Toutefois, comme cela a été précisé dans le chapitre 2, la déclaration du degré de relation n'est recevable que si, parallèlement, sont apportées, pour toutes les étapes du développement et de l'administration de l'examen ou du test, des preuves de bonnes pratiques, d'une validité interne convenable et de procédures adéquates assurant la qualité.

Ce chapitre a trois objectifs :

Contribuer à sensibiliser encore plus :
- à l'importance d'une bonne analyse du contenu de l'examen de langue ;
- au CECRL et particulièrement à ses échelles de descripteurs ;
- aux raisons de relier les examens de langue à un cadre de référence international tel que le CECRL ;
- aux moyens d'utiliser le CECRL pour la planification et la description des examens de langues.

Définir des normes minimales pour :
- la qualité des contenus des spécifications des examens de langue ;
- le processus de mise en relation des examens avec le CECRL.

Apporter aux utilisateurs une aide adaptée pour :
- compléter l'analyse de contenus et le processus de mise en relation proposés ;
- apporter la preuve de la cohérence interne et de la validité du construit ;
- faire une déclaration de niveau qui rendra les résultats des examens en question plus transparents, à la fois pour les utilisateurs de ces résultats et les candidats eux-mêmes.

Les procédures de spécifications exposées dans ce chapitre impliquent 4 étapes :

- assurer une familiarisation convenable avec le CECRL (chapitre 3) ;
- analyser le contenu de l'examen ou du test en question par rapport aux catégories pertinentes du CECRL ; l'utilisateur devra décrire un domaine évalué dans son examen ou son test et qui s'avérerait non traité dans le CECRL;
- mettre en relation l'examen ou le test avec l'échelle de descripteurs adéquate du CECRL, sur la base de l'analyse de contenus ;
- faire une première déclaration sur le degré de mise en relation de l'examen ou du test avec l'un des niveaux du CECRL, en se fondant sur l'analyse de contenus.

Ces procédures impliquent trois types d'activités :

- les activités de familiarisation décrites dans le chapitre 3 ;
- la description détaillée du contenu de l'examen de langue, consignée dans un certain nombre de fiches complétées ;
- l'utilisation des descripteurs adéquats du CECRL afin de relier l'examen de langue à ses niveaux et à ses catégories.

Ces procédures liées aux spécifications donnent aux concepteurs d'examens l'occasion :

- d'être encore plus sensibles à l'importance d'une bonne analyse du contenu d'un examen ;
- de se familiariser avec l'utilisation du CECRL pour la planification et la description des examens de langue ;
- de décrire et d'analyser en détail le contenu d'un examen ou d'un test ;
- de fournir la preuve de la qualité de leur examen ou de leur test ;
- de fournir la preuve de la relation de leur examen ou de leur test avec les niveaux du CECRL ;
- d'apporter des conseils aux rédacteurs d'items ;
- d'accroître, pour les enseignants, les évaluateurs, les utilisateurs d'examens et les candidats, la transparence des contenus, de la qualité et de la relation d'un examen ou d'un test avec le CECRL. Les fiches à compléter ont une fonction de sensibilisation (processus) et seront utilisées pour étayer la déclaration qui sera faite (produit final).

Les procédures décrites ici ont été spécialement conçues pour ce Manuel. Il en existe cependant d'autres. Les utilisateurs de ce Manuel peuvent consulter des procédures d'analyses descriptives permettant de relier un examen à un cadre de référence (par exemple Alderson et al. 1995, Chapitre 2 ; Davidson et Lynch, 1993, 2002 ; Lynch et Davidson, 1994, 1998).

4.2. Description générale de l'examen

La première étape consiste en une définition et une description claire de l'examen ou du test que l'on va relier au CECRL. La validité interne est-elle acceptable ? Pourrait-on recommander un travail d'approfondissement de certains domaines afin d'accroître ou de confirmer la qualité de l'examen et donc le sérieux des résultats de la mise en relation avec le CECRL ? L'expérience acquise lors des études de cas qui ont guidé la rédaction de l'avant-projet du présent manuel a montré que cette démarche permettait de remettre en question certains aspects opérationnels de l'examen et reflétait bien jusqu'à quel point l'examen et les procédures qui lui sont associées, remplissait ses objectifs. Ce processus de sensibilisation ne peut être entrepris par une seule personne (chercheur ou membre de l'équipe). Cet exercice met parfois en évidence un manque de cohérence entre les spécifications officielles de l'examen –qui n'ont peut-être pas été modifiées depuis des années- et l'examen lui-même –tel qu'il a été administré récemment. L'exercice est assurément plus facile s'il existe des spécifications formelles de l'examen. S'il n'en existe pas, le procédé consistant à compléter les fiches de ce chapitre aidera l'utilisateur à prendre en compte certains aspects qui devraient faire partie intégrante de ces spécifications.

On trouvera les fiches suivantes en annexe, partie 2 :

A1 : Description générale de l'examen
A2 : Conception de l'examen
A3 : Correction
A4 : Notation
A5 : Communication des résultats
A6 : Analyses des données
A7 : Justification des décisions

Avant de compléter les fiches, l'utilisateur doit se munir d'une part des spécifications et d'autre part des copies des trois derniers examens administrés aux candidats. S'il s'agit de relier au CECRL une suite d'examens de différents niveaux, une fiche par examen devra être complétée.

La fiche A1 permet de définir les buts et les objectifs de l'examen ainsi que sa population cible. Elle permet également d'avoir une vue d'ensemble des activités communicatives évaluées, des différentes épreuves ainsi que des renseignements fournis et de la façon dont les résultats sont communiqués aux utilisateurs (candidats et centres d'examens).

Les fiches A2 à A6 décrivent les étapes les plus importantes du cycle de conception, développement et administration d'un examen. On y consignera des informations sur la conception, la correction, la notation, la façon de communiquer les résultats et les analyses de données :

- Fiche A2 : processus de conception
- Fiche A3 : critères de correction et barèmes de notation pour chaque épreuve
- Fiche A4 : notation et procédures de définition des points de césure pour chaque épreuve
- Fiche A5 : communication des résultats
- Fiche A6 : analyses et procédures de révision
- Fiche A7 : (justification des décisions). Le concepteur d'examen pourra ici expliquer et justifier ses décisions. Par exemple, pour quelles raisons certains domaines sont évalués et d'autres non ? Pourquoi une pondération particulière est-elle utilisée ? Pourquoi la double correction n'est-elle qu'exceptionnellement mise en œuvre ? Pour quelle raison ne fournit-on pas les résultats par épreuve ou par capacité langagière ? Cela relève-t-il d'un problème de fiabilité ou d'une décision politique ?
- Fiche A8 : elle permet de consigner l'estimation initiale de l'institution quant au niveau global du CECRL évalué par l'examen.

Estimation initiale du niveau global du CECRL		
☐ A1	☐ B1	☐ C1
☐ A2	☐ B2	☐ C2
Brève justification, références à de la documentation :		

Fiche A8 : Estimation initiale du niveau global du CECRL

Le processus détaillé de spécification est exposé dans les fiches A9 à A22 (cf. annexe A, parties A2-A5).

La fiche A23 présente les résultats du processus de spécification sous la forme d'un graphique illustrant les catégories et les niveaux pertinents du CECRL couverts par l'examen analysé. Cette fiche est traitée et illustrée au § 4.5.

Les procédures sont strictement les mêmes pour un examen de langue générale et pour un examen sur objectifs spécifiques. Le CECRL prend en effet en compte les différents

domaines (public, personnel, éducationnel et professionnel). De même, si les activités de communication langagière sont regroupées dans les catégories « Réception, interaction, production et médiation » plutôt que sous les quatre capacités langagières traditionnelles, c'est afin de pouvoir prendre en compte efficacement les objectifs spécifiques éducationnels et professionnels.

4.3. Outils disponibles pour la spécification

Ces outils liés au CECRL sont de trois types. Outre le CECRL lui-même, traduit, à la date de cette publication, en 36 langues, on trouvera :

- Les tableaux et les fiches annexés à ce Manuel.
- Les grilles d'analyses de contenus qui permettent de détaillé de façon extrêmement fine les tâches proposées dans l'examen, en les classant selon des critères standards.
- Les référentiels pour les différentes langues, particulièrement utiles pour les spécifications linguistiques.

4.3.1. Tableaux et fiches

Ce chapitre propose une série de tableaux tirés des échelles de descripteurs du CECRL et accompagnés de fiches à compléter. Le CECRL étant extrêmement détaillé, le nombre de fiches est considérable. Elles sont disponibles dans les parties A2 à A5 des annexes ainsi qu'en téléchargement sur le site www.coe.int/lang

Il existe des fiches et des tableaux associés pour chaque activité langagière communicative (chapitre 4 du CECRL) ainsi que pour les aspects de la compétence langagière communicative (chapitre 5 du CECRL). Les fiches apportent une analyse détaillée de l'examen ou du test en question et permettent de les relier aux sous-échelles appropriées du CECRL. Pour la plupart des fiches, une brève description, une référence et/ou une justification sont demandées.

Dans les études de cas qui ont conduit à la rédaction de ce manuel, plusieurs utilisateurs ont indiqué que compléter ces fiches s'avérait être une bonne méthode pour reconsidérer ce que recouvre un examen et pour réévaluer sa fiabilité.

4.3.2. Grilles d'analyse de contenus

Les grilles d'analyse de contenus du CECRL pour la réception orale et écrite ainsi que pour la production orale et écrite ont été conçues pour que les utilisateurs de ce Manuel puissent décrire leur examen de façon bien plus détaillée que ce que permettent de faire les sous-échelles du CECRL et les tableaux de l'annexe A, cités au paragraphe 4.2. En effet, chaque tâche individuelle proposée dans l'examen y sera répertoriée.

Dans les études de cas qui ont conduit à la rédaction de ce Manuel, certains utilisateurs ont plébiscité ces grilles, les trouvant beaucoup plus utiles que les fiches utilisées actuellement. Ceux qui souhaiteront s'aider du Manuel pour développer un nouvel examen ou pour analyser de façon critique un examen ou un test précis trouveront sans doute ces grilles particulièrement utiles.

Des grilles vierges ainsi que des grilles complétées à titre d'exemple peuvent être téléchargées sur le site www.coe.int/portfolio

4.3.2.1. Les grilles d'analyse de contenus pour la réception orale et écrite

Les grilles d'analyse de contenus pour la réception orale et écrite du CECRL sont en ligne et permettent aux concepteurs d'examens et de tests d'analyser les épreuves de réception orale et écrite afin de les relier au CECRL[4].
La grille permet de consigner, à partir d'une série de choix tirés directement ou indirectement du CECRL, les caractéristiques de chaque tâche, de chaque support, de chaque item de l'examen ou du test : source, type de discours, niveau de difficulté estimé, etc.). Une excellente connaissance du CECRL est naturellement nécessaire pour pouvoir utiliser les grilles de façon totalement efficace. Une composante « familiarisation » avec le CECRL est par conséquent comprise dans cette procédure afin d'apporter des conseils plus approfondis.

Un lien avec la version en ligne des grilles est disponible à l'adresse www.coe.int/portfolio
Le lien direct est www.lancs.ac.uk/fss/projects/grid
On trouvera une version papier des grilles en annexe B.

La version papier permet de compléter les grilles avec les nouvelles catégories liées aux programmes (**curriculum/syllabus**).

Si les grilles ont été conçues pour l'analyse des épreuves de réception orale et écrite, elles sont cependant également utilisables pour la conception d'épreuves de réception. Dans certaines études de cas, elles ont été utilisées pour la formation à la « standardisation » (cf. chapitre 5).

4.3.2.2. Les grilles d'analyse de contenus pour la production orale et écrite

Les grilles d'analyse des tâches de production orale et écrite du CECRL ont été aussi conçues pour aider les utilisateurs à décrire de façon standardisée les caractéristiques des tâches de leurs examens, et leur relation avec le CECRL. Les grilles[5], modifiables en tant que de besoin, sont toutes disponibles sur le site du Conseil de l'Europe. Deux modes d'utilisation sont possibles pour chacune des deux grilles : un pour l'analyse et l'autre pour la présentation du rapport. Pour plus d'information sur les grilles, cf. l'annexe B2.

Grilles pour l'analyse (« Données d'entrées ») : l'utilisation de ces deux grilles convient lors d'ateliers dans lesquels les participants les complètent pour une série de tâches données. L'objectif est alors de préciser les caractéristiques des tâches, les performances attendues (longueur de la réponse, type de discours, registre, etc.), les outils de classement et les commentaires faits aux candidats. Un exemple de tâche est accompagné de cette analyse, d'un échantillon de réponses et de la note attribuée ainsi que d'un commentaire. Les grilles sont utiles pour former les concepteurs de tâches à la standardisation des tâches présentées pour différentes langues, au même niveau.

---

[4] Un groupe de travail constitué de J. Charles Alderson (coordinateur du projet), Neus Figueras, Henk Kuijpers, Günther Nold, Sauli takala et Claire Tardieu, a développé, sur financement du ministère néerlandais de l'Education, un outil permettant de décrire et de classer les tâches de réception orale et écrite en suivant au plus près les descripteurs du CECRL. A l'aide d'un second financement de ce ministère, le groupe a ensuite conçu une version électronique de cet outil, disponible à l'adresse www.lancs.ac.uk/fss/projects/grid Cet outil était à l'origine connu sous le nom de « la grille néerlandaise ». Pour plus d'information, cf. Partie B1 en annexe ainsi que Alderson et al. (2006). Un rapport détaillé est disponible auprès du coordinateur du projet, à l'adresse c.alderson@lancaster.ac.uk

[5] Les grilles pour la production orale et écrite ont été produites dans le groupe « Intérêt spécifique pour le Manuel » de ALTE, en coopération avec le Conseil de l'Europe. La genèse des grilles remonte aux listes de contrôle pour l'analyse de contenus de ALTE. Conçues en 1993 grâce à une subvention LINGUA (93-09/1326/UK-III), leur objectif était de faciliter la comparaison du matériel d'examen entre les différentes langues. La conception des grilles décrites ici a pris en compte le travail réalisé dans le projet néerlandais « Dutch Construct Project » qui a produit les grilles de production orale et écrite.

Le fait de compléter les grilles permet de passer aisément des chapitres « Spécification » et « Standardisation » de l'interprétation des niveaux du CECRL à des exemples concrets (cf. chapitre 5). Elles peuvent également servir à sélectionner des exemples qui seront utilisés pour le calibrage (cf. chapitre 5).

Grilles pour la présentation (« Données de sortie ») : l'objectif de cette forme simplifiée des grilles est de rendre compte de la description des tâches issue de la grille d'analyse présentée ci-dessus (pour la réception orale et écrite). Elles fournissent une information détaillée qui peut constituer la base de bons guides de documentation et d'examen, à condition qu'elles soient complétées par les références adéquates aux critères qualitatifs du CECRL pour chaque échantillon calibré (table 3 du CECRL et Manuel table C.2).

4.3.3. Ouvrages de référence

Dans les procédures de spécification, l'analyse de contenus se réfère principalement au CECRL lui-même. Cependant, en tant que *cadre commun*, le CECRL ne traite par définition d'aucune langue en particulier. Les ouvrages de référence suivants, qui détaillent les spécifications de contenus pour des langues précises, peuvent donc être utiles :

- La série de spécifications de contenus reliés au CECRL, conçue en collaboration avec le Conseil de l'Europe pendant les années 1970-1990, donc ***avant*** l'élaboration du CECRL. Pour l'anglais, la série de spécification est : A1 : « Breakthrough - Découverte»[6] ; A2 : « Waystage – Survie » (van Ek et Trim, 2001a) ; B1 : « Threshold Level – Un Niveau seuil » 5van Ek 1976 ; van Ek et Trim 2001b) ; B2 : « Vantage Level – Compétence opérationnelle effective » (van Ek et Trim 2001c).
- La série des référentiels de niveaux, reliés au CECRL, qui ont été conçus pour différentes langues ***depuis*** la parution du CECRL. On trouvera une liste à jour sur le site www.coe.int/lang qui inclut les ouvrages suivants :
    - Pour l'allemand : Glaboniat, M., Müller, M., Scmitz, H., Rusch, P., Wertenschlag, L. (2002/5) *Profile DEUTSCH (A1-A2. B1-B2. C1-C2.)*, Berlin: Langenscheidt.
    - Pour le français : Beacco etal. (2004, 2006, 2007, 2008) *Niveau B2/A2/A1/A1.1 pour le français : un référentiel*.
    - Pour l'espagnol : Instituto Cervantes (2007) *Niveles de referencia para el español – Plan curricular del Instituto Cervantes : A1, A2-B1, B2-C1, C2).*
    - Pour l'italien : Parizi, F. et Spinelli, B. (publication à venir) Profilo della Lingua Italiana, Fireze : La Nuova Italia.

4.4. Procédures

Avant de compléter les fiches proposées en annexe A ou sur le site www.coe.int/lang , les procédures impliquent que vous consultiez le CECRL, les annexes de ce Manuel et les autres ouvrages de référence cités ci-dessus.

1. **Choix de la commission** : la première étape est la mise en place d'une commission d'experts, si possible mixte (appartenant à l'institution / organisation et extérieurs), et la désignation d'un coordinateur. Ce groupe d'experts internes et externes devrait être constitué de représentants des différentes étapes de la conception d'un examen ou d'un test de langue.

2. **Familiarisation** : avant de mettre en œuvre les procédures de spécification il est essentiel que la commission se familiarise avec le CECRL lui-même. La commission doit donc commencer son travail par les activités de familiarisation du chapitre 3.

3. **Choix de la méthode** : une fois cette étape effectuée, le groupe doit prendre connaissance des multiples fiches et tableaux associés ainsi que des outils de

[6] *Breakthrough*, le niveau « découverte », n'a pas été publié mais est disponible auprès des secrétariats du Conseil de l'Europe et de ALTE.

spécification cités au paragraphe 4.2. Il décidera alors du choix de la méthode et des fiches et tableaux qui seront complétés. Il n'est pas prévu que *toutes* les fiches de l'annexe A soient complétées. Il est rappelé que seules les fiches correspondant aux contenus de l'examen doivent être complétées ; le groupe doit choisir les fiches pertinentes pour l'analyse de l'examen en question. Exemple : si un examen comporte uniquement des tâches lexicales, seules les fiches correspondantes seront complétées et seule l'échelle du niveau de vocabulaire sera examinée. Autre exemple : si un examen mesure plusieurs compétences linguistiques dans différentes capacités langagières, on devra alors compléter un plus grand nombre de fiches et examiner plus d'échelles.

La norme minimale est que les fiches suivantes soient complétées :

- Les fiches de la phase 1 (Description générale : A1 à A7)
- La Fiche A8 (Première estimation du niveau global de l'examen)
- Certaines des fiches numérotées de A9 à A22- qui correspondent à l'examen ou au test en question
- La Fiche A23 (Représentation graphique de la relation de l'examen avec les niveaux du CECR)
- La fiche A24 (Confirmation de l'estimation du niveau global de l'examen)
- Les preuves pertinentes qui permettent d'étayer la déclaration

4. **Activités langagières communicatives** : on complètera normalement en premier les fiches portant sur les activités langagières communicatives (fiches A9-A18). Comme cela a été précisé ci-dessus, chacune des fiches peut être complétée par la personne appropriée de l'institution impliquée. On peut cependant souhaiter procéder de façon plus interactive. L'information consignée dans les fiches sera plus fiable si plus d'une personne est impliquée. Chaque membre de la commission va donc compléter tout ou partie des fiches sélectionnées. Un consensus devra ensuite être obtenu grâce à la confrontation des fiches complétées.

   Le tableau 4.1 présente une vue d'ensemble des fiches et des échelles du CECRL qui y sont reliées. A la fin de la plupart des fiches, il est demandé aux utilisateurs de comparer l'épreuve en question avec la sous-échelle correspondante du CECRL.

Tableau 4.1 : Fiches et échelles du CECRL pour les activités langagières communicatives

Fiche	Activités de communication langagière	Fiche	Echelle
A9	Réception orale	✓	✓
A10	Réception écrite	✓	✓
A11	Interaction orale	✓	✓
A12	Interaction écrite	✓	✓
A13	Production orale	✓	✓
A14	Production écrite	✓	✓
A15	Combinaisons de compétences intégrées	✓	
A16	Compétences intégrées	✓	✓
A17	Médiation orale	✓	
A18	Médiation écrite	✓	

Tableau 4.2 : échelles du CECRL pour les aspects de la compétence langagière communicative

Aspects de la compétence langagière communicative								
	RECEPTION		INTERACTION		PRODUCTION		MEDIATION	
	Réception orale	Réception écrite	Interaction Orale	Interaction Ecrite	Production Orale	Production Ecrite	Médiation Orale	Médiation Ecrite
Compétence linguistique								
Etendue linguistique générale	✓	✓	✓	✓	✓	✓	✓	✓
Etendue du vocabulaire	✓	✓	✓	✓	✓	✓	✓	✓
Maîtrise du vocabulaire			✓	✓	✓	✓	✓	✓
Correction grammaticale			✓	✓	✓	✓	✓	✓
Maîtrise du système phonologique			✓		✓		✓	
Maîtrise de l'orthographe				✓		✓		✓
Compétence sociolinguistique								
Correction sociolinguistique	✓	✓	✓	✓	✓	✓	✓	✓
Compétence pragmatique								
Souplesse			✓	✓			✓	✓
Tours de parole			✓					
Développement thématique	✓	✓		✓	✓	✓	✓	✓
Cohésion et cohérence	✓	✓			✓	✓	✓	✓
Aisance à l'oral			✓		✓		✓	
Précision	✓	✓			✓	✓	✓	✓
Compétence stratégique								
Reconnaître des indices et faire des déductions	✓	✓					✓	✓
Tours de parole (reprise)			✓					
Coopérer			✓	✓				
Faire clarifier			✓	✓				
Planifier					✓	✓		✓
Compenser			✓	✓	✓	✓	✓	✓
Contrôler et corriger			✓	✓	✓	✓	✓	✓

5. **Compétence langagière communicative**: On complétera ensuite les fiches qui concernent les aspects de la compétence langagière communicative (fiches A19-A22). Le Tableau 4.2 donne une vue d'ensemble des différentes compétences communicatives pour lesquelles il est possible de consigner des informations. Cette partie est organisée différemment. Un tableau des descripteurs du CECRL est fourni. Les utilisateurs doivent ensuite renseigner la fiche correspondante sur la base d'une analyse des épreuves de l'examen ou du test en question. A la fin de chaque fiche, les utilisateurs comparent l'examen et l'échelle correspondante du CECRL. Une description ainsi qu'une indication du niveau de chacun des aspects pertinents des compétences retenues dans le CECRL sont demandées. Le même groupe d'experts peut compléter les fiches de façon interactive.

   Les fiches sont proposées dans cet ordre :
   - Réception : fiche A19
   - Interaction : fiche A20
   - Production : fiche A21
   - Médiation : fiche A22

   Aucune échelle du CECRL n'est fournie pour la médiation. Les utilisateurs se réfèreront aux descripteurs pour la réception et la production.

4.5. Déclaration du niveau : représentation graphique de la relation de l'examen avec les niveaux du CECRL

Une fois l'examen analysé en fonction des catégories du CECRL, le résultat obtenu doit être présenté sous la forme d'un graphique montrant clairement la relation avec les niveaux du CECRL. Cette représentation permet de visualiser le contenu de l'examen étudié, rapporté aux sous-échelles appropriées du CECRL pour ce qui concerne les activités langagières communicatives et les aspects de la compétence linguistique (cf. ci-dessous un exemple de fiche A23 complétée).

C2								
C1								
B2.2								
B2								
B1.2								
B1								
A2.2								
A2								
A1								
Panorama	Réception orale	Réception écrite	Conversation sociale	Echange d'information	Notes Messages et Formulaires	Socio linguistique	Pragmatique	Linguistique

Fiche A23 : Représentation graphique de la relation de l'examen aux niveaux du CECRL (exemple)

Dans le graphique ci-dessus, l'axe Y (vertical, à gauche) représente les niveaux du CECRL. Sur l'axe X on représentera la compétence langagière générale et les activités langagières communicatives ainsi que les aspects de la compétence linguistique. Chaque colonne a comme intitulé une catégorie pertinente du CECRL. Les cases qui représentent l'examen ou les épreuves traités seront ombrées. Si l'examen est d'un niveau plus élevé dans certaines catégories, on le montrera en ombrant les cases correspondantes comme dans l'exemple de la Fiche A23 ci-dessus.

L'intitulé des colonnes de la Fiche A23 peut ne pas correspondre à celui qui a été donné aux épreuves de l'examen. Quelques intitulés peuvent correspondre aux épreuves mais il est possible d'en ajouter d'autres, en tant que de besoin. Il se peut, par exemple, que l'examen étudié ne propose pas d'épreuve spécifique pour la compétence linguistique mais que le concepteur de l'examen veuille cependant indiquer aux utilisateurs le niveau de compétence linguistique attendu.

Les démarches présentées dans ce chapitre mettent l'accent à la fois sur le *processus* et sur le *résultat*. On encourage les praticiens à suivre un processus d'analyse de contenus et de mise en relation avec le CECRL. On recommande vivement de réexaminer chaque hypothèse sur le niveau avancée au cours du processus. Il est fort probable que l'estimation initiale donnée dans la fiche A8 doive être modifiée. L'utilisateur doit reconsidérer les

analyses et proposer un jugement raisonné. L'estimation (fiche A8) est confirmée ou révisée dans la fiche A24.

Les chapitres suivants fournissent des outils qui permettent de renforcer la déclaration de niveau. Une recherche plus poussée et une analyse plus approfondie lors d'étapes ultérieures peuvent entraîner une révision de la déclaration avancée. L'exactitude de la déclaration est subordonnée à un large processus de vérification argumentée. On recommande vivement aux concepteurs d'examens d'impliquer leurs collègues dans des débats et des échanges tout au long du processus.

Estimation confirmée (déclaration) du niveau global du CECRL		
☐ A1	☐ B1	☐ C1
☐ A2	☐ B2	☐ C2
Brève justification, références à de la documentation. Si cette fiche présente une conclusion différente de l'estimation initiale consignée dans la fiche A8, merci de commenter les raisons principales de ce changement.		

Fiche A24 : Estimation confirmée (déclaration) du niveau global du CECRL

Les utilisateurs de ce Manuel peuvent se demander :

- S'il est important de réunir et/ou d'analyser des informations ou des données avant d'entreprendre l'étape de spécification.
- S'ils utiliseront les grilles d'analyse de contenus du CECRL.
- Si tous les examens ou les tests peuvent être reliés au CECRL.
- Si le fait d'achever l'étape de spécification présage des changements dans le plan initial d'utilisation de ce Manuel.
- Si l'expérience acquise à l'issue de l'étape de spécification implique, dans l'examen ou le test analysé, des changements qui pourraient intervenir lors de la prochaine réforme programmée.
- Comment ils décideront que l'étape de spécification a été achevée de façon satisfaisante.

Chapitre 5 : Formation à la standardisation et au calibrage

5.1. Introduction

5.2. La formation nécessaire

5.3. Planification préalable

5.4. Animation des stages
5.4.1. Arriver à un consensus et le vérifier

5.5. Formation avec des performances orales et écrites
5.5.1. Performance orale
5.5.2. Performance écrite

5.6. Formation avec des tâches et des items de capacités de réception écrite, orale et de compétences linguistiques
5.6.1. Familiarisation nécessaire
5.6.2. Formation à la définition des points de césure (standard setting)

5.7. De la formation au calibrage

5.7.1. Echantillons nécessaires
5.7.2. Arriver à un consensus et le vérifier
5.7.3. Analyse des données
5.7.4. Documentation

5.1. Introduction

Le but de la démarche de mise en relation des examens avec les niveaux du CECRL est de permettre une catégorisation des candidats en termes de niveaux de compétences du CECRL, de telle façon que cette catégorisation reflète de façon fiable ce que signifient les niveaux du CECR. Si on considère qu'un étudiant est au niveau B1, il faut être tout à fait certain que cet étudiant est vraiment représentatif des descripteurs de ce niveau. Il s'agit là de la validité. Les procédures qui suivent renvoient à la définition des points de césure (standard setting), (voir partie B dans le supplément de référence de ce manuel).

Il existe deux grandes façons d'attribuer des niveaux à des candidats. Il peut s'agir soit d'un simple jugement global de la part du professeur ou de l'examinateur, soit des notes qui sont attribuées au résultat de l'examen. La première option est en général choisie pour les capacités de production, alors que la deuxième concerne généralement les capacités de réception. La distinction n'est pourtant pas aussi tranchée. Dans des épreuves de production écrite, parmi les deux ou trois tâches proposées, chaque tâche peut être notée en fonction de critères analytiques. La totalité des notes obtenues par un candidat peut être traitée de la même façon que le résultat d'une épreuve de réception écrite comportant un certain nombre d'items séparés. Pour éviter tout malentendu, on utilisera respectivement les termes d'examen indirect (examens avec des résultats à base de notes) et d'examen direct (examens évalués de façon globale).

- **Examens directs**. Dans des examens évalués de façon globale le jugement sur le niveau (les six niveaux du CECRL) est direct et il est pour cette raison important d'aider les évaluateurs à émettre des jugements valides. Le principal outil utilisé pour ce genre particulier de définition des points de césure est appelé ***calibrage***. Le calibrage consiste à proposer un (ou plusieurs) échantillons représentatifs illustrant des performances à un niveau donné à la fois pour la formation à la standardisation et comme outil de référence pour les décisions ultérieures concernant des performances de candidats.

- **Examens indirects.** Pour les examens avec des résultats à partir de notes, il faut établir des performances standards. La performance standard est la limite entre deux niveaux de l'échelle continue, indiquée par un examen, et qui est représentée par une note de césure. Une note de césure de 30, par exemple, signifie qu'une note de 30 ou plus acquise à l'examen signifie qu'un certain niveau ou un niveau plus élevé est atteint (par exemple B1), alors qu'un résultat moins élevé correspondra à un niveau plus bas que le niveau du score de césure (dans ce cas B1). On appelle généralement le processus pour arriver à une note de césure la ***définition des points de césure***. Dans le cas des capacités de réception (écrite et orale) ou des compétences sous jacentes (grammaire, lexique), il est important de prendre des décisions sur ces notes de césure.

Les procédures de ***calibrage*** et de ***définition de points de césure*** supposent des décisions collectives qui doivent être soigneusement préparées par une formation adéquate. Le but principal de ce chapitre est d'aider à cette formation.

Comme le calibrage est la suite logique de la formation, il fait partie de ce chapitre.

La définition des points de césure est un thème complexe, largement discuté, souvent sujet à controverse et qui a fait l'objet de nombreuses publications. C'est pour cette raison que les procédures permettant de définir les points de césure sont présentées séparément au chapitre 6. Le coordinateur choisira, parmi l'éventail des méthodes exposées dans le chapitre 6, le supplément de référence et les nombreuses publications sur la question, la ou les méthodes qui conviennent le mieux au contexte ou au but recherché.

Néanmoins, bien que les procédures à suivre vont dépendre de la ou des méthode(s) choisies pour la définition des points de césure, elles seront, dans la majorité des cas, identiques à celles qui sont décrites dans les différentes parties de ce chapitre.

5.2. La formation nécessaire

Les écrits publiés sur la définition des points de césure évoquent très souvent l'importance du groupe d'experts qui recommande la ou les note(s) de césure ou la performance standard, et traite longuement des enjeux que constituent la façon de former ce groupe ; le nombre d'évaluateurs impliqué ; leur parcours professionnel ; les connaissances et l'expertise dans le domaine concerné que ce groupe devrait avoir ; le moment et la durée de la formation.

Des renseignements utiles et détaillés sur la façon d'organiser et de planifier les activités préalables liées aux procédures de définition des points de césure sont fournis par Kaftandjieva, dans la partie B du supplément de référence à ce manuel (2004), Hambleton et Pitoniak (2006) et CCCCizek et Bunch(2007).

L'objectif de cette partie est de décrire une suite de procédures :

(a) pour aider le groupe d'experts à atteindre une compréhension commune des niveaux du CECRL ;

(b) pour vérifier que la compréhension commune est vraiment atteinte ;

(c) pour maintenir cette norme dans le temps.

Les indications qui suivent s'appuient sur les expériences décrites dans les rapports décrivant comment les différentes approches et procédures ont été appliquées lors de l'expérimentation du manuel, ainsi que sur les publications disponibles.
La formation à la standardisation liée aux niveaux du CECRL comprend quatre étapes :

- effectuer les activités de familiarisation décrites dans le chapitre 3 ;

- travailler avec des performances et des tâches d'examens représentatives afin d'atteindre une compréhension adéquate des niveaux du CECRL ;
- transmettre une compétence à relier les tâches d'examens locaux et des performances à ces niveaux ;
- s'assurer que cette compréhension est partagée par l'ensemble du groupe et se déroule de façon cohérente.

Avant de commencer la formation, le facilitateur/coordinateur désigné (appelé désormais coordinateur) doit lire attentivement ce manuel et prendre en compte les ouvrages de référence recommandés et considérés comme étant pertinents dans ce contexte.

Afin de faciliter la visualisation du travail de formation à la standardisation, on trouvera un Tableau récapitulatif (Tableau 5.5) à la fin de ce chapitre. Les institutions peuvent utiliser le Tableau 5.5 afin de faire une estimation du montant du budget à prévoir pour l'ensemble du processus. Le tableau peut également servir d'aide-mémoire fonctionnel aux coordinateurs pour planifier et contrôler le processus.

L'ordre dans lequel sont présentées les étapes du processus de standardisation n'est pas aléatoire. La formation avec des échantillons de performances orales et écrites - qui sont évaluées directement - est plus aisée pour les participants que la formation avec les items de réception orale et écrite. La réception écrite est l'aptitude la plus difficile à évaluer et devrait donc être traitée à la fin. Plusieurs études de cas lors de l'expérimentation du manuel montrent un niveau d'accord entre les experts et un éventail de résultats plus réduit avec des échantillons de production qu'avec des items de réception. Nous considérons que cet ordre est le plus efficace et le recommandons mais il est bien entendu possible de le modifier selon les besoins et les contraintes de la situation.

Des directives détaillées pour la planification, incluant des tableaux représentatifs, des chiffres et des documents se trouvent dans le chapitre 13 : organiser des activités de définition points de césure, par Cizek et Bunch (2007).

Une fois la formation terminée et un consensus adéquat obtenu sur l'évaluation des échantillons illustratifs (avec une fourchette s'étalant au maximum sur un niveau et demi A2+ à B1+), le travail de calibrage (échantillons de production) ou la définition de points de césure (pour des examens indirects avec des résultats à base de notes) peut commencer, avec des performances d'apprenants locaux.

5.3. Planification préalable

Le coordinateur est responsable :

- de la logique à suivre, basée sur ce manuel et sur les références appropriées ;

- des décisions quant aux types d'expertises auxquels il faut faire appel, quant aux personnes à impliquer et à leurs rôles ainsi que l'étape du processus à laquelle elles interviendront ;

- des décisions quant au nombre et à la composition du groupe d'évaluateurs. Un groupe de douze à quinze personnes est un minimum. L'expérience tirée de l'expérimentation du manuel et d'autres projets de définition de points de césure montre qu'il est intéressant de faire appel à des évaluateurs externes à l'institution ainsi qu'à des experts/parties prenantes représentant des points de vue différents.

- de la mobilisation d'experts locaux habitués à :

    - travailler avec le CECRL ;
    - produire des programmes et des spécifications d'examens ;

- évaluer des capacités langagières de production en fonction de critères définis ;
- concevoir des examens de langue et rédiger des items;
- coordonner et former des groupes d'enseignants et d'examinateurs ;

- de la collecte de copies d'échantillons représentatifs du CECRL et de la documentation appropriée ;

- des instructions qu'ils donneront pour recueillir, dans un format défini localement, le matériel qui sera utilisé :
  - les échantillons locaux d'écrits et les vidéos de performances orales d'étudiants qui seront utilisées pour calibrer les performances locales sur des échantillons standards du CECRL et sur le CECRL lui-même ;
  - les tâches d'examens locaux qui serviront de documents de travail dans les stages sur l'évaluation.

- de la décision d'utiliser ou non les niveaux plus du CECRL. Des descripteurs calibrés pour les niveaux A2+,B1+ et B2+ sont disponibles ;

- de la préparation, de l'élaboration et de la reproduction du matériel qui sera utilisé aux différentes étapes de la démarche (voir tableau 5.5 pour les détails) :

  - les descripteurs de niveaux du CECRL ;
  - les tableaux du CECRL et les outils d'évaluation (par exemple le tableau 3 du CECRL – tableau C2 du manuel) [7] **;**
  - une sélection d'échantillons de performances et de tâches représentatives du CECRL [8]
  - une sélection d'échantillons de performances et/ou d'items d'examens locaux ;
  - les fiches de compte rendu et les documents utilisés pour recueillir l'information sur les stages.

- de la vérification du nombre de salles disponibles pour les travaux de groupes ainsi que des moyens - tables et matériel audio pour pouvoir travailler sur des échantillons d'écrits ou des items de réception orale ;

- du recueil et de l'analyse des données venant des stages de formation à la standardisation, de la présentation et de la reproduction de résultats significatifs (par exemple, la difficulté empirique de la valeur des items ; les évaluations d'échantillons par d'autres groupes) afin de les réutiliser éventuellement dans d'autres stages au moment approprié ;

- de l'organisation même des stages de façon la plus adaptée à la situation locale. Le coordinateur devra décider du nombre de participants par stage ainsi que des dates et du programme les plus appropriés. Cela comprend :

  - une décision concernant le statut des participants (enseignants/examinateurs/rédacteurs d'items), les stages auxquelles ils

[7] Tableau 3 CECRL : les niveaux communs de référence : aspects qualitatifs de l'utilisation du langage parlé anglais pages 28-29 ; français : page 28)

[8] Merci de vous reporter à la liste actualisée de matériel disponible sur le site www.coe./int/portfolio Vous y trouverez des échantillons de production écrite et orale d'adultes en allemand, anglais, français et italien. L'espagnol est prévu ultérieurement. Un deuxième CD de tâches et d'items d'examens est en préparation ; ce CD comprend un large éventail de matériel venant des études de cas lors de l'expérimentation du manuel. A la suite du séminaire de juin 2008 de calibrage de performances orales organisé au CIEP, un DVD a été édité avec des performances calibrées, de jeunes de 16 à 18 ans en cinq langues en parallèle : allemand, anglais, espagnol, français et italien.

participeront, et les implications sur la préparation des stages selon le public concerné ;

-
- la nécessité de s'assurer d'une bonne ambiance et du regroupement d'experts adéquat ;

- la planification appropriée du temps (voir ci-dessous) afin de donner l'occasion d'une réflexion et d'une discussion vaste et approfondie qui contribuera à l'obtention d'un consensus sur les évaluations ;

- le résumé des conclusions.

- de l'organisation de la documentation et du compte rendu du travail effectué durant les stages de formation afin d'assurer la crédibilité du système et de fournir un support servant à la diffusion de stages et à des sessions ultérieures.

- de la planification du contrôle continu, de la diffusion et des actions de suivi.

La durée nécessaire

Elle dépendra :

- du degré d'expertise des participants : participation éventuelle à des stages d'évaluation ;
- de leur familiarisation avec les échelles d'évaluation ;
- de leur expérience dans la rédaction d'items de production écrite et l'estimation du degré de difficulté d'un item ou d'une tâche ;
- du degré de familiarisation et d'une pratique préalable par exemple avec la grille « néerlandaise » du CECRL

Avec des participants expérimentés, il est possible d'assurer en une journée la formation pour les capacités de production, le matin étant consacré à la production orale et l'après-midi à la production écrite. Le jour suivant, il est possible de commencer à travailler sur les échantillons de performances. Autre possibilité : consacrer le premier jour à la formation et aux activités de standardisation de performances orales et le jour suivant aux productions écrites. Chaque jour doit débuter par des performances représentatives standardisées (le matin), et continuer l'après-midi avec des échantillons locaux.

La durée nécessaire à la formation pour les capacités de réception ne dépendra pas seulement de la familiarité que les participants ont avec le processus de notation, de sélection et de rédaction d'items et de tâches d'examens, de la quantité d'informations en retour qu'ils ont reçu sur les difficultés des items/tâches, mais aussi du nombre de capacités à évaluer. Il est possible d'appliquer le format décrit ci-dessus pour la production orale et écrite de chaque capacité. Si – comme cela est recommandé dans le manuel- la première capacité de réception est la réception écrite, une formation avec des échantillons d'items d'examens représentatifs peut avoir lieu le matin et peut être suivie de l'évaluation d'items d'examens locaux.

5.4. Animation des stages

La formation devrait se dérouler au cours de stages pendant lesquels les participants se familiarisent avec le CECRL, analysent et évaluent des performances ou des items d'examens et parviennent à un consensus sur le placement à un niveau du CECRL.
Pendant les stages, le coordinateur désigné doit :

- s'assurer que les participants arrivent à une bonne compréhension de ce qu'est le CECRL et vérifier jusqu'à quel point ils prennent conscience de la contribution du CECRL à l'amélioration de leur travail. On utilisera dans ce but les activités de Familiarisation du Chapitre 3 ;

- s’assurer, lors de l’évaluation d’échantillons de performances, qu’une progression logique est suivie afin de parvenir à un consensus et de le renforcer :

    - amorce et exemplification ;
    - évaluation individuelle ;
    - évaluation en tandem ;
    - discussion en grand groupe ;

- recueillir des informations et donner régulièrement un retour d’informations de manière aussi claire et visuelle que possible ;

- vérifier, comme cela est précisé dans les instructions, qu’un consensus satisfaisant sur l’interprétation des niveaux du CECRL est atteint, d’une part en ce qui concerne les descripteurs et d’autre part en ce qui concerne les performances ou les tâches qui les rendent opérationnels.

A l’issue de la formation, les coordinateurs ont la responsabilité de s’assurer que les participants ont à leur disposition tout le matériel nécessaire avant le début du processus de calibrage/définition de points de césure.

5.4.1. Arriver à un consensus et le vérifier

Tout au long du stage, on recommande aux coordinateurs de susciter les commentaires et les discussions et de faire une synthèse des évaluations en tenant compte du contexte afin de parvenir à un véritable consensus.

Comme dans tout stage de formation d’évaluateurs, on demande aux stagiaires d’évaluer le niveau correct d’un échantillon standard. Ce niveau est connu du coordinateur, mais n’est pas donné aux stagiaires avant leur évaluation. Il ne sera donné qu’à la fin du stage, par le coordinateur. Contrairement aux activités de calibrage et de définition de points de césure qui suivent, dans cette étape le groupe n’est pas invité à trouver un consensus sur le niveau sans tenir compte de preuve antérieure, mais doit plutôt arriver à la réponse correcte déjà trouvée en appliquant les critères.

Cela suppose un certain savoir faire de la part du coordinateur qui doit (a) conduire le groupe vers la réponse correcte au cours de ces expériences initiales importantes et, (b) éviter de mettre en cause les participants trop sévères ou trop indulgents dans leur interprétation avant qu’ils n’aient eu le temps de s’investir dans la formation – car cela pourrait les perturber et déstabiliser leurs jugements ultérieurs. Il ne faut pas sous estimer le temps que demande cette évolution. Il est essentiel de prendre tout le temps nécessaire à la formation avant de passer au travail sur les échantillons locaux.

Il y a deux écoles sur la façon de conduire le groupe vers le consensus qui convient.

La première est une approche qui **prend en compte les sensibilités** où l’on évite d’embarrasser les participants en respectant l’anonymat des évaluations. Cette approche garantit que les participants qui enregistrent leur évaluation individuelle avant la discussion, ne sont pas « intimidés » et que le consensus qui émerge progressivement est un consensus authentique. Avec cette approche, les individus sont influencés par les évaluations des autres : si un(e) participant(e) est « marginal(e) », il/elle s’en rend compte et peut se recentrer.

- La discrétion est également préservée si le coordinateur distribue des bulletins d’évaluation sans faire de commentaires. Si l’on veut identifier les évaluateurs dans un recueil de données pour des analyses ultérieures on peut utiliser des surnoms (Astérix ou Mickey par exemple) ou des numéros de code imprimés sur les bulletins. La projection rapide des bulletins anonymes

au rétro projecteur ou dans un tableau de synthèse expose les « marginaux » sans les gêner – à moins qu'ils ne décident d'argumenter !

- Le vote électronique peut être utilisé avec les mêmes effets. Les séminaires de calibrage qui ont donné lieu à l'édition des DVD allemands, français, italiens et portugais ont choisi cette approche. Le vote s'est fait en 2 fois : un vote individuel avant les discussions et un vote après les discussions pour confirmer le consensus.

La deuxième école préconise une **approche plus vigoureuse** : les opinions divergentes doivent s'exprimer et faire l'objet d'un débat si l'on veut parvenir à un vrai consensus. Le consensus sera ici plus délibéré, résultat d'une discussion argumentée - et peut être obtenu par un orateur convaincant. C'est la raison pour laquelle il est bon que l'animateur s'assure que les participants connaissent les échantillons standards et la raison pour laquelle on a attribué tel niveau leur a été attribué ainsi que leur rapport avec les descripteurs.

Les participants apprécient le travail en tandem ou en petits groupes. Le coordinateur peut passer d'un groupe à un autre et écouter les discussions, ramener éventuellement un groupe dans la bonne direction et demander qu'un compte rendu soit fait par un rapporteur de chaque groupe. L'avantage principal du travail en tandem ou en petits groupes est d'obliger de façon naturelle les participants à utiliser les critères définis pour justifier leurs jugements. La façon la plus simple d'enregistrer les résultats du groupe est de les recueillir au fur et à mesure et de les présenter, sur une grille, au rétroprojecteur.
Quel que soit le type d'approche choisie, le coordinateur devra calculer le pourcentage de participants qui s'accordent sur les différentes évaluations ou les coefficients de corrélation entre évaluateurs. Le coordinateur devra décider de l'opportunité de communiquer ces chiffres aux participants s'il considère que cela contribue à la formation et à une meilleure convergence des évaluations.

Il est également intéressant de présenter un schéma de dispersion des évaluations. Des graphiques sont facilement produits avec le vote électronique. Une autre façon de faire est de saisir les évaluations dans une source de données reproduites dans un histogramme formaté avec Microsoft Excel. Une troisième méthode consiste à utiliser les boîtes à moustaches produites par le programme d'analyses d'épreuves SPSS.

5.5. Formation avec des performances orales et écrites

Il se peut que des échantillons de performances et/ou de tâches d'examens représentatifs ne soient pas encore disponibles dans la langue concernée. Dans ce cas, nous recommandons de travailler avec les échantillons de la langue que le groupe a en commun – à condition que les groupes aient un niveau minimum B2/C1 de compétence dans cette langue. Dans ce cas, il faut indiquer dans la documentation qu'il s'agit d'une formation indirecte.

La première étape de la démarche est l'analyse et l'évaluation de performances orales représentatives du CECRL. Elle est suivie (si cela convient) par des performances écrites représentatives. La majorité des échantillons oraux ont un format identique qui comprend, pour chaque candidat, une phase de production orale (un monologue suivi au cours de laquelle un candidat explique quelque chose à un autre candidat qui lui pose des questions) suivie d'une phase d'interaction (au cours de laquelle les deux candidats discutent d'un sujet de façon spontanée)[9]
Pour l'évaluation de la performance écrite, il est important d'examiner des échantillons à la fois d'interaction écrite (par exemple, des notes, des lettres) et de production écrite (par

[9] Ce format a été adopté pour le projet Suisse de recherche qui a élaboré l'échelle de descripteurs du CECRL et qui est montré dans le DVD d'origine pour l'anglais comprenant des performances de ce projet. Cette approche, qui ne correspond pas à une situation d'examen, évite les effets produits par l'examinateur. Il a été adopté par les concepteurs du DVD d'apprenants adultes en français, italiens et portugais et pour le DVD du conseil de l'Europe/CIEP avec des apprenants adolescents en allemand, anglais, espagnol, français et italien.

exemple des descriptions, des histoires, des critiques) d'un candidat. C'est plus particulièrement important pour les niveaux élémentaires.

Il est important de noter que dans les échantillons représentatifs, c'est la compétence du candidat dans son ensemble, à partir de la performance dans sa totalité que l'on évalue, et non pas les performances séparées (monologue/interaction). Dans la documentation, on trouve des raisons argumentées justifiant tel ou tel niveau d'un candidat, avec des références explicites aux critères du CECRL (CECRL tableau 3/Tableau C2 pour la performance orale ; tableau B4 pour la performance écrite).Cela signifie que les tâches d'évaluation ont pour but de générer des échantillons représentatifs et complémentaires de la capacité du candidat à s'exprimer oralement dans la langue. Sur la base de *toutes* les preuves à disposition, l'expert utilise les descripteurs génériques critériés (CECRL tableau 3/tableau C2) pour juger de la compétence du candidat dans la mesure où elle peut être déduite d'un échantillonnage immanquablement limité et incomplet. Le résultat – la compétence apparaissant à travers la performance – est ce qu'on appelle habituellement en français la compétence.

5.5.1. Performance orale

Il est essentiel pour ce stage, que les participants utilisent une grille d'évaluation comportant les descripteurs du CECRL, telles que celles fournies dans l'annexe B. Nous recommandons fortement l'utilisation du tableau 3 du CECRL[10] (indiqué comme étant le tableau C2). De plus, les experts peuvent considérer comme étant utiles :

- une échelle globale simplifiée d'après le tableau 3 du CECRL (tableau C1) ;
- des copies de la grille supplémentaire basée sur le tableau 3 du CECRL (tableau C3) si les niveaux plus sont utilisés ;
- les échelles de descripteurs du CECRL pour l'interaction et la production générales ;
- l'échelle du CECRL pour la maîtrise phonologique, si cela s'avère nécessaire[11] ;
- une fiche standard d'évaluation pour noter leurs commentaires et le niveau attribué à chaque performance (exemples donnés dans les Fiches C2 et C3

Ce stage est organisé en trois étapes :

Phase 1 : Illustration. Le coordinateur commence le stage par deux ou trois performances orales représentatives pour exemplifier les niveaux. Il projette l'échantillon et invite les participants à commenter la performance avec leurs voisins. Au moment opportun, le coordinateur reconstitue le grand groupe et lui fait expliciter pourquoi cette performance illustre le niveau décrit sur la grille du Tableau 3 du CECRL (Tableau C2) et non un niveau inférieur ou supérieur.

On recommande de passer toute la séquence de l'échantillon, même si cela doit prendre 15 minutes. La performance d'un candidat dans la phase de l'interaction peut être très différente (en mieux ou moins bien) de sa performance dans la phase de production et – comme cela est mentionné dans l'introduction, c'est l'ensemble de la compétence dans la capacité concernée qui doit être évaluée – et non une de ses performances.
La sélection des échantillons : les recommandations suivantes prennent en compte l'expérience tirée de l'expérimentation du manuel et de l'animation des stages qui ont donné lieu à des DVD avec des échantillons représentatifs et les projets qui s'y rapportent.

[10] CECRL Tableau 3 : niveaux communs de référence ; aspects qualitatifs de l'utilisation de la langue parlée pour l'anglais page 28-29 ; le français page 28.
[11] La prononciation ne fait pas partie du tableau 3 du CECRL parce qu'il a été conçu pour une utilisation dans des contextes internationaux et les évaluateurs habitués à travailler dans un contexte monolingue, national peuvent avoir tendance à se laisser influencer par leur manque de familiarité avec des accents de personnes parlant d'autres langues maternelles.

- Il est judicieux de commencer par les niveaux B1 ou B2 et de montrer des échantillons de performances à des niveaux très proches pour stimuler la discussion sur les limites entre deux niveaux, en se référant aux critères (CECRL tableau 3/manuel tableau C2).

- Le premier de ces exemples doit présenter une performance de profil relativement « plat » parmi les catégories du Tableau 3 du CECRL/ Tableau C2 du Manuel - à savoir, un locuteur qui serait par exemple de niveau B1 dans toutes les catégories Etendue, Correction, Aisance, Interaction, Cohérence.

- Un de ces échantillons standards introductifs devrait montrer un profil moins régulier, par exemple si le locuteur est au niveau B1 dans certaines catégories mais en B2 ou au moins en B1+ dans d'autres. Si la question des « profils inégaux » n'est pas traitée assez tôt au cours de la formation, elle peut poser ultérieurement des problèmes.

  Afin de mettre en évidence le fait que certains candidats peuvent avoir des profils très inégaux et qu'il faut examiner séparément les différents aspects qualitatifs (Etendue, Correction, Aisance, Interaction, Cohérence), les coordinateurs peuvent envisager d'évaluer plusieurs performances en ne prenant en compte *qu'un seul aspect*. Cela neutralise la tendance naturelle des évaluateurs à laisser leur impression générale avoir une influence sur leur jugement dans chaque aspect (« effet de halo »).

L'utilisation des instruments de mesure : Les conseils suivants prennent en compte l'expérience tirée de l'expérimentation du manuel, l'animation des stages qui ont donné lieu à des DVD avec des échantillons représentatifs et les projets qui s'y rapportent.

- On peut demander aux participants d'utiliser d'abord uniquement l'échelle globale (Tableau C1) qui simplifie la grille du Tableau 3 du CECRL (Tableau C2 du manuel) afin de se rendre parfaitement compte de leur impression globale sur le niveau des candidats avant d'examiner les catégories de la grille du Tableau 3 du CECRL (Tableau C2 du manuel).

- S'étant fait une première impression sur le niveau des performances, ils devraient alors consulter les descripteurs détaillés de ce niveau sur la grille du Tableau 3 du CECRL (Tableau C2), lire les descripteurs pour les niveaux immédiatement supérieurs et inférieurs de chaque catégorie et utiliser la grille pour tracer le profil de la performance du candidat.

- Si les niveaux « plus » sont utilisés, les participants devraient à ce moment là consulter la grille supplémentaire (tableau B3) pour décider si le candidat est un exemple « fort » du niveau – une performance « de niveau plus ».

- Ils devraient alors utiliser les descripteurs du tableau 3 (tableau C2) et si nécessaire la grille des niveaux « plus » supplémentaires (tableau C3) comme fil directeur de la discussion avec leur voisin.

- Au cours de la discussion, les participants voudront peut être consulter aussi les échelles supplémentaires de descripteurs mentionnées ci-dessus.

- **Phase 2 : Pratique***.* Dans cette seconde phase, le rôle du coordinateur est d'aider les stagiaires à voir s'ils ont encore tendance à être trop sévères ou trop indulgents. Si le vote s'est fait par bulletin, le coordinateur utilisera une fiche de synthèse sur transparent (par exemple la Fiche B3) ou un graphique pour enregistrer les évaluations. Tout au long de cette étape, le coordinateur doit faire visualiser aux participants leur comportement en tant que groupe et animer la discussion comme indiqué plus haut, sans embarrasser les individus. Si l'on n'a pas utilisé le vote anonyme, une technique efficace consiste à écouter les discussions des groupes et, lorsque tout le monde est regroupé, à faire donner « la réponse » par les groupes avec lesquels on a la meilleure chance qu'elle soit correcte.

On recommande au coordinateur de mener une discussion au niveau de tout le groupe pour justifier de l'attribution d'un candidat à tel niveau plutôt qu'au niveau supérieur ou inférieur, en se référant de façon explicite aux critères des descripteurs. Cela évite que des participants réutilisent des notions préétablies des niveaux du CECRL (souvent de simples traductions d'un autre système) et montre la nécessité de prendre comme seule référence les critères des descripteurs.

La sélection des échantillons : on recommande là aussi l'utilisation de deux à trois échantillons.

L'utilisation des instruments de mesure
Avant le stage, les coordinateurs doivent décider s'ils continueront à utiliser l'échelle globale (Tableau B1) après l'étape d'Illustration. Elle est utile en ce sens (a) qu'elle donne aux participants un point de départ pour la lecture de la grille (Tableau 3 du CECRL ; Tableau C2 du Manuel) et, (b) qu'elle aide les participants à faire la part de leur première impression par rapport à un jugement réfléchi – en particulier si l'on a consigné séparément les deux réactions comme dans la fiche d'enregistrement proposée (Fiche C2)

- **Phase 3 : Evaluation individuelle.** Les stagiaires évaluent individuellement le reste des performances, rendent leurs bulletins d'évaluation et discutent ensuite sur ce que représentent les niveaux du CECRL auxquels ces performances ont été affectées. On recommande vivement de continuer à analyser les performances par blocs de trois. De la sorte, on focalisera mieux la discussion sur la standardisation– plutôt que d'entrer dans une discussion sur les mérites de certaines performances. Le dernier bloc de trois devrait faire l'objet d'un accord presque général. La grande majorité des participants devrait en effet être d'accord sur le niveau avec une dispersion inférieure à un niveau et demie. Par exemple, pour une performance que l'on s'accorde à situer en B1+, la dispersion des résultats ne devrait pas excéder l'éventail de B1 à B2 ; pour une performance que l'on s'accorde à placer en B1, la dispersion devrait être de A2+ à B1+.

  Le stage se terminera lorsqu'on aura atteint ce niveau d'accord dans le groupe et que le coordinateur (et les participants) seront satisfaits du degré de consensus atteint pour l'évaluation d'échantillons standards de performances orales.

  On recommande là aussi l'utilisation de deux à trois échantillons.

Avant le stage, les coordinateurs doivent décider s'ils continuent à utiliser l'échelle globale (Tableau C1) après l'étape d'Illustration. Elle est utile en ce sens (a) qu'elle donne aux participants un point de départ pour la lecture de la grille (Tableau 3 du CECRL ; Tableau C2) et, (b) qu'elle aide les participants à faire la part de leur première impression par rapport à un jugement réfléchi – en particulier si l'on a consigné séparément les deux réactions comme dans la fiche d'enregistrement proposée (Fiche C2). Cependant, il peut sembler plus simple d'éliminer une des deux fiches avec lesquelles les experts travaillent. L'expérience montre qu'une fois que les experts se sont habitués à utiliser le tableau 3 du CECRL (tableau C2), ils n'ont à vrai dire pas besoin de l'échelle (tableau C1) pour arriver à une première impression globale.

La sélection des échantillons**:** on recommande qu'au moins une performance par niveau du CECRL soit analysée, évaluée et discutée au cours du stage**.**

L'utilisation des instruments de mesure
Pendant la discussion, le coordinateur décide si l'utilisation d'autres échelles orales du CECRL et la justification de façon plus détaillée du niveau attribué. contribuent à une meilleure compréhension du niveau.

5.5.2 Performance écrite

On recommande une démarche semblable à celle qui a été préconisée pour la performance orale.

La grille d'évaluation à laquelle se reporter est le Tableau C4 de la section C de l'annexe. Cette grille est un prolongement du tableau 3 du CECRL, avec deux colonnes sur la Description et sur l'Argumentation qui ne doivent être utilisées que pour les textes de ce type.

- **Phase 1 : Illustration** : Le stage commence par deux ou trois performances écrites standards que le coordinateur utilise pour exemplifier les niveaux. Pour chaque échantillon, à un moment donné, le coordinateur reconstitue le grand groupe et lui fait expliciter comment cette performance illustre le niveau décrit sur la grille du Tableau C4 et pourquoi elle n'est pas du niveau inférieur ou supérieur.

Tableau 5.1 : Gestion du temps pour l'évaluation d'échantillons de performance orale

Nombre recommandé de participants : 30 participants au maximum	
Première étape : Familiarisation	60 minutes
Deuxième étape : Travail sur des échantillons standards :	
Phase 1 : Illustration avec environ trois performances standards	60 minutes
Pause	
Phase 2 : Pratique sous contrôle du coordinateur avec environ trois performances standards	60 minutes
Phase 3 : Etape libre avec environ trois performances standards	60 minutes
Déjeuner	
Troisième étape : Calibrage des échantillons locaux :	
Evaluation individuelle et discussion de groupe sur environ trois performances	60 minutes
Evaluation individuelle d'environ cinq performances supplémentaires	60 minutes
Pause	
Planification du suivi et mise en réseau	60 minutes
Synthèse et clôture	30 minutes

Documents et matériel à préparer

Photocopies pour tous les participants :

- *Grille d'évaluation du CECRL Tableau 3/du Manuel Tableau C2*
- *Echelle d'évaluation simplifiée ci-dessus : Tableau C1 (si nécessaire)*
- *Grille de « niveaux plus » en supplément du tableau 3 (si nécessaire)*
- *Feuilles d'évaluation pour les participants : exemples des Fiches C2 – C3*
- *Choix et copies d'échelles complémentaires pertinentes ou des Tableaux A1-A3*

Auxquels s'ajoutent :

- *des vidéos de performances standards ;*
- *le manuel ;*
- *des fiches de synthèse pour le coordinateur et des transparents (Fiche B 4) ;*
- *des vidéos locales (enregistrées et/ou sélectionnées selon les instructions des Etudes de cas).*

La sélection des échantillons

- Le premier de ces exemples représentatifs doit présenter une performance de profil relativement « plat » parmi les catégories du Tableau B4 (à savoir, un locuteur qui serait par exemple de niveau B1 dans toutes les catégories Etendue, Correction, Aisance, Interaction, Cohérence et également d'un bon niveau pour décrire et argumenter).

- Comme pour les exemples de performance orale, le coordinateur peut envisager d'évaluer plusieurs performances écrites dans une même catégorie afin que les participants se rendent compte de « l'effet de halo ».

- On recommande que l'un de ces premiers échantillons montre un profil moins régulier, par exemple que le locuteur soit au niveau B1 dans certaines catégories mais en B2 ou au moins en B1+ dans d'autres. Si la question des « profils inégaux » n'est pas traitée assez tôt au cours de la formation, elle peut poser ultérieurement des problèmes.

L'utilisation des instruments de mesure

Le coordinateur invite les stagiaires à lire les textes et à examiner la performance par rapport aux critères du Tableau C4.

- **Phase 2 : Pratique** : Dans cette seconde phase – où l'on utilisera de nouveau trois échantillons – le rôle du coordinateur est d'aider les stagiaires à voir s'ils ont encore tendance à être trop sévères ou trop indulgents. Si le vote s'est fait par bulletin, le coordinateur utilisera une fiche de synthèse (par exemple la Fiche C3) pour rapporter les évaluations sur transparent.

  Tout au long de cette phase, le coordinateur doit faire visualiser aux participants leur comportement en tant que groupe et animer la discussion comme indiqué plus haut, sans embarrasser les personnes. Si l'on n'a pas utilisé le vote anonyme, une technique efficace consiste à écouter les discussions des groupes et, lorsque tout le monde est regroupé, à faire donner « la réponse » par les groupes avec lesquels on a la meilleure chance qu'elle soit correcte.

  - **Phase 3 : Evaluation individuelle.** Les stagiaires évaluent individuellement le reste des performances et discutent ensuite des niveaux du CECRL auxquels ces performances ont été affectées.

    On recommande très vivement de continuer à analyser les performances par blocs de trois. De la sorte, on focalisera mieux la discussion sur la standardisation – plutôt que d'entrer dans une discussion sur les mérites de certaines performances. Le dernier bloc de trois devrait faire l'objet d'un accord. C'est-à-dire que la grande majorité des participants devraient être d'accord sur le niveau avec une dispersion inférieure à un niveau et demie. Par exemple, pour une performance que l'on s'accorde à situer en B1+, la dispersion des résultats ne devrait pas excéder l'éventail de B1 à B2 ; pour une performance que l'on s'accorde à placer en B1, la dispersion devrait être de A2+ à B1+.

    Le stage se terminera lorsqu'on aura atteint ce niveau d'accord dans le groupe.

La sélection des échantillons**:** comme pour les performances orales, on recommande qu'au moins une performance par niveau du CECRL soit analysée, évaluée et discutée au cours du stage**.**

L'utilisation des instruments de mesure

Comme au cours des discussions sur les échantillons de production et d'interaction orales, le coordinateur peut décider d'utiliser des échelles spécifiques (par exemple, production écrite générale, écriture créative, essais et rapports) pour faciliter un accord et mieux justifier l'attribution d'un niveau donné. Le coordinateur peut également distribuer les Tableaux A2 et A3, en parallèle avec le Chapitre 4 sur la Spécification.

Tableau 5.2 : Gestion du temps pour l'évaluation d'échantillons de performance écrite

Nombre recommandé de participants : 30 participants au maximum	
Activités d'introduction (Familiarisation)	60 minutes
Travail sur des échantillons standards : *Phase 1 : Illustration avec environ trois performances représentatives* *Pause* *Phase 2 : Pratique sous contrôle du coordinateur avec environ trois à cinq performances représentatives* *Phase 3 : Etape libre avec environ trois à cinq performances représentatives*	60 minutes 60 minutes 60 minutes
Déjeuner	
Calibrage des échantillons locaux : *Evaluation individuelle de performances d'un niveau élevé, moyen et faible et discussion de groupe* *Evaluation individuelle d'environ cinq performances supplémentaires*	60 minutes 60 minutes

Tableau 5.3: Documents et matériel à préparer pour l'évaluation de la production écrite

Documents et matériel à préparer

Photocopies pour tous les participants :

- *Grille d'évaluation (Tableau C4)*
- *Feuilles d'évaluation pour les participants (exemples des Fiches C2 et C3*
- *Choix et photocopies des échelles complémentaires pertinentes*

Auxquels s'ajoutent :

- *des textes de performances standards*
- *des fiches de synthèse pour le coordinateur et des transparents (Fiche C 4)*
- *des textes produits localement (enregistrés et/ou sélectionnes selon les instructions des Etudes de cas)*

5.6. Formation avec des tâches et des items de capacités de réception (écrite et orale) et de compétences linguistiques

L'objectif des activités décrites dans cette partie est de s'assurer que les participants puissent établir le lien entre leur interprétation des niveaux du CECRL et les items d'examens représentatifs afin de pouvoir ultérieurement utiliser cette compréhension commune pour :

- relier les épreuves ou les items pertinents produits localement aux niveaux du CECRL

- acquérir, comme une plus value, une compétence pour l'élaboration d'items d'examens pouvant éventuellement être considérés comme étant reliés aux Niveaux du CECRL.

Les techniques décrites peuvent être utilisées pour les items et les tâches d'examens évaluant des capacités de réception et peuvent être, le cas échéant, transférées à

l'évaluation d'autres aspects de l'utilisation de la langue tels que la grammaire et le vocabulaire.

Les tâches qui impliquent des capacités intégrées (par exemple, écouter un texte et répondre à des questions puis utiliser l'information donnée pour faire un résumé) devront être considérées du double point de vue de la difficulté des aspects réceptifs et productifs de la tâche. Il y a généralement une différence délibérée de difficulté entre les deux parties de la tâche, et il faut traiter cette question au cours de la formation. La difficulté des items peut varier (et on peut la faire varier systématiquement si on le souhaite) en fonction du texte lu ou écouté, de l'aptitude à la compréhension que l'on teste et de la réponse que le candidat doit donner pour manifester sa compréhension.

Comme pour les échantillons de performance, une formation avec des tâches et des items représentatifs affectées de valeurs de difficulté connues doit être d'abord mise en place et suivie ensuite du processus d'analyse d'items produits localement (Chapitre 6).

La formation avec des items calibrés prépare, dans l'ordre suivant, à :

1. se rendre pleinement compte de l'étendue des sous échelles de descripteurs du CECRL pour des domaines particuliers disponibles dans le CECRL (voir chapitre 4);

2. identifier la pertinence du contenu des items analysés en fonction de ce que recouvre le construit par rapport aux niveaux et aux échelles du CECRL. Comme cela est mentionné dans la partie 4.3.2, les recherches du projet hollandais de construit du CECRL (Alderson et al 2006) et la grille d'analyse de contenu du CECRL qui en est résulté pour les réceptions orale et écrite peuvent être très utiles.

3. estimer le niveau de chaque tâche et item en fonction des descripteurs pertinents du CECRL ;

4. examiner les raisons possibles de divergences entre les niveaux estimés et les niveaux établis empiriquement ;

5. confirmer le niveau de difficulté en les confrontant aux données empiriques.

Il est essentiel de commencer la formation avec la réception écrite. De même qu'il est plus facile de travailler sur des performances orale et écrite (que l'on peut observer en direct) que de travailler sur des compétences de réception (qu'on ne peut pas observer), il est de loin beaucoup plus facile d'organiser un travail de groupe sur la lecture et la relecture de textes et d'items imprimés (que l'on peut voir) que d'écouter à de nombreuses reprises des items et des textes (que l'on ne peut pas observer).

Une fois le processus d'évaluation des items de réception écrite achevé, il sera plus facile d'organiser le stage sur la capacité à la réception orale et de travailler sur des textes de réception orale car les stagiaires auront déjà l'habitude de la tâche à accomplir.
Le coordinateur doit décider de l'organisation des stages et estimer leur durée qui dépendra du contexte et de la formation antérieure des participants.

5.6.1 Familiarisation nécessaire

Même si les stagiaires ont déjà participé au stage de Familiarisation décrit dans le Chapitre 3, il est nécessaire d'organiser une activité consistant à trier les descripteurs de la capacité que l'on étudie avant de commencer l'évaluation de la difficulté et la définition des points de césure.

Le CECRL fournit des échelles globales générales (par exemple « Réception », « Compréhension générale de l'écrit », « Compréhension générale de l'oral ») mais aussi des échelles spécifiques qui décrivent les différentes activités langagières de réception (par

exemple « Comprendre en tant qu'auditeur ») et de stratégies (« Reconnaître des indices et faire des déductions »).

Les coordinateurs devront faire le choix des échelles les plus pertinentes pour l'examen dans le contexte où on le passe. Le travail doit toujours commencer par l'analyse et la discussion des échelles générales (par exemple « Compréhension générale de l'oral »). Puis les coordinateurs peuvent rassembler les sous-échelles les plus appropriés au contexte pour la capacité donnée (par exemple « Comprendre en tant qu'auditeur ») ou utiliser les reformulations des descripteurs du CECRL pour l'auto-évaluation, utilisées dans le projet DIALANG (CECRL, Annexe C) et demander aux participants de trier les descripteurs selon les 6 niveaux du CECRL (voir Partie 3.2. 1, Activité f).

La standardisation des items qui évaluent des compétences linguistiques devra se faire selon une approche sensiblement différente de celle adoptée pour la réception orale ou pour la réception écrite parce qu'il est nécessaire de préciser le type de composantes que l'on peut s'attendre à trouver aux différents niveaux. Le CECRL fournit des descripteurs généraux pour des éléments de la compétence de communication langagière (CECRL Partie 5.2 ; Manuel Tableaux A1-A3) mais les spécifications linguistiques de ce type sont propres à chaque langue. La partie 4.3 passe en revue les différents outils disponibles. Le projet DIALANG a aussi élaboré un ensemble de spécifications pour 14 langues, comprenant des conseils aux rédacteurs d'items.

5.6.2. Formation à la définition des points de césure (standard setting)

Le processus de standardisation se déroule en trois étapes suivant des procédures pour la formation semblables à celles utilisées avec les échantillons de performances standards :

- **Phase 1 : Illustration** : Première évaluation du niveau d'un texte et des tâches et des items qui y correspondent. Cette activité préliminaire aidera les participants à s'entendre sur les niveaux du CECRL pour la capacité évaluée.

  Il est essentiel d'examiner à la fois le **niveau du texte d'origine** et la **difficulté de chaque item** qui l'accompagne. Un texte n'a pas un « niveau ». C'est la compétence des candidats, telle qu'elle se manifeste dans leurs réponses aux questions, que l'on peut relier à un niveau du CECRL. Ce qu'on peut dire au mieux d'un texte, c'est qu'il convient pour son utilisation dans un examen visant un niveau donné.

Tableau 5.4 : Sources de références dans le CECRL

Domaine	**Référence dans le CECRL**
Situations, catégories de contenu, domaines	Tableau 5 dans le CECRL 4.1
Thèmes de communication	Les listes, dans le CECRL 4.2
Activités de communication	Les listes, dans le CECRL 4.3
Activités de communication et stratégies	Les listes, dans le CECRL 4.4.2.2
Textes et types de textes	Les listes, dans le CECRL 4.6.2 et 4.6.3
Caractéristiques du texte : longueur de la tâche, cohérence et structure des tâches	L'information dans le CECRL 7.3.2.2
Tâches	La description, dans le CECRL 7.1,7.2 et 7.3

A ce propos, la grille d'analyse de contenu du CECRL pour la réception orale et écrite décrite dans le chapitre précédent, est un instrument très utile pour faire prendre conscience de l'importance des traits qui affectent le niveau de difficulté.

Nous recommandons aux utilisateurs de se reporter aux fiches complétées du Chapitre 4 (Spécifications) et d'examiner la difficulté du texte et de la tâche par rapport aux parties appropriées du CECRL. Pour la réception écrite, par exemple, on utilisera la Fiche A10. Le Tableau 5.4 indique les parties du CECRL auxquelles se reporter.

Cette activité doit se faire individuellement dans un premier temps. Le coordinateur devra faire prendre conscience aux stagiaires, comme pour le travail avec les performances des apprenants, des convergences ou des divergences de leurs évaluations. Les points suivants nous semblent particulièrement importants :

- Il est très important que les stagiaires lisent ou écoutent réellement le texte et répondent individuellement aux items qui s'y rapportent avant d'évaluer la difficulté des questions traitées et le niveau du CECRL qu'elles représentent le mieux.

- Après avoir répondu aux items, ils devraient être en mesure de comparer leur propre réponse et la réponse correcte (et les catégories qualitatives dans le barème des items polychotomiques) à chaque item. Pour s'assurer d'une compréhension claire du corrigé type ou du barème, il est bon qu'une discussion précède l'évaluation de la difficulté des questions.

- Il est également essentiel que le coordinateur donne des consignes claires sous forme d'instructions précises distribuées aux stagiaires. L'item est la mise en œuvre d'un descripteur de « capacité à faire » du CECRL. Le problème est donc de savoir à quel niveau l'apprenant doit se trouver pour être capable de répondre correctement ou de façon acceptable à la question..

  L'instruction précise que les évaluateurs reçoivent va dépendre de la méthode appliquée pour la définition des points de césure. L'exemple suivant se rapporte à la méthode Basket (partie 6.7.2)

  Pour des items notés 1-0 (items dichotomiques)

  « A quel niveau du CECRL un candidat peut-il donner une réponse correcte à l'item suivant ? »

  Pour des items polychotomiques :

  « A quel niveau du CECRL un candidat peut-il donner une réponse correcte à l'item suivant avec des niveaux de résultats XXX (par exemple 2,1,0) ? »

- Les participants notent individuellement leur évaluation des items, et justifient ensuite leurs décisions par deux ou en petits groupes,

- A la fin de l'activité, le coordinateur révèle « le » niveau sur lequel la ou les question(s) sont effectivement calibrées.

  - **Phase 2 : Pratique suivie** : Une fois les étapes d'illustration et de discussion achevées et un accord obtenu sur la façon d'envisager le processus, on demandera aux participants d'évaluer individuellement différents textes accompagnés des tâches et des items qui leur correspondent, de les relier aux niveaux du CECRL et d'identifier les descripteurs du CECRL que chaque tâche ou item met en œuvre.

De même que pour l'évaluation des échantillons de production orale et écrite, on peut poursuivre le travail avec 4 à 6 items ou 2 ou 3 mini tests (un texte avec plus d'une question). On demandera aux stagiaires de :

- lire les textes et répondre aux questions correspondantes ;

puis remplir une grille (voir ci-dessous) donnant leur évaluation de chaque item afin :

- d'identifier les descripteurs du CECRL que l'item met en œuvre ;
- de classer chaque question sur l'un des six niveaux du CECRL

Fiche B5 : Fiche d'évaluation des questions/items (DIALANG)

Compétence_____ _	**Descripteur mis en œuvre** *(Enumérer les sous échelles et le niveau)*	**Niveau attribué**	**Commentaires** *(En incluant les références à la Fiche A9)*
Question/item 1			
Question/item 2			
Question/item 3			
Question/item 4			
Question/item 5			
Etc.			

Le travail de groupe devrait prendre en compte les aspects suivants:

- le type d'items (réponse sélectionnée, réponse construite) et comment cela affecte la difficulté de l'item ;
- la mise en œuvre de divers descripteurs du CECRL dans le texte et dans la tâche ;
- la preuve dont on dispose pour justifier le calibrage de chaque item sur un des niveaux du CECRL ;
- les autres aspects pertinents des caractéristiques de l'item, du texte, de la réponse, portés dans la colonne « commentaires ».

Il faut remarquer à cet égard que les évaluateurs ont tendance à surestimer la difficulté des questions à réponse sélectionnée (par exemple à choix multiple), qui ont tendance à être plus faciles que les évaluateurs ne le pensent souvent. De la même façon, ils sous estiment la difficulté des réponses construites (par exemple, répondre à une question, compléter une phrase) qui ont tendance à être plus difficiles que les évaluateurs ne le pensent. Demander aux participants de répondre effectivement aux questions avant de se lancer dans des discussions sur la difficulté peut permettre de réduire le problème. Quoi qu'il en soit, se centrer sur l'interaction entre un texte et un type d'item pour définir la difficulté – en fonction de la mise en œuvre d'un descripteur du CECRL – est une sensibilisation nécessaire à ce moment de la formation.

Il nous semble utile d'attirer l'attention des participants sur le rôle de la complexité de la langue, de la longueur du passage qu'il faut examiner pour trouver la réponse correcte, de la vraisemblance des options dans les questions à choix multiples, etc., comme facteurs de la difficulté de l'item. Les coordinateurs devraient susciter à nouveau des commentaires et des discussions et faire une synthèse claire des évaluations en les présentant sous forme schématique non seulement pour que les stagiaires puissent les visualiser mais aussi pour une documentation ultérieure.

- **Phase 3** : Evaluation individuelle : Les stagiaires continuent à travailler individuellement avec le reste des items puis discutent des niveaux du CECRL auxquels ils ont été calibrés. De même que pour les performances de production écrite et orale, il est recommandé de travailler sur des groupes de 4 à 6 items. Cela permet de centrer la discussion sur la standardisation plutôt que sur les propriétés des items ou des différents textes. Le dernier groupe d'items devrait faire l'objet d'un large consensus.

On recommande aux participants de travailler de la même façon qu'avec les échantillons des performances orale et écrite (en utilisant la grille pour consigner leurs évaluations) jusqu'à ce la dispersion des résultats ne dépasse pas un niveau et demi (par exemple de A2+ à B1+).

Le coordinateur peut utiliser une fiche d'évaluation globale comme la Fiche C4 afin de reporter sur un transparent les évaluations des items faites par les stagiaires et de faire apparaître visuellement les variations de leur accord. Cette fiche sera nécessaire à la documentation.

Une fois que l'on a achevé la formation à la standardisation (Parties 5.4 et 5.5) et que l'on considère que le consensus sur l'évaluation des échantillons standards est satisfaisant, l'étape de travail sur les performances locales peut commencer. La partie qui suit (5.6.) fait un compte rendu pour chaque étape de la façon de calibrer des échantillons locaux de performances orale et écrite. Les procédures à suivre sont très semblables à celles de la formation (5.4).

Pour établir des seuils fonctionnels sur des examens conçus localement pour les réceptions écrite et orale ou pour des compétences sous jacentes, le choix des procédures de définition de standards parmi celles qui sont décrites dans le chapitre 6 de ce manuel (ou d'autres écrits sur la définition des standards) aura une influence sur les procédures à suivre. On recommande aux utilisateurs de ce manuel de lire le chapitre 6, et de choisir une méthode ou plus d'une, et, en suivant le canevas de la formation décrite dans cette partie, d'élaborer leurs propres procédures, étape par étape, qui soit appropriée au contexte. La documentation disponible pourra être très utile pour rédiger les procédures, mais il est nécessaire de prendre en compte les points décrits dans la partie suivante pour le calibrage en rapport avec la sélection d'items, l'analyse de données et la documentation.

5.7. De la formation au calibrage

L'application de la compréhension des niveaux du CECRL au calibrage des échantillons locaux (de performances orale ou écrite) ou de tâches/d'items locaux (pour les examens portant sur la réception orale et écrite et la compétence linguistique évalués avec des notes) doit avoir lieu aussitôt que possible après la formation à la standardisation. On recommande très vivement qu'elle ait lieu au cours du même stage, l'après-midi même ou le jour suivant. Le coordinateur sera le mieux placé pour juger si cela est faisable ou s'il vaut mieux le faire plus tard.

Si le calibrage d'échantillons locaux se fait au cours d'un stage à part, on recommande, au cours d'une phase d'harmonisation de montrer aux stagiaires des extraits d'une ou deux performances standards évaluées au cours de la session précédente et on leur rappelle la discussion qui a eu lieu.

Les procédures à suivre pour le calibrage sont les mêmes que celles suivies à la formation.

5.7.1 Echantillons nécessaires

Même si cela retarde le projet, il est important d'investir du temps et de l'énergie pour recueillir un jeu d'échantillons locaux de bonne qualité. Une fois calibrés sur le CECRL, il y a des chances pour que ces échantillons prennent tout leur sens en termes de référence. On recommande en conséquence de faire un choix réfléchi des items pour garantir la qualité, la représentativité (en ce qui concerne les candidats) et le contenu couvert par l'examen.

Le processus de collecte peut être très semblable à celui du processus de production d'items :

- définition des critères de sélection ;
- identification des échantillons de candidats ;
- travail en atelier pour étudier et filtrer les échantillons en fonction de leur qualité ;

- sélection ;
- vérification de la représentativité du jeu d'échantillons sélectionnés ;
- apport éventuel d'échantillons supplémentaires pour « compléter » l'ensemble ;
- documentation des caractéristiques des échantillons pour le calibrage grâce à un outil tel que les grilles du CECRL pour les tâches de productions écrite et orale (annexe B2).

Il est essentiel que les échantillons locaux de performances utilisés pour le calibrage comprennent, pour les mêmes candidats, des discours de types différents couvrant l'éventail des activités décrites dans le CECRL.

Pour la performance orale, cela suppose une activité avec des étapes qui illustrent différents types de discours. La technique de tournage pour filmer les échantillons représentatifs a été conçue pour éviter l'influence de l'examinateur et pour fournir un échantillon équilibré à la fois de production et d'interaction orale.

Pour la production écrite, différents types de textes sont suggérés. Il est préférable que les échantillons d'écrit incluent à la fois des productions libres (par exemple, une lettre amicale, une description) et des activités plus formelles où les candidats suivent un modèle appris (par exemple, la lettre de confirmation d'une réservation d'hôtel). Ceci est particulièrement important, notamment aux niveaux élémentaires.

Il est essentiel de veiller à ce que les échantillons de productions recueillis pendant le processus de production, d'administration, d'enregistrement ou de documentation soient de bonne qualité et utilisables. Dans le cas des vidéos, cela implique un son et une image de qualité[12]; dans le cas de textes écrits cela signifie que les performances n'ont pas été biaisées par des circonstances extérieures dues à une prolongation du temps imparti, l'usage de dictionnaires, une mauvaise écriture, etc.

Comme cela a été suggéré dans la partie antérieure, le fait de compléter les grilles du CECRL de tâches écrite et orale permet de s'assurer que la sélection des échantillons est équilibrée et que les éléments de base de la documentation sont disponibles.

En général, les procédures à suivre sont celles décrites dans les parties 5.3 et 5.4 pour la formation à la standardisation avec des échantillons représentatifs. . Cela comprendra :

- l'utilisation des mêmes outils que ceux utilisés pour la formation (Tableaux C1, C2 et aussi C3 (niveaux plus) ; le tableau C4 pour les performances écrites ; les échelles du CECRL et/ou les tableaux A1, A2 et A3 pour les textes et les items de réception et de compétence linguistique) ;
- une évaluation individuelle suivie de discussion en petits groupes conduisant le grand groupe au consensus ;
- une discussion sur la dispersion dans les évaluations individuelles renouvelée jusqu'à ce que l'on parvienne à un accord acceptable (dispersion égale à un niveau et demi).

Un point important est ici à souligner : les évaluations individuelles doivent être enregistrées avant toute discussion. En fait, l'expérience des séminaires de calibrage qui ont débouché sur l'édition de DVD représentatifs montre que c'est la dispersion des évaluations qui est affectée par les discussions (les marginaux se conformant à la norme) et non la moyenne et donc le résultat. Néanmoins, le signe du succès d'un séminaire de calibrage est que l'évaluation d'individus rassemblés et le consensus final arrivent aux mêmes niveaux du

[12] Si une vidéo est ultérieurement copiée sur un « master » lui-même copié pour distribution, les utilisateurs auront alors une copie de troisième génération qui aggrave tous les défauts sonores. C'est la raison pour laquelle on recommande de *toujours* utiliser, même avec la technologie du DVD digital, un micro extérieur et *non* celui de la caméra. Avec un micro externe d'étendue moyenne (1-2 m), il est parfaitement possible d'obtenir une bonne qualité de son sans passer par un studio d'enregistrement.

5.7.2. Parvenir à un consensus et le vérifier

CECRL pour un échantillon ou un item. La publication de données non biaisées fait partie des preuves qui peuvent être fournies [13]

5.7.2. Arriver à un consensus et le vérifier

Si l'on ne parvient PAS à un accord, le coordinateur doit discuter avec les stagiaires de la raison de ce problème incompatible avec leur maîtrise de l'évaluation des échantillons représentatifs. Le coordinateur devra se prononcer sur la cause du problème et faire le nécessaire pour le résoudre. Parmi les raisons possibles et les solutions :

Problème	**Action possible**
• Les échantillons locaux ne proposent qu'une tâche et cette tâche est trop différente des échantillons du CECRL	→ Vérifier qu'un éventail assez large de discours est disponible. Trouver d'autres échantillons plus proches du CECRL
• La grille d'évaluation (Tableau C2) ne semble pas appropriée pour les échantillons (par exemple contexte professionnel, tâche étroitement définie)	→ Réviser la grille en consultant les échelles du CECRL
• Certains stagiaires commencent à appliquer d'autres normes quand ils évaluent « leurs » apprenants	→ Proposer un échantillon local et un échantillon du CECRL pour essayer de forcer les praticiens à appliquer les mêmes normes

5.7.3. Analyse des données

Les évaluations des échantillons standards du CECRL devraient être analysées statistiquement afin de (a) confirmer la relation avec les niveaux et, (b) calculer la fiabilité d'un même évaluateur (cohérence) et des évaluateurs entre eux (cohérence).

Le degré de l'accord entre les participants doit être évalué et le niveau moyen des échantillons confirmé par l'analyse des évaluations au cours du processus de calibrage. L'avantage principal est que les évaluateurs dont le comportement n'est pas cohérent peuvent être identifiés et qu'on peut exclure leurs évaluations de l'analyse.

Plusieurs méthodes permettent d'atteindre ce but. Outre les corrélations de fiabilité entre les évaluateurs il y a, par exemple, le modèle multiple de Rasch mis en œuvre dans des programmes tels que FACETS.

[13] Ce n'est pas toujours le cas pour la définition de points de césure d'examens indirects et notés. Comme la définition des points de césure est un processus indirect, elle se fait dans beaucoup de méthodes par paliers successifs. Dans les derniers paliers on donne en général des informations pour orienter les stagiaires vers des évaluations plus précises – et l'ensemble des jugements individuels initiaux ne coïncideront pas avec les résultats finaux d'un séminaire de définition de points de césure couronné de succès. Les informations qui sont habituellement transmises pour aider les stagiaires à définir des points de césure comprend la difficulté empirique des items ; pour les conséquences que les seuils fonctionnels établis par les jugements peuvent avoir sur le pourcentage de personnes ayant atteint le niveau concerné, etc, ainsi que pour d'autres informations, veuillez vous reporter au chapitre 6.

5.7.4. Documentation

A la fin du stage, il est essentiel que le jeu d'échantillons calibrés soit archivé, accompagné des comptes rendus du stage. Lors d'un stage ultérieur de formation, il sera extrêmement utile de pouvoir donner une explication justifiant qu'un échantillon donné ait été classé à un certain niveau. A cet égard la documentation qui accompagne les échantillons représentatifs des DVD peut servir de modèle.
L'enregistrement sonore des débats lors du stage peut être un document utile pour préparer des notes de ce type sur chaque échantillon calibré. Le coordinateur peut aussi décider de demander à l'un des stagiaires de l'aider en prenant des notes sur la raison du classement de certains échantillons à des niveaux donnés. On peut alors standardiser ces notes et en faire un ensemble cohérent pour la documentation et les distribuer aux participants à l'issue du stage.

Les utilisateurs du manuel peuvent se demander :

- comment s'assurer de la constitution d'un panel équilibré et représentatif pour le projet ;
- quelle taille un panel peut et doit raisonnablement avoir ;
- quelle est la stratégie la plus appropriée au contexte (en termes de ressources, de planification, d'application, d'analyse) ;
- si le projet a pour but de calibrer des échantillons « locaux » pour une utilisation ultérieure comme échantillons représentatifs d'un contexte spécifique ;
- comment s'assurer de la qualité d'un tel matériel « local' en vue du calibrage (et de formations ultérieures) ;
- sous quelle forme présenter la documentation sur le matériel local et comment la distribuer ;
- quelle durée de formation est nécessaire ;
- si tous les participants doivent être mis au même niveau au départ ou s'il est possible de donner à certains des tâches à accomplir avant le stage ;
- s'ils vont utiliser les niveaux plus (il y a des arguments pour et contre ; l'important est de ne pas modifier l'approche une fois que le processus est en cours) ;
- s'ils vont utiliser les grilles d'évaluation du CECRL dans l'annexe C ou élaborer d'autres outils plus spécifiques du CECRL ;
- comment publier et diffuser les résultats du processus de standardisation dans le champ de l'évaluation ;
- comment s'assurer d'une bonne diffusion locale et du suivi.
-

Tableau 5.5 : Formation à la standardisation et calibrage : récapitulatif				
Activité	***Matériel nécessaire***	***Durée***	***Effectif***	***Suggestions***
FAMILIARISATION	• Questionnaires de contrôle fondé sur des rappels du cadre de référence • Photocopies de ces listes • Photocopies des Tableaux 1 et 2 du CECRL • Versions abrégées du Tableau 2 du CECRL, autres échelles	2 heures	Coordinateur Possibilité de grands groupes	Utiliser le programme d'auto formation en ligne s'il est disponible
FORMATION (Capacités de production)	• Vidéos de performances standards (au minimum 8) • Ecrits standards (au minimum 8) • Photocopies d'échelles spécialisées de compétence : • Tableau 3 du CECR/ Tableaux B1 – B3 (performance orale) • Tableau B4 (performance écrite) Photocopies de • Fiches de notation des stagiaires (Fiches B2 et 3) • Fiches de notation du coordinateur (Fiche B4) Photocopies d'autres échelles supplémentaires si pertinentes	3 à 4 heures par capacité Introduction, 30 min Echantillons standards, 90 min Echantillons locaux, 90 min	Coordinateur 30 stagiaires maximum	Traiter deux compétences par jour ou passer une demi-journée sur la formation et une demi-journée sur le calibrage d'une seule capacité.
FORMATION **(Capacités de réception)**	Photocopies d'échelles spécialisées de compétence : • Compréhension générale de l'écrit • Compréhension générale de l'oral Photocopies de • Fiches de notation des stagiaires (Annexe 2) • Fiches de notation du coordinateur (Annexe 3) Photocopies d'autres échelles supplémentaires si pertinentes • Modèles d'items calibrés	3 à 4 heures par capacité: Introduction, 30 min Echantillons standards, 90 min Echantillons locaux, 90 min	Coordinateur 30 stagiaires maximum	Il est possible de traiter deux compétences par jour car les stagiaires seront maintenant familiarisés avec les niveaux du CECRL et les activités de standardisation
CALIBRAGE D'ECHANTILLONS DE PERFORMANCES **(Production)**	• Vidéos locales (au minimum 8) • Ecrits produits localement (au minimum 8) • Photocopies d'échelles spécialisées de compétence • Tableau 3 du CECR/ Tableaux B1-B3 (performance orale) • Tableau B4 (performance écrite) Photocopies de • Fiches de notation des stagiaires (Fiches B2 et 3) • Fiches de notation du coordinateur (Fiche B4) Photocopies d'autres échelles supplémentaires si pertinentes	3 à 4 heures par capacité: Introduction, 30 min Echantillons standards, 90 min Echantillons locaux, 90 min	Coordinateur 30 stagiaires maximum	Traiter deux compétences par jour ou passer une demi-journée sur la formation et une demi-journée sur le calibrage d'une seule compétence

Chapitre 6 : Procédures de détermination des scores de césure

6.1. Introduction

6.2. Aspects généraux

- **6.2.1. Organisation**
- **6.2.2. Concepts**

6.3. La méthode de Tucker-Angoff

- **6.3.1. Procédure**
- **6.3.2 Le candidat aux compétences minimales**
- **6.3.3. Les déclarations de probabilité**
- **6.3.4. Regroupement des normes individuelles et approximation**

6.4. Deux variations de la méthode de Tucker-Angoff

- **6.4.1. La méthode du « oui-non »**
- **6.4.2. Extension de la méthode de Tucker-Angoff**

6.5. La méthode des groupes contrastés et la méthode des cas limites

- **6.5.1. La méthode des groupes contrastés**
- **6.5.2. La méthode des cas limites**

6.6. La méthode du corpus de productions

- **6.6.1. Formation, précision de l'étendue et localisation par agrandissement**
- **6.6.2. Calcul des scores de césure : régression logistique**

6.7. La méthode d'appariement au descripteur de l'item et la méthode du panier

- **6.7.1. La méthode d'appariement au descripteur de l'item**
- **6.7.2. La méthode du panier**

6.8. La méthode du marque-page

- **6.8.1. Le travail du panel d'experts**
- **6.8.2. Contenu des livrets d'items ordonnés**
- **6.8.3. Aspects techniques**

6.9. Variante de la méthode du marque-page selon le Cito

6.10. Déclinaisons particulières

- **6.10.1. Définition des scores de césure sur plusieurs compétences**
- **6.10.2. Définition des scores de césure et ajustement de tests**
- **6.10.3. Définition des scores de césure sur plusieurs langues**

6.11. Conclusion

6.1. Introduction

Lé résultat élémentaire de la participation à un test est un score numérique. Dans le cadre de tests constitués d'une forte proportion d'items, en réception écrite et en réception orale par exemple, ce score correspond généralement au nombre de bonnes réponses. Dans le cadre des capacités productives, la performance est principalement évaluée à partir d'un nombre définis de critères pour lesquels le candidat reçoit un nombre de points (par exemple de zéro à quatre ou cinq). Le cas échéant, le score au test est le nombre total de point acquis par le candidat sur l'ensemble des critères et l'ensemble des tâches qu'il ou elle a accompli. Sur la base de ce score une décision est prise quant aux compétences du candidat, dont la plus importante, celle relative à l'échec/réussite : est-ce que la performance du candidat au test est satisfaisante ? Si la certification est reliée au CECRL une autre décision doit alors être prise : savoir si le candidat a atteint ou non un niveau particulier du CECRL (B2 par exemple). Ces décisions (échec/réussite et niveau du CECRL) impliquent la détermination d'un *score de césure* qui définit une *performance normée*. Pour la décision échec/réussite, le score de césure est le score minimal au test qui conduit à la décision "réussite"; les scores inférieurs à ce score de césure conduisent eux à la décision « échec ». De même, un score de césure pour le niveau B2 correspond au score minimal qui conduira à positionner la compétence du candidat au niveau B2 ou plus; les scores inférieurs sont alors interprétés comme infra-B2 (c'est-à-dire B1 ou moins que B1).

Certains tests nécessitent plusieurs points de césure. En reliant l'examen au CECRL, on pourrait par exemple souhaiter disposer d'un score de césure pour A2, B1 et B2. Ceci est particulièrement important. Un score de césure doit être conçu comme une frontière entre deux catégories adjacentes d'une seule et même échelle. Ainsi, dans l'exemple dont il est ici question, il faudra considérer que chaque candidat sera classé soit en A2, en B1 ou en B2 et que deux scores de césure sont alors nécessaires : l'un qui marque la frontière entre les niveaux A2 et B1 et l'autre pour la frontière entre les niveaux B1 et B2. En général, le nombre de points de césure est égale au nombre de classification moins un.

Pour éviter toute confusion entre les catégories (niveaux) et les scores de césure (les limites entre ces niveaux), on dénomme souvent le point de césure par les deux catégories adjacentes qu'il sépare. Dans l'exemple précédent avec trois catégories, les points de césure pourront être indiqués comme A2/B1 et B1/B2. Il est primordial de rester vigilant à l'égard de la labellisation des deux catégories aux extrémités de l'échelle : la labellisation de la catégorie la plus faible, dans cet exemple en A2, pourrait impliquer que tout candidat dont le score est inférieur au score de césure A2/B1 est de niveau A2, incluant également les candidats ayant un score de zéro. C'est pourquoi il est préférable de rendre la labellisation explicite, pour l'exemple ci-dessus, il conviendrait de retenir « A2 ou inférieur à A2 ». De même, l'utilisation de « B2 ou supérieur à B2 » serait plus judicieuse pour la catégorie supérieure de cet exemple.

La détermination du score de césure ou de la performance normée relève souvent d'une décision collégiale. Le groupe qui réalise une pareille décision est généralement appelé panel ou groupe de décision. La participation d'un panel dure classiquement plusieurs jours. La plus grande partie du temps est consacrée à des activités qui sont décrites dans les chapitres précédents. Pour relier les examens au CECRL, les panélistes doivent être familiers du CECRL (Chapitre 3), ils doivent s'assurer que l'examen recouvre lui même les spécifications du CECRL (chapitre 4), et enfin, ils doivent être entrainés à la façon d'appliquer les descripteurs du CECRL à l'examen (Chapitre 5). Dans ce présent chapitre, l'attention sera portée sur des aspects plus formels du groupe de décision : le type de jugement établi par les panelistes, le type d'information disponible et la manière dont les jugements sont traités et compilés pour parvenir à un ou plusieurs scores de césure. De telles procédures ont souvent été formalisées et sont connues sous le nom de procédures de détermination des scores de césure.

La définition des scores de césure peut avoir des conséquences importantes pour les individus et pour les décisionnaires politiques. Cette détermination exige un jugement

prudent ; autrement dit «la définition des scores de césure est probablement le pan de la psychométrie qui associe plus des aspects culturels, politiques et artistiques en un mélange de ses produits que n'importe quel autre» Cizek (2001, p. 5).

6.2. Aspects généraux

Une part essentielle des procédures de détermination des points de césure tient en l'organisation efficace des rencontres. Généralement, une partie voire la totalité des phases de familiarisation, de spécification et de standardisation décrites dans les chapitres précédents de ce manuel forment un ensemble cohérent avec les procédures de définition des scores de césure (au sens strict du terme) qui sont discutées dans ce chapitre. Ainsi, la procédure considérée dans son ensemble nécessite des ressources et exige une organisation efficace. Une excellente introduction est proposée dans les premiers chapitres de Cizek & Bunch (2007). Dans cette section, l'attention est donc restreinte à la détermination des scores de césure, et les autres éléments fondamentaux seront seulement brièvement exposés.

6.2.1. Organisation

Les procédures de définition des points de césure par un panel durent généralement deux à trois jours, et démarrent par une ou deux sessions de familiarisation, de discussion sur les spécifications du test, d'entraînement avec du matériel servant d'illustration. Elles passent ensuite par une étape cruciale au cours de laquelle tous les experts du panel jugent le test constitué des items considérés. Après la remise d'instructions appropriées, les membres du jury rendent leur jugement, généralement au cours de deux ou trois tours séparés par des phases de discussions, puis de mises en commun et de données supplémentaires.

Pendant les sessions entre les phases d'évaluation, deux types d'informations principales sont fournies. Après la première phase d'évaluation, une information indiquant le comportement des membres du jury est remise, montrant que certains d'entre eux rendent de véritables jugements déviants. Ce type d'information est appelé information *normative*, et doit en principe permettre, en premier lieu, de détecter et d'éliminer les malentendus au sujet des instructions. C'est une bonne expérience que de permettre aux membres du jury de discuter de cette information en petit groupe. Le risque, avec ces échanges, est alors d'orienter le groupe vers le point de vue de la personnalité la plus dominante de ce groupe (voir les suggestions dans la section 5.4.1). C'est la tâche et le rôle du leader du groupe (le facilitateur) que de conduire les discussions de telle sorte que les membres du jury ne se sentent pas influencés par cette personne.

Après le deuxième tour, une information de nature différente nommée *impact* est souvent donnée. Cette information indique les conséquences des jugements des panélistes et repose sur le calcul de la proportion des candidats qui atteindraient ou non chaque catégorie selon les scores de césure provisoires déterminés par le résultat des tours précédents. Bien entendu, pour être en mesure de pouvoir réaliser cette opération, on devra avoir collecté les scores d'un échantillon représentatif de candidats.

Le paragraphe précédent pourrait prêter à confusion. La détermination des points de césure telle qu'elle est décrite dans ce chapitre aborde les performances considérées dans une approche critériée : il est demandé à des juges expérimentés de formuler les minima requis (en termes de performances au test) pour réussir l'examen ou pour obtenir le niveau « B2 », qui sont supposés être guidés par l'application d'un système général (dans notre cas le CECRL) à un test ou à un examen. On serait en droit de penser que le pourcentage de participants qui réussissent l'examen n'est pas important. Mais on ne devrait pas oublier que la procédure conduisant à la définition des scores de césure en situation de fort enjeu est souvent ancrée dans un contexte social et politique, et qu'il est alors prudent de confronter les panélistes aux conséquences sociales de leurs décisions. Après avoir informé les panélistes, il est possible qu'un certain nombre d'entre eux changent d'avis et deviennent plus stricts ou plus indulgents, par rapport à leurs jugements précédents, et ce pour des

raisons opportunistes. Si cela se produit, cela n'implique pas nécessairement que ce changement d'opinion devienne la décision finale. Au contraire, une déviation importante dans les standards après mesure de l'impact pourrait être utilisée pour engager une discussion plus approfondie dans le but de trouver un consensus raisonnable et rationnel entre deux décisions très différentes ; ce qui pourrait justifier l'organisation d'un quatrième tour de jugement.

On doit conserver à l'esprit que la présentation des informations normatives et celles relatives à la mesure de l'impact nécessite un travail préparatoire conséquent. Cette préparation doit être telle que les calculs afférents (qui dépendent des jugements effectués par les panélistes) peuvent être entrepris efficacement (par exemple pendant la pause du déjeuner) pour que l'information soit disponible pour le tour suivant.

Pour la grande majorité des procédures d'établissement des points de césure décrites dans la littérature, de nombreuses variations ont été testées, adaptées à des besoins spécifiques ou inspirées par des carences d'expériences antérieures. Les applications illustrent ce qui tient essentiellement en la même procédure : l'organisation des échanges (en séance plénière ou en petits groupes), etc. Ces variations peuvent toutefois différer par le nombre de tours de jugements. Il n'est pas nécessaire de suivre à la lettre les détails des procédures décrites, des variantes répondant au mieux à un dispositif particulier peuvent être introduites. Dans la suite de ce chapitre, les détails des procédures et les variantes possibles ne seront pas abordés; les traits essentiels et les caractéristiques de chaque méthode doivent être considérés comme l'élément à retenir.

Pour jauger la validité et l'efficacité d'une procédure appliquée à un projet donné, il est crucial qu'une documentation détaillée et adéquate de l'ensemble des étapes de la procédure soit disponible. Sans cette description technique détaillée, l'évaluation professionnelle des résultats devient délicate et l'on ne peut plus prétendre avoir élaboré un argumentaire.

6.2.2. Concepts

En insistant sur le fait que les scores de césure ne peuvent être correctement définis en se contentant de suivre mécaniquement une méthode donnée, ce chapitre proposera une discussion de quelques aspects fondamentaux qui sont soulevés par une variété de méthode de détermination des scores de césure. Parmi ces concepts, on trouvera :

- les déclarations de probabilité ;
- la probabilité de maîtrise ou la probabilité de réponse ;
- la notation à crédit partiel ;
- les concepts liés à la TRI (paramètres de difficulté, niveau de difficulté, discrimination) ;
- les tables de décisions ;
- les livrets d'items ordonnés (OIB en anglais), et
- la zone seuil.

Il est délicat d'introduire ces concepts dans le résumé. Ainsi, ils seront présentés dans le chapitre lorsqu'ils deviendront, pour la première fois, nécessaires à la description d'une méthode particulière. L'ordre dans lequel ils sont abordés est celui qui aidera l'utilisateur à suivre le développement de ces concepts. Il n'y a pas de relation entre l'ordre de présentation des méthodes et leurs caractéristiques qualitatives. Le chapitre présente une variété de méthode pour définir les scores de césure afin de proposer un choix, mais comme les situations de détermination des points de césure diffèrent, il ne préconise pas l'utilisation d'une méthode particulière plus qu'une autre.
Les méthodes pour établir les points de césure sont parfois divisées en deux sous-ensembles ; d'une part celles centrées sur le test, et d'autre part celles centrées sur le candidat. Trois méthodes de cette dernière catégorie sont discutées. La méthode des groupes contrastés et la méthode des cas limites qui utilisent directement le jugement des candidats par un correcteur qui les connait bien. La méthode du corpus de productions, qui

requiert des jugements holistiques sur l'ensemble du travail d'un échantillon de candidats, est utilisée pour déterminer leur score au test ou à l'examen, et ce, pour des réponses à des questions à choix multiples, des réponses construites, ou encore pour des productions plus conséquentes. La caractéristique importante de ces méthodes centrées sur le candidat tient au fait que les candidats *spécifiques* sont reportés dans des catégories (échec ou réussite, niveau B1, B2 ou en cas limite) par un jugement holistique.

Parmi les plus anciennes méthodes, comme celles de Tucker-Angoff ou de Nedelsky[14], il est demandé aux panélistes d'effectuer un jugement sur chaque item. Ces jugements reposent sur les caractéristiques des items perçues par le panel d'experts. La procédure, dans son ensemble, peut être appliquée sans aucune donnée empirique de candidats. Pour ces méthodes, la mention « centrée sur le test » est tout à fait appropriée. Avec la popularité grandissante de la théorie de réponse à l'item (TRI), des méthodes ont été développées. Pour celles-ci, la distinction entre les méthodes centrées sur le test et celles centrées sur le candidat est moins claire. Dans ces méthodes, l'information disponible pour les panélistes est directement issue des performances d'un groupe de candidats. Généralement, cette information est formalisée par la mesure de difficulté de l'item. La disponibilité d'une telle information est censée aider le panel d'experts et les dispenser de la délicate tâche de fournir une estimation de la difficulté qui repose exclusivement sur les caractéristiques perçues d'un item.

Les méthodes discutées dans ce chapitre pourraient ainsi être classées en trois groupes. Le premier serait relatif aux méthodes « centrées sur le candidat » (C-C), le deuxième serait relatif aux méthodes « centrées sur le test » (C-T) dans la mesure où elles peuvent être mises en œuvre sans aucune donnée empirique, et le troisième serait relatif aux méthodes de la « TRI » en ce sens où le panel d'experts utilise un résumé des données empiriques (classiquement fourni par l'analyse dans le cadre de la TRI).

Le tableau 6.1 ci-après offre un aperçu des méthodes discutées, leur classification est donnée dans la colonne « classe » et la section où elles sont traitées est indiquée dans la colonne « section ». Dans la section 6.10. des sujets particuliers sont discutés.

La qualité de la définition des scores de césure est sujette à de grande variation. Quelle que soit la méthode retenue ou la combinaison de plusieurs d'entre elles, nous ne pouvons pas considérer que les scores de césure ont été correctement définis uniquement parce que certaines procédures auraient été respectées. Il est nécessaire de rassembler des *preuves évidentes de qualité des résultats* des procédures et d'en faire part de façon suffisamment détaillée et transparente. Cette question concernant la validité sera traitée plus longuement dans le dernier chapitre de ce manuel.

[14] Cette méthode est probablement la plus ancienne des méthodes de détermination des scores de césure. Elle n'est pas discutée dans ce manuel. L'ouvrage de Cizek and Bunch (2007, Chapter 4) en offre une bonne description.

Tableau 6.1: vue d'ensemble des méthodes discutées

Méthodes	Section	Classe
La méthode de Tucker-Angoff	6.3.	C-T
La méthode du Oui / Non	6.4.1.	C-T
Extension de la méthode de Tucker-Angoff	6.4.2.	C-T
La méthode des groupes contrastés	6.5.1.	C-E
La méthode des cas limites	6.5.2.	C-E
La méthode du corpus de productions	6.6.	C-E
La méthode d'appariement du descripteur de l'item	6.7.1.	C-T
La méthode du panier	6.7.2.	C-T
La méthode du marque-page	6.8.	TRI
Une variante de la méthode du marque page par le Cito	6.9.	TRI

6.3. La méthode de Tucker-Angoff[15]

Bien que cette méthode a été introduite en 1971 comme une remarque dans un chapitre consacré aux tests, à l'étalonnage, à la standardisation et à l'ajustement, qu'Angoff a écrit pour une seconde édition du livre de référence *Educational Measurement* (Thorndike, 1971), c'est encore, après plus de 35 années, l'une des méthodes les plus répandues pour déterminer les scores de césure. De nombreuses variations de cette méthode ont été proposées ; dans ce chapitre deux d'entre elles seront abordées. Nous commençons avec celle aujourd'hui appelée « la méthode d'Angoff », même si Angoff la présentait seulement dans une note de bas de page comme étant une variation de la procédure exposée dans le corps du texte.

6.3.1. Procédure

Un concept de base, qui apparait également dans de nombreuses autres procédures d'établissement des points de césure, est le concept du « candidat aux compétences minimales », également désigné parfois comme le « candidat limite », le « candidat à la frontière » ou encore le « candidat réussissant à peine ». Là où un point de césure doit être utilisé, par exemple pour le CECRL au niveau B1, le candidat aux capacités minimales est celui qui a les compétences pour être apparié au niveau B1, mais de telle sorte que la perte, si infime soit-elle, d'une partie de ses compétences suffirait à ne plus le catégoriser dans ce niveau de qualification. La tâche des panélistes est de conserver à l'esprit un tel profil de candidat ou d'un ensemble de candidats durant tout le travail de jugement qu'ils doivent effectuer.

Pour chaque item du test, le panel d'experts doit indiquer avec quelle probabilité un candidat aux compétences minimales répondrait correctement. De la sorte, les données collectées au cours d'un tour de jugement peuvent être représentées comme celles qui figurent dans le tableau 6.2. ci-dessous, où 15 juges formaient un panel pour déterminer les scores de césure pour un test de 50 items.

L'étape suivante de la procédure consiste en **l'addition** des probabilités sur l'ensemble des items et pour tous les juges. Pour le juge 1 par exemple, cette somme équivaut à 17.48. La probabilité d'une réponse correcte à un item binaire étant équivalente à son score attendu (voir la section C dans le Supplément au manuel), la somme des probabilités sur l'ensemble des items équivaut au score attendu au test pour le candidat aux compétences minimales, selon le juge 1. Dans l'exemple, nous voyons que ces sommes diffèrent d'un juge à l'autre, il en est toujours ainsi dans les séances qui conduisent à la détermination des scores de césure. Par conséquent, reste à résoudre raisonnablement le problème de la concaténation des sommes individuelles des juges en une décision finale. Le plus souvent c'est le calcul de

[15] Dans la littérature, cette méthode est communément appelée la méthode d'Angoff, mais Angoff lui-même a attribuée celle-ci à son collègue d'ETS, Ledyard Tucker.

la moyenne des sommes qui est opéré, et la moyenne est considérée comme le point de césure recherché.

Pour résumer : trois composantes sont essentielles dans cette procédure. Le concept du candidat aux compétences minimales acceptables, la détermination d'une probabilité pour une réponse correcte par un tel candidat (qui doit être renseignée pour chaque item et par chaque membre expert du panel) et la concaténation des sommes des probabilités pour l'ensemble des panélistes. Chacun de ces aspects sera commenté au cours des sections suivantes.

Tableau 6.2: données de base dans la méthode de Tucker-Angoff

	Juge 1	**Juge 2**	**…**	**Juge 15**
Item 1	0.25	0.32	…	0.35
Item 2	0.48	0.55	…	0.45
Item 3	0.33	0.38	…	0.28
…	…	…	…	…
Item 49	0.21	0.30	…	0.35
Item 50	0.72	0.80	…	0.90
Somme	**17.48**	**19.52**	…	**18.98**

6.3.2. Le candidat aux compétences minimales

Le concept du candidat aux compétences minimales acceptables ou du candidat frontière est au cœur de cette approche. Dans la phase d'entraînement des panélistes, une attention particulière doit être réservée à l'explicitation de ce concept pour en fournir une définition raisonnable, et garantir que la représentation interne d'un tel profil de candidat par les membres du panel d'experts est i) communément intégrée par les panélistes et ii) en accord avec l'objectif et les interprétations des résultats du test.

Supposons qu'une procédure soit élaborée pour définir les points de césure relatif au niveau B1, c'est-à-dire pour rechercher un score de césure entre les niveaux A2 et B1. Pour être certain que ce score de césure reflète les limites et rien d'autre que cela, on doit s'assurer que les membres experts du panel ont une maîtrise précise de la signification de A2 et B1, ou plus généralement, que ces membres sont très familiers du CECRL. Plus encore, les panélistes devraient avoir une idée claire et consistante de la déclinaison du CECRL à chaque item. En ce sens, ils doivent connaître les descripteurs pertinents (« Etre capable de ») en répondant à chaque item. En particulier, ils doivent avoir une idée précise des descripteurs critiques, en l'occurrence ceux qui permettent d'établir la meilleure distinction entre les niveaux A2 et B1. Le processus pour parvenir à une bonne compréhension des différences critiques entre les niveaux A2 et B1, pour ce qui concerne chaque item de l'examen, est chronophage et fastidieux. Des conseils pour l'organisation de cette activité peuvent être consultés dans les chapitres précédents.

Dans quelques unes des variations de la méthode de Tucker-Angoff, il est suggéré que les membres experts du panel aient à l'esprit un candidat **concret**, qu'ils considèrent comme un candidat à la frontière des niveaux visés, par exemple un candidat qu'ils connaissent bien. L'argument mis en avant pour cette procédure est qu'elle devrait aider les membres experts du panel à se construire une représentation **stable** du candidat qui se situe à la frontière entre deux niveaux, et ce au fur et à mesure qu'ils parcourent la liste des items. Bien que ce soit admis, travailler avec des candidats réels présente deux inconvénients. Le premier est qu'un tel candidat est généralement connu par un seul des panélistes et qu'il devient difficile

d'utiliser les caractéristiques de celui-ci dans les échanges en petits groupes, parce que précisément à l'exception d'un seul panéliste personne d'autre ne le connait. Le second inconvénient est plus problématique. Le recours à un candidat réel devient critique si chacun pense à un candidat en particulier. En effet, il sera très difficile de réduire les écarts avec la représentation correcte de la personne se situant à la frontière des niveaux et sur laquelle les panélistes devraient s'appuyer pour effectuer leur jugement. Ce problème peut se produire au démarrage de la procédure de détermination des scores de césure mais également pendant les phases d'entraînement et d'échanges en petits groupes. Dans tous les cas de figure, il est recommandé que le travail réalisé à partir du « candidat concret » ne se fasse pas au détriment de la phase minutieuse d'entraînement.

6.3.3. Les déclarations de probabilité

Pour chaque item, les membres experts du panel doivent statuer sur la probabilité avec laquelle le candidat à la frontière des niveaux visés donnerait une réponse correcte. Parce que les panélistes sont souvent peu familiarisés avec les probabilités, ils peuvent être sceptiques vis-à-vis de ce type de tâche. Il est alors vivement conseillé de rendre cette opération plus concrète. Par exemple, on peut leur suggérer d'imaginer 100 personnes de même niveau que le candidat qu'ils sont en train de considérer et qui répondent à l'item. La question serait alors : combien d'entre-eux vont réussir l'item ? Le nombre indiqué par les panélistes est ensuite divisé par 100 pour être considéré comme la probabilité de réponse correcte du candidat. Cette probabilité est communément nommée le coefficient d'Angoff.

L'utilisation du nombre 100 dans l'exemple ci-dessus présente deux avantages. Premièrement, la réponse proposée par les membres du panel peut être directement interprétée comme un pourcentage, et deuxièmement le nombre de réponse possible (de 0 à 100) est suffisamment grand pour garantir avec précision l'expression des probabilités. Supposons qu'un des membres du panel ait à l'esprit une probabilité de 2/3, soit de 0.6666$\underline{6}$. Pour répondre à la question avec 100 personnes, il dira probablement 67 [16].

Il y a deux aspects à prendre en compte quand les panélistes doivent attribuer une probabilité. Le premier, avec des questions à choix multiples, est que la probabilité d'une réponse correcte peut être importante, même si le niveau de compétence du candidat est nettement inférieur à celui de la personne à la frontière des deux niveaux. La raison est liée au choix heureux par ignorance. Il est utile de le rappeler aux experts du panel, par exemple en les invitant à ne pas statuer sur une probabilité inférieure à la réponse au hasard (obtenue en divisant un par le nombre de réponses possibles). Il s'agit d'un point important pour les échanges entre les tours de jugement et pendant la phase d'entraînement.

L'autre aspect est lié à la tendance à éviter les jugements extrêmes ; ce qui signifie que lorsque qu'on dispose de suffisamment d'information pour statuer sur une probabilité extrême, il existe une tendance, dans le comportement humain, qui consiste à éviter ce travers en donnant des valeurs plus grandes que les valeurs réelles quand celles-ci sont très faibles, ou en donnant des valeurs plus faibles que les valeurs réelles quand celles-ci sont très fortes. Si une telle tendance se produit quand on utilise cette procédure, l'effet différera en fonction du niveau général de difficulté du test ou de l'examen. Pour un test très facile et pour le candidat se situant à la frontière des niveaux, les probabilités seront très fortes pour de nombreux items. Si ces probabilités sont systématiquement biaisées vers des valeurs inférieures, par le réflexe humain décrit ci-avant, alors le point de césure sera plus faible (plus d'indulgence) que sans cette tendance. Inversement, pour un test très difficile pour ce même candidat : les faibles probabilités seront surestimées, et le point de césure recherché sera biaisé à la hausse.

[16] Ce n'est pas la même chose que 100x2/3, mais l'erreur est suffisamment petite pour ne pas causer de biais systématique dans le résultat final. Si on utilisait 10 au lieu de 100 (ou d'arrondir la probabilité à une décimale, c'est-à-dire que les réponses seraient 0, 0.1, 0.2….1) on constaterait une erreur systématique sur le résultat final, en particulier si le standard recherché est proche de l'une ou l'autre des bornes du score. (Reckase 2006a; 2006b.)

Bien entendu, il est très difficile de mesurer à quel point cette tendance conservatrice se réalise dans un projet donné de définition des points de césure, mais l'on peut tenter d'éliminer ces phénomènes de deux manières. La première s'applique à tous les jugements des méthodes de définition des scores de césure : être modeste au regard des ambitions. Il est illusoire de penser qu'il est possible d'élaborer un test et d'obtenir des points de césure pour les six niveaux du CECRL (de A1 à C2) dans un seul et même test ou examen en utilisant les méthodes centrées sur le test. Avec la méthode de Tucker-Angoff, cela implique que pour les candidats à la frontière des niveaux A1 et A2 il y aurait de nombreux items très difficiles, et inversement, pour les candidats à la frontière des niveaux C1 et C2 il y aurait de nombreux items très faciles (nécessaires pour le point de césure A1/A2). Même une faible tendance à attribuer des probabilités de façon conservatrice pourrait avoir un effet substantiel sur les scores de césure, en étant trop sévère pour les faibles niveaux et trop indulgents pour les niveaux plus élevés.

La seconde façon pour éviter des biais systématiques dans l'estimation des probabilités est de fournir aux panélistes ce que Cizek et Bunch appellent le *feedback de réalité*. Cela peut être réalisé de la manière suivante et sous la condition que les données réelles du test soient disponibles. Après le premier tour de détermination des scores de césure, les points de césure provisoires sont calculés. Supposons que dans un test constitué de 50 items comme celui utilisé dans l'exemple du tableau 6.2., la moyenne des probabilités est 18.52, ainsi le point de césure correspondra à un score de 18 ou 19. Si ce point de césure n'est pas trop éloigné du point de césure définitif, il est raisonnable de considérer les candidats avec un score avoisinant celui du point de césure provisoire comme des candidats à la frontière des niveaux délimités par ce point de césure. Pour ces candidats on peut calculer la proportion des réponses correctes à chaque item et donner les résultats de ces calculs comme élément de feedback aux panélistes quand ils seront en préparation du deuxième tour de jugement. Ces proportions sont des estimations empiriques de la proportion des réponses correctes pour les candidats qui se situent au point de césure. Les experts du panel pourraient la comparer à leurs propres estimations et être conduits à fournir des ajustements raisonnables. A partir des déterminations des probabilités au tour suivant, il peut être constaté si, et dans quelle mesure, les déterminations conservatrices ont été ajustées dans la direction souhaitée.

Pour définir un voisinage raisonnable au point de césure provisoire, on souhaite souvent avoir un compromis entre la largeur de l'étendue autorisée et le nombre de candidats ayant un score dans cet intervalle. Supposons qu'on fixe le point de césure provisoirement à un score de 19 points, et supposons que seulement 15 candidats aient obtenu ce score. La proportion des réponses correctes pour chaque item dans ce petit groupe aura un fort écart-type parce qu'ils sont peu nombreux. Elargir la définition du voisinage de 17 à 21, par exemple, augmenterait considérablement ce nombre, mais d'un autre côté, si le point de césure est réellement à 19, il pourrait être litigieux de considérer légitimement les candidats avec un score de 17 ou 21 comme étant à la frontière des niveaux. Une stratégie possible est de définir le voisinage comme le point de césure provisoire plus ou moins l'écart-type de mesure. Pour éviter les biais, il est important que l'étendue du voisinage soit symétrique autour du point de césure provisoire.

6.3.4. Regroupement des normes individuelles et approximation

Additionner les probabilités associées aux items d'un membre du panel fournit le point de césure individuel de cet expert. La moyenne des ces estimations individuelles peut être appréhendée comme le point de césure de l'ensemble du panel. Mais il n'en est pas ainsi. D'une certaine manière les moyennes sont des mesures fragiles. En particulier la moyenne est sensible aux valeurs extrêmes, atypiques, qui peuvent provenir d'un ou deux experts du panel, qui s'obstinent à donner des points de césure extrêmes, ou qui n'auraient pas compris la procédure. Pour limiter l'influence de telles extrêmes sur la décision de groupe on peut utiliser des indices plus robustes. Le plus populaire d'entre eux est la médiane, mais un autre, très utile, est la moyenne tronquée. Une moyenne tronquée est la moyenne d'un jeu de donnée où un certain pourcentage de données est exclu du calcul. Les données exclues

sont les plus extrêmes (aussi bien du côté supérieur que du côté inférieur). Si les experts du panel sont au nombre de 20 et que le pourcentage de troncature est fixé à 10%, alors la plus forte et la plus faible des valeurs sont exclues et la moyenne sera ainsi calculée sur les 18 valeurs restantes.

Généralement le point de césure d'un membre individuel du panel, comme celui du groupe d'expert est une moyenne, une moyenne tronquée ou la médiane et correspondra à un nombre décimal. Mais en pratique, la participation individuelle à un test ne peut pas résulter en un score à valeur décimale. Ainsi, le résultat décimal devra être arrondi à l'entier immédiatement supérieur ou inférieur. Arrondir à l'entier le plus proche peut apparaitre comme un problème trivial : dans l'exemple, cela signifie qu'il conviendrait d'arrondir le 18.55 à 19, mais en réalité la question est bien plus complexe qu'il n'y parait.

Pour comprendre, on devrait prendre en compte, dès lors que l'on met en œuvre une procédure visant à déterminer les scores de césure, et quelle que soit l'attention méthodologique qu'on y aura accordé, qu'on aboutira inévitablement à des erreurs de classification parce que les scores eux-mêmes ne sont pas parfaitement fidèles. Ces erreurs de classification peuvent se répartir de deux façons : un candidat avec un score vrai égal au- ou supérieur au- score de césure peut être classé comme n'ayant pas atteint le point de césure (approximation par défaut), a contrario, un candidat avec un score vrai inférieur au score de césure, peut, par l'erreur de mesure, être catégorisé comme ayant atteint le point de césure (approximation par excès). Les erreurs de classification ont des conséquences au niveau individuel et parfois au niveau de la société. De façon plus importante encore, les approximations par défaut pourraient être différentes des approximations par excès. Si l'on considère ces dernières comme plus délicates, alors il est préférable de rendre plus sévère le point de césure et d'arrondir vers l'entier supérieur. Les conséquences des erreurs de catégorisation sont discutées plus en détail dans le chapitre suivant.

Une dernière mise en garde pour ce qui concerne l'arrondissement est évoquée ici. Les nombres arrondis, et les calculs effectués sur les nombres arrondis, peuvent avoir des conséquences imprévues et non souhaitées. Par conséquent, l'arrondissement devrait être effectué le plus tardivement possible. C'est une mauvaise pratique, par exemple, que d'arrondir les valeurs individuelles des points de césure (la ligne du bas dans le tableau 6.2.) de chaque expert du panel à l'entier le plus proche, puis de calculer la moyenne des valeurs arrondies pour enfin arrondir de nouveau le résultat. Un simple exemple permet de mieux comprendre ce phénomène : supposons qu'avec trois juges dont les points de césure sont respectivement 17.01, 17.51 et 17.53. La moyenne est de 17.35, soit de 17 après arrondissement. Si on arrondissait les valeurs d'emblée à 17, 18 et 18, la moyenne serait alors de 17.67 et la valeur arrondie serait, elle, fixée à 18.

6.4. Deux variations de la méthode de Tucker-Angoff

Dans les mises en œuvre de la méthode de Tucker-Angoff, l'attribution des probabilités aux réponses correctes est souvent perçue comme étant compliquée à comprendre et à mettre en œuvre. Une variation de cette méthode est appelée la méthode du Oui/Non[17] qui permet d'éliminer cette difficulté.

La proposition originelle d'Angoff était exclusivement dédiée aux tests constitués d'items **binaires** (dichotomiques). Dans de nombreux tests, en particulier pour ceux incluant des capacités productives, les items ont des scores polytomiques pour lesquels on peut obtenir, par exemple 0, 1, 2 ou 3 points. La méthode de Tucker-Angoff peut (en principe) être appliquée à de telles situations. Dans cette section, les deux variations sont brièvement discutées

[17] En fait, c'est ce qu'Angoff avait proposé à l'origine de sa méthode de détermination des scores de césure. La méthode discutée dans la section précédente a été proposée en note de bas de page.

6.4.1. La méthode du « Oui/Non »

La description la plus claire qu'on peut avoir de cette méthode est le texte originel rédigé par Angoff lui-même.

« Une procédure pour décider du score minimal qui détermine la réussite peut être mise en place de la façon suivante : en conservant à l'esprit l'hypothèse du candidat aux compétences minimales acceptables, on pourrait parcourir les items du test les uns après les autres et décider si une telle personne est en mesure de répondre correctement à chacun des items. Si un score de un est attribué à chaque item répondu correctement par le dit candidat et qu'un score de zéro est attribué pour chaque item échoué, la somme des scores obtenus aux items sera égale au score brut du candidat aux compétences minimales acceptables » (Angoff 1971 pp. 514–515).

Au lieu d'attribuer des probabilités (des nombres variant de zéro à un), les experts du panel attribuent un (en disant oui) ou zéro (en disant non). Bien que de bons résultats aient été rapportés avec cette méthode (voir Cizek and Bunch 2007, pp. 88–92 pour quelques résultats), la méthode peut conduire à de sévères biais.

Pour le voir, on peut considérer les réponses données (0 ou 1) comme des probabilités arrondies. Considérons désormais un test plutôt homogène qui serait relativement facile pour le candidat à la frontière entre les niveaux. Cela signifierait pour tous ces items que le candidat a une probabilité supérieure à 50% de répondre correctement aux items, ainsi un expert du panel devrait rationnellement indiquer Oui pour chaque item. Pourtant, s'il procède de la sorte, son point de césure individuel correspondra au score maximal du test alors que le candidat en question obtiendrait en moyenne un score légèrement supérieur à la moitié du score maximal.

Nous pouvons donc en déduire un principe plus général sur la signification de la définition des points de césure. Dans l'exemple précédent, il est clair qu'un résultat significatif peut être obtenu si le candidat situé à la frontière des niveaux peut répondre correctement à certains items du test (avec une probabilité supérieure à .5) et peut échouer aux autres items (avec une probabilité inférieure à .5). Cela empêcherait que le score de césure soit très extrême (proche de zéro ou proche du score maximal). En d'autres termes, cela signifie que le test devrait inclure suffisamment d'information sur l'habileté du candidat situé à la frontière des niveaux et que cela conduit aux mêmes conclusions que précédemment. Si des points de césure doivent être définis pour des compétences distinctes (par exemple pour A1/A2 et B2/C1) en utilisant le même test, on devra alors collecter suffisamment d'information sur plusieurs étendues d'habileté ; ce qui est en règle générale difficilement réalisable, à moins que le test ne soit très long. Ignorer ce principe peut conduire à des résultats aberrants, comme ceux montrés dans l'exemple suivant. Supposons qu'un test soit construit pour établir une distinction entre les niveaux B2 et C1. Utiliser ce test pour fixer les points de césure entre les niveaux A1 et A2 conduirait probablement à un score de césure de zéro avec la méthode du Oui/Non (un candidat à la frontière des niveaux A1 et A2 ne répondra pas correctement aux items), et conduira à une conclusion aberrante : le candidat sera au niveau A2 s'il obtient un score de zéro à ce test.

6.4.2. Extension de la méthode de Tucker-Angoff

Une généralisation de la méthode aux tests constitués d'un ensemble d'items dichotomiques et polytomiques est facile à comprendre si on appréhende la probabilité d'une réponse correcte à un item dichotomique comme le score attendu pour cet item (voir Section C du Supplément au manuel). Pour ce qui concerne les items polytomiques, il est plus difficile de déterminer les probabilités de réponse, parce qu'il convient d'attribuer la probabilité d'obtenir le score 0, 1, 2, etc. jusqu'au score maximal de l'item. On peut néanmoins circonscrire ce problème en déterminant le score attendu pour un item polytomique. L'instruction qui serait donnée aux experts du panel dans cette situation serait la suivante : *« Imaginez que 100 candidats qui se situent à la frontière des niveaux visés répondent à cet item, pour lequel on*

peut obtenir jusqu'à 4 points, quel serait, selon vous, le score moyen obtenu par ces 100 candidats ? »

Au lieu de renseigner les probabilités dans un tableau comme le tableau 6.2., on peut compléter par le score moyen attendu tel qu'il est déterminé par les panélistes. Les autres opérations de la procédure (sommation et concaténation) demeurent les mêmes que pour la méthode de Tucker-Angoff appliquée aux items dichotomiques.

Le seul problème qui demeure avec cette méthode est qu'on doit s'assurer que les experts du panel aient correctement intégré la signification d'un score moyen. En particulier, ils devraient comprendre que la moyenne peut être un nombre décimal bien que les scores individuels ne peuvent être que des valeurs entières. Une bonne façon est de les former à établir, pour eux-mêmes, une table de fréquences des scores observables pour les 100 candidats à la frontière des niveaux visés et ensuite d'en calculer la moyenne. Un exemple d'une telle table est fourni dans le tableau 6.3., pour un item dont le score maximal est de 3. La tâche des panélistes consiste alors à renseigner la colonne des fréquences dans le tableau (et de vérifier que la somme est égale à 100). La troisième colonne (score multiplié par la fréquence) se complète mécaniquement. A partir de l'exemple dans le tableau 6.3., on déduit immédiatement que le résultat du score attendu est 75/100 = 0.75. Si l'on pense qu'il est risqué de faire remplir la troisième colonne par les panélistes eux-mêmes, on peut se contenter de préparer un tableau plus simple (sans cette troisième colonne) et confier aux experts du panel le soin de compléter la colonne fréquence. Le calcul restant peut alors être effectué dans un second temps par une tierce personne.

Tableau 6.3: calcul du score attendu pour les 100 candidats limites

Score	Fréquence	Score * Fréquence
0	45	0
1	35	35
2	20	40
3	0	0
Somme	100	75

Pour conclure : La méthode de Tucker-Angoff et ses nombreuses variations sont typiques des méthodes *centrées sur le test*, par le fait que la tâche principale des membres experts du panel est de se concentrer sur les caractéristiques des items et de catégoriser ces items en regard de la compétence d'un candidat défini comme étant à la frontière des niveaux visés. Cette catégorisation est absolue (dans la méthode du Oui/Non) ou probabiliste. D'un point de vue purement formel, on pourrait dire qu'au cours de l'application de cette méthode, les experts du panel n'ont pas besoin d'être formés ou d'avoir une expérience particulière avec des candidats réels du test. Cependant, en pratique, sélectionner un tel panel d'experts conduirait à des résultats inacceptables. Même avec des enseignants bien expérimentés, la tâche reste abstraite, et les enseignants considèrent qu'il est difficile de répondre aux exigences de la tâche. C'est pourquoi, toutes les variations autour de cette méthode utilisent aujourd'hui plusieurs tours de jugements et proposent des éléments prégnants sur les performances réelles des candidats pour modérer la détermination des scores de césure. Mesurer l'impact des données offre une idée des conséquences des décisions sur les groupes de candidats et peut conduire à des ajustements non négligeables. Fournir des données réelles, comme la proportion pour un groupe de candidats limites, à partir des points de césure provisoires, peut donner une indication qui aidera à ajuster les estimations de probabilité vers des valeurs plus réelles. Et même avec ces informations, l'accent principal de la méthode porte sur les propriétés du test, la qualification de cette méthode comme centrée sur le test reste donc justifiée. Au cours de la section suivante, deux méthodes centrées sur les candidats seront décrites.

6.5. La méthode des groupes contrastés et la méthode des cas limites

Ces deux méthodes sont très contrastées par rapport à celle de Tucker-Angoff. En effet, les jugements des experts du panel reposent premièrement (et quasi-exclusivement) sur les performances au test de **candidats réels**. Elles sont donc prototypiques des méthodes **centrées sur les candidats.**

La nécessité commune aux deux méthodes est de disposer de scores au test pour un échantillon de candidats. Comme c'est le cas également pour toutes les méthodes visant à déterminer les scores de césure, une attention particulière doit être apporté à l'échantillon sélectionné pour qu'il soit représentatif de la population cible. En outre, les candidats doivent être bien connus par (au moins) un des experts du panel. En pratique, on aura généralement recours pour le panel à des enseignants/formateurs des candidats de l'échantillon retenu. Ainsi, chaque candidat de l'échantillon est bien connu par au moins un des panélistes.

6.5.1. La méthode des groupes contrastés

La tâche des experts du panel est de placer chaque candidat dans l'une des deux catégories (dans le cas d'un seul score de césure) ou dans *k+1* catégories quand il y a *k* scores de césure à déterminer. Si l'objectif de la procédure est de déterminer le score de césure, par exemple, pour le point de césure B1/B2, chaque candidat est catégorisé par les panélistes soit en B1 (ou moins) soit en B2 (ou plus).
Une fois cette information disponible, une table de fréquence munie de deux colonnes est construite. Les lignes représentent le score au test et les deux colonnes indiquent les fréquences des scores pour les groupes de candidats catégorisés en B1 ou en B2. Un exemple qui repose sur des données fictives pour un test constitué de 50 items est proposé dans la figure 6.1., où les deux distributions de fréquences sont représentées graphiquement. L'échantillon total est constitué de 400 candidats, 88 ont été catégorisés en B1 et 312 en B2. Les distributions présentent ici des caractéristiques souvent observées : elles sont très irrégulières (en lien avec la taille modérée de l'échantillon) et elles sont fortement en recouvrement. Il n'est donc pas intuitif de placer le score de césure. Par ailleurs, les deux groupes diffèrent considérablement du point de vue de leur effectif, comme c'est souvent le cas dans les décisions de type réussite/échec.

Le score moyen des candidats B1 est de 16.78 et de 34.24 pour les candidats B2. Un score de césure acceptable, au moins provisoirement, est la valeur qui se situe au milieu de ces deux moyennes, donc 25.51 = (16.78+34.24)/2. Il faut néanmoins rester prudent en prenant en compte cette valeur (ou une valeur arrondie de cette dernière) pour le point de césure.

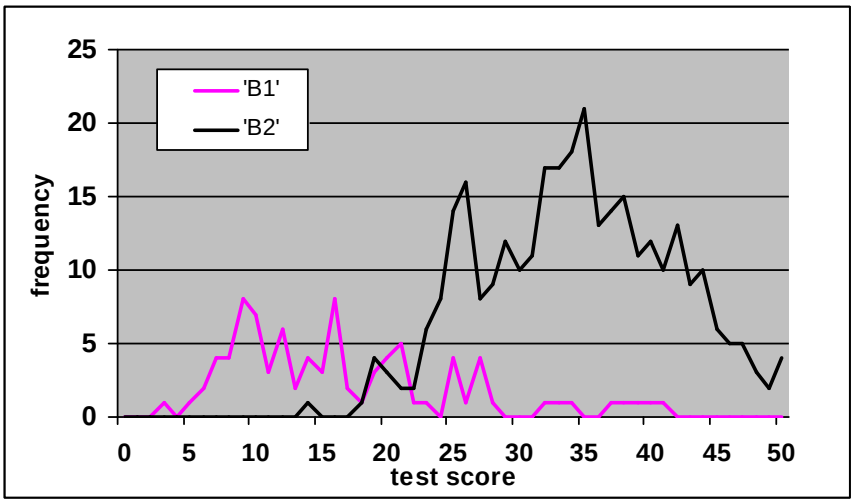


Figure 6.1. Distribution de fréquence pour les scores au test dans les deux groupes contrastés

Pour ce qui concerne la distribution des candidats B1, sept (sur un total de 88) obtiennent un score particulièrement élevé (au-delà de 30 points) et peuvent être considérés comme au dessus du point de césure. Cela vaut donc la peine de vérifier si ces sept candidats ont été catégorisés par le même enseignant ou non. Le cas échéant, cela pourrait devenir un sujet de discussion dans le panel pour voir si cet enseignant/formateur n'a pas été trop sévère dans ses jugements. Même sans ces cas atypiques, le recouvrement entre les distributions est souvent observé.

Une technique efficace pour opérer un choix rationnel est de construire une **table de décision pour plusieurs scores de césure.** Cette technique est illustrée peu après. Dans le tableau 6.4., le tableau de fréquence correspondant à la figure 6.1. est partiellement représenté : les scores faibles (inférieurs à 20) et les scores élevés (à compter de 28) ont été regroupés ; les autres scores sont représentés.

Tableau 6.4: distribution de fréquence correspondant à la figure 6.1.

Score	B1	B2
0–20	63	9
21	5	2
22	1	2
23	1	6
24	0	8
25	4	14
26	1	16
27	4	8
28–50	9	247

Le tableau 6.5. est directement dérivé du tableau 6.4. Prenons le score de césure 24 par exemple. Dans le tableau 6.4. on peut voir que 18 candidats, catégorisés en B1 par leurs formateurs réussissent le test et sont ainsi considérés comme des candidats B2 sur la base de leur score. Ces 18 candidats sont des « faux positifs ». De façon similaire, 19 candidats catégorisés en B2 par leurs formateurs échouent au test, ils sont donc des « faux négatifs ». Rassemblés, ces candidats sont au nombre de 37 et sont tous incorrectement classés. Sur un total de 400 personnes, cela représente 9.3%.

Tableau 6.5: tables de décision pour cinq scores de césure

Score de césure	21		22		23		24		25	
Catégorisé en :	**B1**	**B2**	**B1**	**B2**	**B1**	**B2**	**B1**	**B2**	**B1**	**B2**
Inférieur au score de césure	63	9	68	11	69	13	70	19	70	27
Score de césure ou supérieur	25	303	20	301	19	299	18	293	18	285
Total	88	312	88	312	88	312	88	312	88	312
% d'erreur de catégorisation	8.5		7.8		8.0		9.3		11.3	

Avec le tableau 6.5. on peut s'apercevoir que le pourcentage d'erreur de catégorisation varie avec les variations du score de césure. Il atteint sa valeur minimale pour un score de césure à 22 et varie très peu si le score de césure est fixé à 23. Ainsi, 22 et 23 pourraient être préférés aux valeurs provisoires 25 et 26 déterminées par le milieu des deux moyennes.

Il y a un autre aspect de cette procédure que l'on ne doit pas omettre. Le *nombre* d'erreurs de classification, c'est-à-dire le nombre de faux positifs et de faux négatifs. Pour chaque score de césure dans le tableau 6.5. ce nombre diffère mais dans des proportions toutes relatives. Cette comparaison n'est pas celle qui doit être effectuée parce que le nombre de candidats classés en B1 et B2 par les enseignants/formateurs varie considérablement. Prenons par exemple le score de césure 24, pour lequel le nombre de faux positifs et de faux négatifs est quasi-identique. Les faux positifs représentent 18 sur 88 soit 20.4% des candidats étiquetés B1 alors que les 19 faux négatifs représentent seulement 6.1% des candidats B2. Pour le score de césure 22, ces pourcentages sont respectivement 22.7% et 3.5%, révélant un dilemme qui peut s'avérer en pratique : le score de césure optimal selon le pourcentage total d'erreur de catégorisation n'est pas, en général, optimal en regard de l'équilibre des faux positifs et faux négatifs. Des considérations prudentes sont nécessaires quant aux conséquences des faux positifs et des faux négatifs avant de parvenir à la décision finale.

Il y a deux considérations à prendre en compte quand on applique cette méthode à des situations à forts enjeux. La première est d'ordre statistique, la seconde est de nature plus méthodologique. Pour ce qui concerne la première, la taille de l'échantillon utilisé dans cet exemple est restreinte, en particulier pour le groupe des candidats B1. C'est ce qui rend les nombres qui figurent dans le tableau 6.5. particulièrement instables statistiquement, sous-entendu qu'une réplication à l'aide d'un échantillon de même taille, le tableau pourrait faire l'objet de modifications considérables et conduire à un choix très différent pour définir le score de césure optimal.

L'autre aspect est encore plus important. Le raisonnement conduit pour élaborer les tableaux et les interpréter repose sur l'hypothèse que le jugement des enseignants/formateurs a une haute valeur de véracité et correspond à la réalité (« si votre formateur vous indique que vous êtes B1, alors vous êtes B1 »). Bien sûr, ce n'est pas le cas et les jugements des formateurs, même bien entrainés au CECRL ne seront pas totalement valides. Il est vrai qu'une surestimation de quelques candidats par un formateur pourrait être compensée par une sous-estimation d'un autre candidat par un autre formateur, mais comme personne n'a de contrôle sur ces aspects là ce point demeure problématique. En effet, les candidats restent « à la merci » des formateurs. Si un ou deux formateurs sont trop indulgents, disons s'ils catégorisent facilement en B2 dans cet exemple, il sera quasiment impossible de détecter une telle indulgence. Même s'ils obtiennent plus de jugements B2 que leurs collègues du panel, cela ne constitue pas une preuve de leur indulgence, parce qu'il est possible qu'ils aient eu des candidats plus compétents. On pourrait essayer de vérifier en utilisant les scores au test, en montrant par exemple que le score moyen de leurs candidats est approximativement le même en moyenne que celui des autres candidats, ils auraient alors à ajuster leurs jugements. Cette pratique reste cependant risquée. En effet, le cœur de la méthode des groupes contrastés demeure une comparaison de deux variables : les scores au test et les jugements des enseignants/formateurs. Pour obtenir une méthode efficace, les données des deux variables devraient être collectées de *façon indépendante*, ce qui signifie par exemple que les enseignants/formateurs doivent donner leurs jugements sur

les candidats sans connaître leur score. Si l'on utilise des informations de l'une de ces variables pour ajuster l'autre, on rompt cette indépendance. En procédant de la sorte on manipule les données (vers une certaine décision) et on compromet l'intégralité de la procédure.

6.5.2. La méthode des cas limites

Cette méthode est très similaire à celle des groupes contrastés : elle repose également sur un jugement du niveau de candidats réels. Les jugements eux-mêmes sont utilisés pour identifier les candidats qui doivent être considérés comme étant à la frontière du point de césure recherché. En reprenant l'exemple de la section précédente, on pourrait tenter d'identifier les candidats qui se situent autour de la frontière entre les niveaux B1 et B2.

Lorsque ce groupe est identifié, le score de césure est défini par la valeur centrale des scores au test de ce même groupe, par exemple la moyenne ou la médiane, puis arrondie correctement.

Le principe de cette méthode est très simple, mais l'on peut rencontrer quelques difficultés lors de la mise en œuvre. Quelques-unes d'entre elles vont être examinées.

La première, et assurément la plus délicate, est une définition claire d'un candidat à la frontière des niveaux visés. Dans le CECRL, les niveaux sont opérationnalisés par les descripteurs « être capable de », mais les cas limites ne sont pas explicitement décrits. Les définir comme « quelque chose entre deux « être capable de » risquerait d'être trop confus pour garantir une compréhension commune du CECRL, de laquelle pourrait surgir des variations incontrôlables et indésirables pour les membres du panel. Une bonne façon de procéder pour guider les membres experts du panel dans leur compréhension des cas limites serait de recourir au point de référence : des exemples de performances limites.

La seconde difficulté est de nature statistique. Il est fréquent que la taille du groupe limite soit modérée, pour ne pas dire petite, de telle sorte que la moyenne ou la médiane du score au test de ce groupe aura un écart-type important. En outre, par cette application en tant que méthode autonome, les informations utiles sur les performances des autres candidats ne sont pas utilisées. On peut alors procéder en combinant la méthode des cas limites et celles des groupes contrastés. C'est l'objet de la discussion qui suit.

Considérons une fois de plus l'exemple de la recherche du score de césure pour B1/B2. Au lieu de demander aux panélistes de catégoriser les candidats comme étant limites ou non, on pourrait leur demander de répartir leurs candidats en *trois* catégories : B1, B1/B2 et B2. Les deux groupes B1 et B2 peuvent alors être utilisés pour la méthode des cas limites pour offrir deux points de césure supplémentaires. Cette information est particulièrement utile pour la validation de la procédure. Il en sera un peu plus question dans le chapitre suivant. Pour mettre en place la table de décision (voir tableau 6.5.), les résultats peuvent être aisément combinés, ou mieux encore, les tables peuvent être préparées séparément et fournir l'information sur le taux d'erreur de catégorisation pour les candidats qui sont définitivement identifiés comme étant non limites (selon les jugements des panélistes) et pour les candidats qui sont jugés limites.

Cette méthode fonctionne de façon satisfaisante quand on peut s'assurer que tous les candidats de l'échantillon sont soit de niveau B1 soit de niveau B2 (ou autour de la frontière entre ces deux niveaux). S'il subsiste un doute que des candidats très faibles ou très forts aient participé à l'examen, il est alors plus prudent d'ajouter une ou deux catégories de jugement, qui pourraient être étiquetées, par exemple, A2/B1 ou moins et B2/C1 ou plus, même si l'on n'envisage pas de rechercher le score de césure pour A2/B1 et B2/C1, ces deux catégories supplémentaires peuvent contribuer à éclaircir le contraste entre les groupes B1 et B2.

Un avantage supplémentaire de cette méthode combinée est d'éviter les choix forcés pour les enseignants/formateurs dans le cas où ils seraient dubitatifs eux-mêmes à l'égard de la catégorie définitive dans laquelle ils doivent placer les candidats.

6.6. La méthode du corpus de productions

La méthode du corpus de productions (Kingston et al 2001) est peut être la plus appropriée pour les jugements holistiques, bien qu'elle puisse être utilisée avec toutes les combinaisons de type d'items et de tâches. Elle est centrée sur les candidats et n'utilise pas la TRI. Vous trouverez ci-dessous une brève liste de ce qui est nécessaire pour appliquer cette méthode :

- Une collection de travail d'un échantillon de candidat. Ce travail peut consister seulement en des réponses à des questions à choix multiples, ou en un mélange de questions à choix multiples, de questions ouvertes et de rédactions voire même en un portfolio. Une condition d'application nécessaire est que le travail (la performance au test, le portfolio) soit validé par un *score numérique.*
- L'échantillon n'a pas besoin d'être représentatif d'une population cible du test. Il doit néanmoins couvrir la plupart de l'étendue des scores possibles, indépendamment de la fréquence relative de ces scores avant la mise en place de la procédure pour déterminer les scores de césure.
- La tâche des experts du panel est de fournir un *jugement holistique* sur chaque échantillon de travail qui leur est présenté. Dans le cadre du CECRL, un tel jugement consistera en l'attribution aux candidats de l'un des niveaux prédéfinis que l'on vise dans la procédure de définition des points de césure. Supposons que l'on veuille définir les points de césure pour A1/A2 et A2/B1, le jugement des experts du panel devra catégoriser chaque production de candidat soit en A1, A2 ou B1 (ou plus).
- Le type de jugement requis de la part des panélistes est le même que celui demandé dans la méthode des groupes contrastés ou dans la méthode des cas limites. La différence essentielle avec ces deux méthodes tient au fait qu'ici tous les panélistes évaluent la même collection d'échantillon de production, de telle sorte que les discussions en groupe entre les tours aient du sens. La méthode du corpus de productions nécessite deux tours, bien qu'il puisse être nécessaire d'en ajouter un troisième.
- Les scores des échantillons de travail des candidats ne sont pas connus des experts du panel.
- Pour convertir les jugements des panélistes en un score de césure, on doit avoir recours à une technique particulière, appelée la régression logistique. La raison est liée à la haute sélection de l'échantillon des travaux utilisés. En effet, l'application des méthodes usuelles (par exemple rechercher le point central entre les moyennes dans le cas de la méthode des groupes contrastés) pourrait conduire à de sévères biais.

Dans la suite de cette section, quelques détails sont proposés pour ce qui concerne l'organisation de la méthode (section 6.6.1.) et sur les techniques d'analyse statistique requises (Section 6.6.2.). Des informations supplémentaires pourront être consultées dans les ouvrages de Kingston et al (2001) et de Cizek and Bunch (2007, Chapitre 9).

6.6.1. Entraînement, précision de l'étendue et localisation par agrandissement

Ces trois termes font référence aux différentes phases de cette procédure mais aussi à différents échantillons de travaux qui vont être utilisés. Concrètement, il faudra fixer les points de césure pour A1/A2, A2/B1 et B1/B2, et le panel devra être constitué de 15 membres.

Le matériel d'entraînement consiste en un petit échantillon d'extraits de réponses des candidats, sélectionné pour qu'il couvre le plus largement possible l'étendue des scores et des niveaux. Dans cet exemple, il serait pertinent de sélectionner deux ou trois cas pour chacun des niveaux A1, A2, B1 et B2, et d'essayer de sélectionner des échantillons de travaux de telle sorte qu'ils représentent une variation conséquente du point de vue des

scores obtenus à l'intérieur des niveaux. Pour cette sélection, on peut compter sur des jugements experts. Pour la phase d'entraînement elle-même, le lecteur est renvoyé au chapitre 5. Kingston et al insistent sur le fait que des échantillons de travaux inhabituels ou contradictoires du point de vue des scores doivent être évités, par exemple un travail avec un score très élevé sur des questions à réponses ouvertes et un très faible score sur d'autres items similaires.

Après le premier entraînement, un premier tour de jugement est organisé, appelé précision de l'étendue. Le matériel présenté aux experts du panel est un échantillon de travaux de candidats représentant la totalité de l'étendue des scores obtenus. Les travaux échantillonnés sont présentés dans des dossiers où chacun d'entre eux contient un petit nombre d'extraits de travaux. Le contenu des dossiers doit être uniforme du point de vue des scores et présenté par ordre croissant des scores. Les dossiers sont également présentés dans un ordre croissant selon les scores des échantillons de travaux qu'ils comprennent. Pour un test avec un score maximal de 55, on pourrait préparer 10 chemises avec trois travaux par chemise, de telle sorte que les travaux présentés se réfèrent à 30 scores différents et soient présentés à l'ensemble des experts du panel.

Tableau 6.6: résumé du tour de précision de l'étendue

Dossier	Score	A1	A2	B1	B2	Total
1	13	15	0			15
	15	15	0			15
	16	14	1			15
2	18	13	2			15
	19	11	4			15
	21	9	6			15
3	23	10	5			15
	24	7	8			15
	26	5	10			15
4	27	3	10	2		15
	28	0	12	3		15
	30	1	11	3		15
5	32		9	6		15
	33		11	4		15
	34		8	7		15
6	35		7	8		15
	36		8	7		15
	37		6	8	1	15
7	39		3	12	0	15
	41		1	14	0	15
	42		1	12	2	15
8	43			10	5	15
	45			11	4	15
	46			8	7	15
9	48			4	11	15
	49			1	14	15
	51				15	15
10	52				15	15
	53				15	15
	54				15	15

La tâche de chaque panéliste est d'attribuer à chaque exemplaire de production l'une des catégories du CECRL : dans cet exemple A1, A2, B1 ou B2. Ensuite, les jugements sont collectés et l'équipe organisatrice prépare une table de fréquence des jugements comme

celle qui figure dans le tableau 6.6. On peut déduire de cette table des informations utiles pour réduire la quantité de travail au cours du second tour.

- Pour l'échantillon des travaux dans le dossier 10, les jugements sont unanimes (B2), on peut ainsi assurément considérer que le point de césure B1/B2 est inférieur à 52, le plus faible des scores contenus dans le dossier 10. De façon similaire pour le dossier 1 où il y a unanimité pour la catégorie A1, on peut en déduire que le point de césure A1/A2 est supérieur à 16.
- Les points de césures pour lesquels les jugements des panélistes sont le plus en désaccord se situent principalement au niveau adjacent entre les catégories. Pour le point de césure A1/A2, le score correspond à 24 (dossier 3), pour A2/B1 c'est pour un score de 34 et 35 (dossier 5 et 6), et pour B1/B2 le désaccord le plus grand est observé pour un score de 46 (dossier 8).

Ces scores indiquent la valeur approximée du point de césure, et pour éviter un travail inutile pour les membres du panel au cours du second tour, de nouveaux dossiers sont constitués avec des travaux dont le score est au voisinage des points de césure provisoires. Dans cet exemple (tableau 6.6.), un échantillon des travaux avec des scores compris entre 21 et 27 pour A1/A2, entre 32 et 38 pour A2/B1 et entre 42 et 48 pour B1/B2 serait un choix approprié. Ces nouveaux travaux devraient être répartis en six dossiers, disons de trois ou quatre échantillons de travaux qui seront évalués de la même manière qu'au cours du premier tour. Cette seconde sélection affine les échantillons étudiés de façon plus prononcée qu'au cours du premier tour ; c'est pourquoi le second tour est appelé localisation par agrandissement (pinpointing).

L'échantillon des productions qui doit être évalué au cours du second tour peut être soit un matériel entièrement nouveau, soit identique à celui utilisé au cours du premier tour, soit un mélange d'une association de l'ancien et du nouvel échantillon. Cette décision d'association précise dépendra principalement du temps nécessaire pour parcourir entièrement le nouvel échantillon de travail, mais il est recommandé d'essayer de constituer une association à part égale d'anciennes et de nouvelles productions. Ce nouvel ensemble offre l'opportunité de juger le degré de généralisabilité de la procédure et l'inclusion de l'ancien échantillon de travaux permet d'évaluer la consistance des jugements des panélistes.

6.6.2. Calcul des scores de césure : régression logistique[18]

La technique utilisée pour calculer les points de césure est appelée régression logistique. Comme pour tous les types de régression, il y a une variable dépendante et une ou plusieurs variables indépendantes. Ici, nous disposons d'une seule variable indépendante : le score au test. La variable dépendante est le jugement des membres du panel, qui peut prendre deux valeurs pour un point de césure, disons A2/B1 : la performance est accomplie au regard du point de césure (symbolisé par une valeur de un) ou non accomplie (valeur fixée à zéro). Le modèle de régression appliqué n'est pas le modèle linéaire usuel entre les variables dépendante et indépendante, mais un modèle linéaire entre la variable indépendante et le *logit de la probabilité d'obtenir 1 sur la variable dépendante.* La formule équivalente est la suivante :

$$\ln \frac{p}{1-p} = a + bs$$

Où ln est le logarithme naturel, s le score au test, a et b les deux paramètres de la régression qui doivent être estimés. Le symbole p représente la probabilité d'atteindre le

[18] La technique abordée dans cette section utilise l'approche générale de la régression logistique, mais la façon dont les coefficients sont estimés n'est pas celle utilisée habituellement dans les techniques de régression logistique. Néanmoins, la technique présentée ici est plus simple à comprendre et ses résultats sont très utiles.

point de césure. Bien sûr, cette probabilité est inconnue, mais on peut l'approximer par la proportion des membres du panel qui ont jugé que le point de césure était atteint.

Dans le tableau 6.7., les résultats du second tour sont présentés pour les sept travaux autour du point de césure provisoire A2/B1. Notez que pour le calcul des proportions, on doit prendre en compte *toutes* les cellules indiquant que le point de césure a été atteint. En particulier, pour le score de 38, dix membres du panel ont indiqué le niveau B1 et un membre du panel a indiqué le niveau B2, ce qui correspond à un total de 11 personnes sur 15, donc à une proportion de 11/15, soit 0.733.

La régression effectue une régression linéaire simple où la variable indépendante est le score et la variable dépendante est donnée par la colonne la plus à droite du tableau 6.9. Si cette table est réalisée sous Excel, la régression peut être conduite directement dans le classeur.

Tableau 6.7: résultats du réajustement

Score	A2	B1	B2	*p*	ln[*p*/(1-*p*)]
32	10	5		0.333	-0.6931
33	11	4		0.267	-1.0116
34	9	6		0.400	-0.4055
35	7	8		0.533	0.1335
36	8	7		0.467	-0.1335
37	6	9		0.600	0.4055
38	4	10	1	0.733	1.0116

L'estimation des coefficients de régression donne :
a = - 10.3744 et b = 0.29358

L'étape finale est le calcul du point de césure lui-même à partir de ces deux coefficients. Le score de césure est conceptualisé comme le score pour lequel la probabilité d'atteindre le point de césure est exactement fixée à .5 et le logit de p = 0.5 est ln[0.5/(1-0.5)] = ln(1) = 0.

Ainsi, nous recherchons le score pour lequel nous avons :

$$\ln\frac{0.5}{1-0.5} = 0 = a + bs\,,$$

Nous en déduisons immédiatement que :

$$\text{cut-off score } = \frac{-a}{b} = \frac{10.3744}{0.29358} = 35.34\,,$$

qui sera arrondi à 35 ou 36. Dans la figure 6.2., les sept points de données (issus du tableau 6.7.) sont représentés graphiquement ensemble avec la droite de régression. Le score de césure doit être lu sur l'axe des abscisses au point où la droite de régression coupe le zéro de l'axe des ordonnées, comme l'indique la ligne verticale en pointillé.

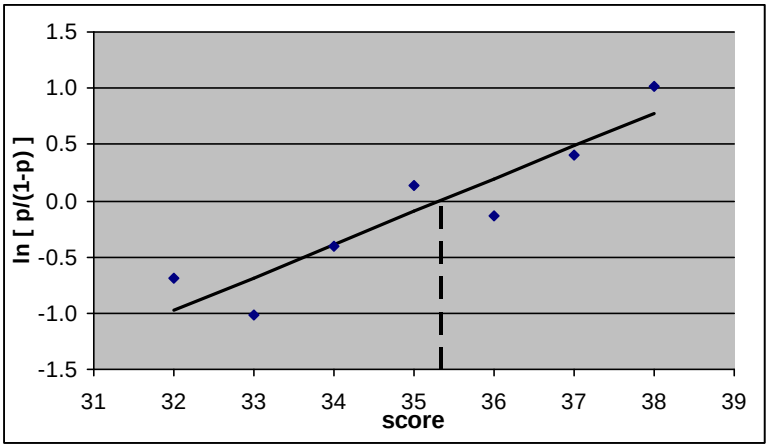


Figure 6.2. Régression logistique

6.7. La méthode d'appariement au descripteur de l'item et la méthode du panier

Dans leur livre sur les procédures de détermination des scores de césure, Cizek et Bunch (2007) rapportent le commentaire suivant pour introduire la méthode d'appariement au descripteur de l'item (p. 193) :

« Les descripteurs de niveau de compétences constituent les fondements de nombreuses méthodes modernes de détermination des scores de césure, et sont l'un des éléments clefs sur lesquels les participants comptent lorsqu'ils réalisent leurs évaluations, et ce quel que soit le pré-requis exigé par la méthode retenue pour conduire cette évaluation».

Et également,

« Dans un sens, cela ne devrait pas être une exagération de déclarer que les points de césures sont plus déterminés par les panels qui ont recours aux descripteurs que par ceux qui notent les items ou les performances. Cette assertion est la plus défendable sous deux conditions très fréquentes : 1. quand les descripteurs sont très bien détaillés et incluent des états très spécifiques de compétences pour un niveau donné de performance ; et 2. quand un panéliste impliqué dans la procédure de définition des scores de césure opérant un jugement sur un item ou une tâche d'un test s'appuie- comme il ou elle devrait - sur les descripteurs pour indiquer comment la performance est liée aux niveaux de compétences. ».

Dans le CECRL, les niveaux de compétences, de A1 à C2 (et qui peuvent être plus détaillés), sont présentés par des descripteurs de type « être capable de », resitués dans un contexte approprié et de façon encore plus élaborée par des exemples de référence. Les chapitres précédents, décrivant les activités incontournables de préparation des notations des panélistes, autant que les spécifications du test, peuvent être considérés comme un parfait exemple de l'accomplissement des conditions mentionnées ci-avant.

Les deux méthodes qui sont discutées dans cette section utilisent directement les descripteurs pour parvenir à un (ou plus généralement plusieurs) score de césure.

6.7.1. La méthode d'appariement au descripteur de l'item

Cette méthode est relativement récente. Elle a été proposée par Ferrara, Perie and Johnson in 2002[19]. La tâche demandée aux membres du panel est de placer chaque item dans les niveaux (A1, A2, etc.) auxquels ils appartiennent en respectant le principe suivant : « à quelle description de niveau de compétence (c'est-à-dire niveau du CECRL ou catégorie)

[19] En fait la méthode a été présentée en conférence en 2002 lors du meeting of the American Educational Research Association à la Nouvelle Orléans et le titre était « *Setting performance standards: the item descriptor (ID) matching method ».*

s'apparentent le mieux les connaissances, les compétences et les processus cognitifs nécessaires pour répondre correctement à cet item ? » (Ferrara, Perie and Johnson 2002, p. 10).

On déduit immédiatement de l'assertion précédente que la méthode est centrée sur le test. La tâche des panélistes est d'attribuer un niveau à chaque item. Les auteurs présentent une liste ordonnée d'items (accompagnée d'une brève description). L'ordre de présentation est un ordre croissant selon la difficulté, et un indice de difficulté est fourni. Un telle liste est appelée livret ordonné d'item (*ordered item booklet* (OIB)) dans la littérature relative aux méthodes de détermination des scores de césure. La méthode a été développée pour des situations où l'analyse dans le cadre de la TRI est utilisée pour estimer les paramètres de difficulté des items.

La procédure pour convertir ces jugements en un score de césure (pour chaque membre du panel) utilise le concept important de la *zone seuil*, qui sera expliquée à l'aide d'un exemple. Dans le tableau 6.8., un exemple fictif de jugement est proposé pour un test considéré comme approprié pour définir les points de césure A2/B1 et B1/B2. Le formulaire est légèrement abrégé parce que les descriptions des items ont été omises. La colonne la plus à droite contient les jugements d'un membre du panel. La colonne étiquetée « difficulté » contient le paramètre d'estimation de la difficulté d'un modèle de la TRI. Plus ce nombre est élevé, plus la difficulté de l'item est importante. La colonne étiquetée « item-ID » identifie l'item dans le test, de telle sorte qu'il peut être recherché pendant la procédure de jugement.

Nous supposerons que tous les jugements de ce membre du panel pour les items 1 à 10 sont soit en A1 soit en A2 et qu'après l'item 21 de tels jugements n'apparaissent plus. On constate, à partir du tableau, que selon le jugement du panéliste il n'y a pas de césure fine entre les items A2 et ceux B1. Les jugements du panéliste sont relativement conformes à l'ordre de difficulté : les items 15 et 18 sont appariés au niveau A2, alors que certains items plus faciles sont placés en B1. L'étendue des items, qui est précédée par une séquence claire et univoque de jugements au plus faible niveau et suivie par une séquence claire de jugements au niveau supérieur, est appelé la *zone seuil.* Dans cet exemple, cette étendue (intervalle) contient les items 14 à 18.

L'idée de base de cette méthode est que l'étendue du seuil, et l'étendue correspondante de la variable sous-jacente (la variable latente), indique une région où les scores de césure doivent être positionnés. Pour la variable latente, les paramètres de difficulté pourraient être utilisés, de telle manière que le score de césure se situe entre -1.63 et -1.20. Le milieu de ces deux valeurs pourrait alors être une option raisonnable. Bien sûr, chaque procédure définit un point de césure dans le domaine du score. Ainsi, le point de césure sur le trait latent doit être converti en un *score de césure*. Cette conversion est technique et est discutée dans la section 6.8.3.

Pour définir l'intervalle seuil, les auteurs de cette méthode proposent que le point de départ soit l'item qui est précédé par au moins trois jugements consécutifs au plus faible niveau. Dans l'exemple donné ici, c'est le cas pour les items 11, 12, 13 qui sont tous appariés au niveau A2. Le point final est l'item qui est immédiatement suivi par au moins trois jugements au niveau supérieur (ici les items 19, 20 et 21 qui sont appariés au niveau B1).

Tableau 6.8: exemple de réponses dans la méthode de l'appariement au descripteur de l'item (formulaire abrégé)

Rang de l'item	Item-id	Difficulté	Jugement
…	…	…	…
11	22	-2.13	A2
12	13	-2.11	A2
13	7	-1.84	A2
14	1	-1.63	B1
15	4	-1.48	A2

16	8	-1.47	B1
17	3	-1.32	B1
18	17	-1.20	A2
19	15	-1.06	B1
20	9	-.97	B1
21	19	-0.94	B1
…	…	…	…

Pour des applications en relation avec le CECRL, le succès de cette méthode semble dépendre de façon très critique de l'étroite relation entre la difficulté et le niveau des items. Idéalement, on pourrait dire qu'un item qui suppose seulement les compétences et habiletés décrites au niveau A2 est plus facile qu'un item conçu pour le niveau B1. Ce serait trop simpliste pour une théorie sur la difficulté des items. Il existe une grande variation dans la difficulté à l'intérieur des niveaux attribués aux items. Cette variation est telle que de nombreux items difficiles parmi ceux des niveaux les plus faibles sont plus difficiles que les items faciles parmi ceux des niveaux élevés. Ce n'est pas sans générer une large zone seuil et sans tendre à faire disparaitre l'aspect intuitif de la méthode.

6.7.2. La méthode du panier

Il s'agit d'une méthode qui présente de nombreuses similarités avec celle de l'appariement au descripteur de l'item et qui a été utilisée pour la détermination des points de césure dans le projet Dialang (Alderson 2005). Elle est présentée, section 5.6., dans la section entraînement à la détermination des points de césure. La similarité tient en la comparaison des ressources exigées par un item en termes de descripteurs, c'est-à-dire au sens des descripteurs « être capable de » du CECRL. La question élémentaire posée aux panélistes n'est pas un jugement sur l'item mais se centre sur un candidat abstrait ayant les compétences d'un niveau défini. La formulation élémentaire de la question est la suivante :

> *"A quel niveau du cadre un candidat peut déjà répondre correctement à cet item ?"*

Si l'envergure d'un test est large, par exemple s'il couvre tous les niveaux de A1 à C2, comme c'était le cas pour Dialang, la même question doit être posée pour chaque item de chaque niveau. Bien qu'une telle procédure présente des avantages indéniables pour examiner la validité de la méthode et ses résultats (voir chapitre suivant), elle est très chronophage et peut présenter des pertes de motivation chez les experts du panel.

C'est pourquoi un raccourci de méthode a été proposé. Les experts du panel doivent mettre chaque item dans un panier correspondant aux niveaux du CECRL. Si un item est placé dans le panier B1, cela signifie qu'une personne de ce niveau devrait donner une réponse correcte à cet item. On suppose ici que si c'est le cas les personnes de niveaux supérieurs devraient également répondre correctement à l'item. Notez que ce jugement n'implique pas que les personnes de niveaux inférieurs ne devraient pas fournir une réponse correcte ; cela signifie simplement (pour les membres du panel) qu'une réponse correcte ne devrait pas être exigée pour les candidats de niveaux inférieurs.

Notez que la tâche des panélistes dans cette méthode abrégée est la même que dans celle de l'appariement au descripteur de l'item. Dans ces deux méthodes un appariement doit être réalisé entre un descripteur (un niveau du CECRL) et les exigences requises par les items. Néanmoins, dans la méthode du panier aucune information sur la difficulté des items n'est fournie aux panélistes.

La méthode pour convertir les jugements en score de césure suppose qu'avec la méthode du panier le panéliste propose les exigences minimales requises pour chacun des niveaux. Supposons que pour un test constitué de 50 items, deux items sont placés dans le panier A1, sept dans le panier A2, 12 dans le panier B1. Pour ce panéliste, ces 21 (= 2+7+12) items

devraient être traités correctement par n'importe quel candidat de niveau B1 ou supérieur. Ce nombre, qui correspond à l'exigence minimale, est interprété comme le score de césure.

Nous proposons maintenant une courte note technique.
Pour un des panélistes, un item pourrait être jugé comme étant si difficile qu'il ne pourrait pas exiger qu'un candidat du niveau supérieur le réussisse. Au regard de la procédure, cela signifie que l'item ne s'ajuste à aucun des paniers envisagés. On peut anticiper de telles situations en ajoutant un panier supplémentaire qui serait étiqueté « supérieur à C2 ». Bien entendu, si un test vise le niveau B1, il n'est pas nécessaire de disposer de paniers pour tous les niveaux. Les trois paniers de niveaux les plus forts pourraient être nommés B1, B2 et supérieur à B2.

Il est possible que l'ajustement de l'exigence minimale au point de césure conduise à des points de césure trop indulgents. Il serait alors raisonnable de penser qu'une personne d'un niveau donné soit également en mesure de répondre correctement à des items qui exigent un niveau supérieur. Ce point n'est pas pris en compte dans la méthode, mais des études comparatives (non encore publiées) indiquent que la méthode du panier tend à produire des points de césure minorés (indulgents) par rapport à ceux obtenus avec d'autres méthodes.

En conclusion de cette section, voici quelques remarques :

- Les deux méthodes discutées dans cette section sont relativement récentes et reflètent l'importance des descripteurs de niveau de compétence, qui dans le cas du CECRL sont opérationnalisés en descripteur du type « être capable de ». Il est difficile d'imaginer que l'une ou l'autre de ces méthodes peut être raisonnablement appliquée dans le cas de procédure du type réussite/échec. En effet, pour chaque niveau de performance (A1, A2,…), la performance est décrite « positivement » (ce qu'on est capable de faire) alors qu'il n'est pas facile de décrire ce qu'on est en droit d'attendre d'une personne qui échouerait.
- En principe, ces deux méthodes peuvent être utilisées pour des items dichotomiques (QCM par exemple, ou items de type vrai/faux) mais aussi pour les questions à réponses ouvertes et pour les tâches (qui produisent des scores partiels, par exemple dans l'intervalle 0-2 ou 0-3). Ces derniers sont plus fréquents dans l'évaluation des compétences productives. On ne devrait pas sous-estimer la charge de travail impliquée par la phase d'entraînement. Par exemple, pour une tâche de production orale, un étudiant peut obtenir jusqu'à trois points. Cette tâche apparaitra trois fois dans la liste des items. La première fois comme étant une combinaison de réponses rapportant 1 point, la deuxième comme étant une combinaison de réponses rapportant deux points et la troisième comme étant une réponse permettant de bénéficier du total des points alloués à cette question, soit trois points. Dans ces trois cas, la description de la tâche sera la même, mais la qualité de la réponse différera. Pour garantir une bonne compréhension des différences, on devrait se référer à la consigne de la tâche (une partie des spécifications du test) et probablement ajouter des échantillons de réponses qui illustrent l'usage attendu de la consigne. Ce point illustre la nécessité d'avoir de bonnes consignes : on ne peut pas obtenir des points de césure qui aient du sens avec une consigne qui dirait : 0 point pour une mauvaise réponse, un pour une réponse pas trop mauvaise, deux points pour une réponse légèrement meilleure et la totalité des points pour une réponse parfaite. Pour sélectionner de bons exemples de réponses (des exemples de référence), on devrait s'assurer que les correcteurs ont également une bonne compréhension des consignes de notation et qu'ils les suivent scrupuleusement. En fait, l'intégralité du processus d'élaboration d'un test ou d'un examen, de la première étape (définition de l'objectif du test) à la dernière (définir les points de césure), est une longue chaine de décisions qui sont en étroite relation. Parce que la détermination du score de césure est la dernière étape, une négligence à l'une ou plusieurs des étapes antérieures pourrait donner le sentiment que le score de césure ne semble pas fonctionner correctement.
- Dans leur discussion au sujet de la méthode d'appariement au descripteur de l'item, Cizek et Bunch déclarent que les items devraient être présentés aux panélistes par ordre

croissant de difficulté, et plus encore, qu'un indice de difficulté devrait être fourni (comme dans le tableau 6.8.). Il est important de noter que pour la tâche confiée aux membres du panel, ces indices ne sont pas utilisés. Ils deviennent importants quand les jugements des panélistes doivent être convertis en un score de césure. Cette conversion n'est généralement pas réalisée par les experts du panel eux-mêmes, mais en aparté par l'équipe qui conduit la procédure de détermination des points de césure. Cette conversion sera discutée dans la section 6.8.3. Il pourrait être préféré de ne pas présenter de telles valeurs numériques, parce qu'elles pourraient être aisément mal interprétées et pourraient détourner l'attention des panélistes de leur tâche principale ; en l'occurrence l'appariement entre les exigences des items et le(s) descripteur(s) d'un niveau du CECRL.

- Bien que les caractéristiques formelles de cette méthode soient simples à mettre en œuvre (le formulaire de jugement est très facile à mettre en œuvre, et celui pour la méthode d'appariement au descripteur de l'item peut être téléchargé depuis le site : www.sagepub.com/cizek/IDMform), il serait illusoire de penser qu'une application rapide et précipitée de la méthode pourrait garantir des résultats pertinents et utiles. Le succès (au regard de la validité, qui sera discutée plus en détail dans le chapitre suivant) dépend de façon prépondérante de trois facteurs :
    - Premièrement, la clarté et la puissance discriminative des descripteurs.
    - Deuxièmement, de façon complémentaire au premier facteur, le degré de compréhension des descripteurs par les experts du panel. Ce point exige une phase de familiarisation avec le CECRL et une bonne standardisation au sens utilisé dans le chapitre précédent.
    - Le troisième facteur exige que les items ou les tâches du test ou de l'examen puissent être décrits de façon univoque et compris selon les descripteurs spécifiques de compétence. Les panélistes doivent comprendre clairement quel « être capable de » doit être appliqué et quel est celui qui ne s'applique pas, et ce pour chaque item ou tâche.
- La dernière recommandation est relative au nombre de tours de jugement et les raisons pour lesquelles il est vivement conseillé d'en avoir plus d'un. Un deuxième tour réalisé avec des données normatives (préparées entre le premier et le deuxième tour), montrant des cas particuliers de désaccord et invitant à l'échange en petits groupes à cet égard, n'est pas conduit pour tendre vers l'unanimité, mais pour stimuler les discussions qui conduiront à une compréhension non ambigüe du CECRL et des relations entre les descripteurs et les exigences de chacun des items ou tâches.

6.8. La méthode du marque-page

La méthode du marque-page (Mitzel et al 2001) est devenue rapidement populaire aux Etats-Unis. La plupart des aspects de cette méthode ont déjà été abordés dans les méthodes précédentes, à l'exception d'un qui sera expliqué plus en détail au cours de cette section. Nous commençons par un aperçu des caractéristiques importantes.

- La méthode est centrée sur le test et est applicable aussi bien pour les items dichotomiques que les items polytomiques (questions à réponses ouvertes).
- Les experts du panel utilisent le concept du candidat aux compétences minimales acceptables ou du candidat à la limite de deux niveaux. La procédure doit être répétée autant de fois qu'il y a de points de césure à fixer (par exemple A1/A2, A2/B1 et B1/B2 d'un même test). La charge de travail est néanmoins plus légère que celle exigée par la méthode de Tucker-Angoff. La raison est expliquée au point suivant.
- Les items ou les tâches sont présentés aux membres du panel par ordre de difficulté croissante. Les tâches à réponse ouverte apparaitront plusieurs fois dans cette liste. Par exemple, si le score peut être 0, 1 ou 2 points, la tâche apparaitra à deux reprises, une fois avec une réponse permettant d'obtenir un point et la seconde fois avec une réponse permettant d'obtenir deux points. L'ordre de difficulté des items n'est pas trivial et sera discuté dans la section 6.9. Notez que cet ordre de présentation est également utilisé dans la méthode d'appariement au descripteur de l'item, discutée dans la section 6.7.1. Les items et les tâches sont disposés dans un livret. Chaque page contient un item (dans

le cas d'items dichotomiques) ou une combinaison d'une tâche à notation partielle pour les questions à réponses ouvertes. Le contenu de chaque page sera décrit plus en détail. Dans la littérature sur la détermination des scores de césure, ce livret est appelé *Livret d'items ordonnées (Ordered Item Booklet* (OIB)).

- Le concept de maîtrise d'une tâche ou d'un item. La maîtrise est ici définie en termes probabilistes. Si un candidat maîtrise un item, on peut s'attendre à ce qu'il/elle réponde correctement avec une probabilité associée élevée. La définition exacte de cette probabilité associée élevée est arbitraire, mais dans la plupart des cas, elle est fixée à 2/3, même si certains autres préfèrent la fixer à 50% alors que d'autres la fixent à 80%. Dans la littérature sur les points de césure, le critère de maîtrise se réfère à la *probabilité de réponse*. Les membres du panel doivent décider si pour un item donné un candidat à la frontière des niveaux (pour un point de césure défini) maîtrise ou non cet item. Pour une probabilité de réponse fixée à 2/3, cela signifie qu'ils doivent décider si la personne répondra correctement dans au moins deux cas sur trois. (Si la probabilité de réponse est fixée à 80%, il faudra considérer une réponse correcte dans 4 cas sur 5). Il est important de s'assurer que les membres du panel aient bien intégré cette notion de probabilité de réponse, et une attention particulière doit être allouée à cette compréhension au cours de la phase d'entraînement. Bien qu'il n'existe pas de raison rationnelle pour retenir une valeur particulière pour ce qui concerne la probabilité de réponse, ce choix a des conséquences définitives sur les points de césure que l'on trouvera. En général, plus la probabilité de réponse est fixée à une valeur élevée, plus le point de césure le sera également.
- Pour les combinaisons de tâches à notation partielle, la probabilité de réponse a une signification particulière. Supposons que le score maximal soit de trois pour une tâche donnée. Si le score partiel est égal à un, la probabilité de réponse se réfère à la probabilité d'obtenir un score de *un ou plus.* Si le score partiel est de deux, la probabilité de réponse se réfère à la probabilité d'obtenir au moins deux points. Enfin, si le score obtenu est de trois, la probabilité de réponse se réfère à la probabilité de les obtenir.

6.8.1. Le travail du panel d'experts

Il est demandé aux experts du panel de commencer avec le point de césure le plus bas (par exemple A1/A2), de progresser dans le livret en allant du plus facile vers le plus difficile, et de décider pour chaque item si la probabilité d'une réponse a atteint le seuil fixé ou si elle est supérieure. Lorsque la réponse est affirmative, cela signifie que le candidat limite maîtrise l'item, selon le point de vue du panéliste. Parce que les jugements s'opèrent en premier lieu sur les items les plus faciles, on s'attend à ce que les réponses soient affirmatives pour quelques items à la suite, mais qu'à partir d'un item donné la réponse devienne négative. Supposons que ce soit le cas à l'item 11, alors un marque-page (ou un symbole similaire) doit être placé à cet endroit. Immédiatement, le membre du panel doit changer de point de césure, en traitant le suivant (par exemple A2/B1 ici) et continuer son travail de jugement à partir de l'item où il se trouve.

S'il y a trois points de césure à définir, le travail est en principe finalisé quand les trois marques-pages sont placés dans le livret. Cette opération devrait être réalisée bien avant le dernier item. Il est cependant d'usage d'inviter les panélistes à examiner l'ensemble des items, et même à considérer la possibilité de déplacer les marques-pages précédents au fur et à mesure qu'ils progressent dans le livret.

A chaque tour, chaque membre du panel indique son point de césure provisoire dans un tableau comme celui présenté dans la figure 6.3., pour une situation correspondant à trois points de césure. Il est préférable de laisser aux participants la possibilité d'indiquer deux numéros de page, comme dans la figure 6.3. Les pages 11/12 pour le point de césure A1/A2 signifient (pour le participant) qu'un candidat aux compétences minimales acceptables au niveau A1/A2 a au moins une probabilité (égale ou supérieure à la probabilité de réponse) de répondre à l'item 11 correctement. Ce ne sera pas le cas pour l'item 12.

Les informations collectées à l'issue du premier tour, par l'équipe organisatrice de la procédure de définition des points de césure, vont être utilisées pour le tour suivant et la décision finale.

Tour 1			
Point de césure:	A1/A2	A2/B1	B1/B2
Numéro des pages:	11/12	24/25	38/39

Figure 6.3: Formulaire d'enregistrement des jugements des panélistes dans le cadre de la méthode du marque page

6.8.2. Contenu des livrets d'items ordonnés

Chaque page du livret d'items ordonnés contient les informations suivantes :

- Le numéro de la page du livret. Ce nombre doit être placé en évidence dans le coin supérieur droit de la page puisque c'est la position que les membres du panel doivent indiquer pour leurs jugements.
- La position de l'item dans le test ou dans l'examen (coin supérieur gauche). Si l'item le plus facile est l'item 5, le coin supérieur gauche portera le numéro « item 5 », alors que le coin supérieur droit aura le statut « 1 » parce qu'il est le plus facile et qu'il a cette position dans le livret. Dans le cas d'items à notation partielle, une double référence est nécessaire. Par exemple, « item 13-2 » fait référence à l'item 13 qui permet de bénéficier de deux points. Si trois points peuvent être acquis à cet item, il y aura alors trois pages portant les références suivantes « 13-1 », « 13-2 » et « 13-3 » respectivement.
- Au centre et en haut de chaque page, la probabilité de réponse et l'échelle de valeur de cette probabilité de réponse seront indiquées de la façon suivante :
    - Pour les items dichotomiques : « Niveau d'accomplissement requis pour une réponse correcte dans 2 cas sur 3 = -1.84 ». La probabilité de réponse est fixée à 2/3 et la valeur sur le trait latent d'obtenir une réponse correcte est -1.84. La section 6.8.3. explique comment on calcule cette valeur.
    - Pour les items à crédit partiel (comme avec les questions à réponse ouverte) le texte sera : « niveau d'accomplissement requis pour une réponse correcte d'obtenir 2 points ou plus dans 2 cas sur 3 = 1.38 ». Cette mention apparaitra avec l'identifiant de l'item « nn-2 ». Pour le score le plus élevé qu'il soit possible d'obtenir à cet item, la précision « ou plus » sera enlevée.

- Le texte de l'item ainsi que :
    - pour les QCM, les réponses alternatives (distracteurs) ;
    - pour les items à crédit partiel, la consigne de notation pour obtenir le score partiel. Il est également conseillé dans cette situation de fournir la règle qui permet d'obtenir un point en moins et un point de plus. Ainsi les panélistes pourront voir les différences de notation sur une seule et même page de l'OIB.
- Les réponses correctes :
    - Pour les QCM, ce sera la clef.
    - Pour les items à crédit partiel, un ou plusieurs échantillons permettant d'obtenir le score spécifié pour aider les membres du panel à se focaliser sur la signification précise de ce score.
- Une référence à un livret source :
    - Pour un test de réception écrite où plusieurs items sont proposés à partir d'un seul texte, il est conseillé de rassembler tous les textes dans un

« livret source ». Par exemple, avec de nombreux paragraphes, il conviendra d'indiquer le(s) paragraphe(s) pertinent(s) dans le coin inférieur droit de la page du livret.

- Pour les tests de réception orale, les choses sont un peu plus compliquées. Un ordinateur doit être disponible pour chaque panéliste pour leur permettre d'écouter les items autant de fois qu'ils en ont besoin.

6.8.3. Aspects techniques

Sur les valeurs de la probabilité de réponse dans la méthode du marque-page.
La méthode de l'appariement au descripteur de l'item et la méthode du marque-page ont été développées dans le cadre de tests calibrés à l'aide de la TRI. Elles font usage des résultats de ces mesures. Nous illustrons ce point avec une situation d'items dichotomiques, calibrés à l'aide du modèle de Rasch. Les détails pour les items à crédit partiel pourront être consultés dans l'ouvrage de Cizek & Bunch (2007, Chapitre 10).

Dans le modèle de Rasch, la fonction de réponse à l'item est donnée par :

$$P(X_i = 1 | \theta) = \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)} \quad (1)$$

Où β_i est le paramètre de difficulté de l'item *i.* (Sa valeur est connue à partir de la détermination du point de césure). Considérons en premier lieu le cas ou l'habileté est égale à la difficulté de l'item ($\theta = \beta_i$), alors nous écrirons la formule (1) de la façon suivante :

$$P(X_i = 1 | \theta = \beta_i) = \frac{\exp(\beta_i - \beta_i)}{1 + \exp(\beta_i - \beta_i)} = \frac{\exp(0)}{1 + \exp(0)} = \frac{1}{1+1} = \frac{1}{2}$$

Ce qui signifie pour un trait latent égal à la difficulté de l'item que la probabilité de répondre correctement équivaut à .5, et inversement. Si la probabilité de réponse est fixée à ½, l'habileté requise pour avoir la maîtrise est égale à la difficulté de l'item.

Si l'on fixe la valeur de la probabilité de réponse à *p,* on recherche alors la valeur pour θ, tel que :

$$\frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)} = p$$

La solution est donnée par :

$$\theta = \beta_i + \ln\left[\frac{p}{1-p}\right]$$

Ou « ln » est le logarithme naturel. Si *p=2/3*, nous avons (2/3)/(1/3) = 2 and ln(2) = 0.693, et nous trouvons $\theta = \beta_i + 0.693$. C'est cette valeur qui sera imprimée sur les pages intérieures du livret comme la valeur du niveau d'accomplissement (voir section 5). Notez que l'augmentation à 2/3 de la probabilité de réponse (comme seuil d'exigence de maîtrise) fait augmenter la valeur de l'échelle à 0.693 logits. Si l'on porte la probabilité de réponse à une valeur de ¾, l'augmentation est ln(3) = 1.098, et pour une probabilité de réponse de 4/5, l'augmentation est la suivante : ln(4)=1.386

Le point de césure provisoire dans la méthode du marque-page
En exemple, les numéros des pages sont représentés dans le tableau 6.9., accompagnés du niveau de réussite pour une probabilité de réponse de .5 (deuxième colonne) et pour une probabilité de réponse de 2/3 (colonne la plus à droite). La différence entre les deux dernières colonnes est ln(2)=0.69. Considérons que la probabilité de réponse est fixée à .5 et que quelques membres du panel ont apposé leurs marques-pages pour A1/A2 en position 13/14. Cela implique, selon ces panélistes, que le candidat aux compétences minimales

acceptables maîtrise (avec une probabilité de réponse à .5) les items 1 à 13, mais pas l'item 14. Autrement dit, le niveau (habileté latente) de ce candidat doit se situer entre -1.84 et -1.63. Généralement, on prend en compte la plus petite de ces deux valeurs provisoires. Notez que ce point de césure provisoire est une valeur exprimée sur l'échelle latente. Ensuite, on rassemble les points de césure provisoires (puis on calcule la moyenne, tronquée ou non, ou la médiane) pour parvenir au point de césure collectif exprimé sur l'échelle latente

Convertir les points de césure sur l'échelle latente en un score de césure
La façon la plus simple d'effectuer cette conversion sur l'échelle latente en score de césure est d'utiliser une table qui offre de bonnes estimations de la valeur latente pour tous les scores possibles du test. Un exemple est proposé dans le tableau 6.10. Supposons que le point de césure soit de -1.35 sur l'échelle latente. A partir de la table, on peut constater qu'un score de 9 (items corrects) conduit à une valeur latente estimée à -1.409, inférieure à celle recherchée, alors qu'un score de 10 a une valeur correspondante de -1.257, supérieure à la valeur recherchée. On en déduit que le score de césure se situe entre 9 et 10 et que cette valeur doit être arrondie en prenant en compte les faux positifs et faux négatifs dont il a été question à la section 6.3.4.

Tableau 6.9: marque-page et niveaux de réussite

Numéro de page	Niveau de réussite pour RP = 0.5	Niveau de réussite pour for RP = 2/3
…	…	…
11	-2.13	-1.44
12	-2.11	-1.42
13	-1.84	-1.15
14	-1.63	-0.94
15	-1.48	-0.79
…	…	…
19	-1.32	-0.63
20	-1.20	-0.51
21	-1.03	-0.34
…	…	…

Tableau 6.10: estimation de la valeur Theta

Score	Estimated theta
…	…
5	-2.153
6	-1.938
7	-1.746
8	-1.571
9	-1.409
10	-1.257
11	-1.114
12	-0.977
13	-0.845
14	-0.717
15	-0.592
16	-0.471
17	-0.351
…	…

Il reste à déterminer quelle est l'estimation de la variable latente que l'on doit utiliser. Dans la section G.7. du Supplément au manuel, plusieurs estimations sont discutées. Il a été montré que l'estimation de la probabilité maximale était sujette à de sérieux biais. Il est donc conseillé d'utiliser l'estimateur Warm, contrairement à ce que Cizek et Bunch suggèrent[20]. C'est particulièrement important dans les cas de scores extrêmes, qu'ils soient faibles ou forts.

Un problème supplémentaire avec la méthode d'appariement au descripteur de l'item
Dans la méthode du marque-page, la valeur de probabilité de réponse est clairement expliquée aux panélistes. Ce point est essentiel parce que plus la probabilité de réponse est élevée plus strict sera le point de césure ; les experts du panel doivent donc être parfaitement conscients de la signification de la probabilité de réponse.

Au contraire, dans la méthode d'appariement au descripteur de l'item, le concept de probabilité de réponse n'entre pas en jeu parce que les panélistes doivent seulement indiquer à quel niveau (A1, A2, etc.) correspond le mieux chaque item. A partir du niveau de difficulté reporté dans le tableau 6.9. (troisième colonne) on peut déduire s'il s'agit des paramètres de difficultés ou du niveau de réussite pour d'autres valeurs de probabilité de réponse que .5. Comme il a été dit précédemment, ces valeurs numériques ne sont pas utilisées pour la tâche de jugement par les membres du panel, au-delà du fait qu'ils indiquent l'ordre des items au regard de leur difficulté. Toutefois, une fois que la *zone seuil* a été déterminée, ces valeurs jouent un rôle majeur par ce qu'elles sont utilisées pour déterminer le seuil provisoire (pour chaque panéliste) et enfin pour calculer le seuil pour l'ensemble des membres.

Nous pouvons entrevoir le problème en imaginant deux groupes de panélistes bien préparés. Dans un groupe, les niveaux de difficultés sont fournis pour une probabilité de réponse correspondant à celle enregistrée via le modèle de Rasch, dans l'autre groupe, les niveaux de difficultés sont fournis par les paramètres de difficulté plus ln(2), autrement dit pour une probabilité de réponse de 2/3. La tâche fondamentale des panélistes étant d'apparier les exigences de l'item aux niveaux du CECRL, on peut s'attendre à ce que les zones seuil des deux groupes ne soient pas systématiquement différentes et qu'elles ne soient pas influencées par les valeurs fournies pour chaque item. Cependant, le calcul du point de césure effectué à partir des valeurs de difficulté différera approximativement de

[20] Dans la littérature, il est conseillé d'utiliser la fonction caractéristique du test pour convertir les valeurs latentes en score. Dans le modèle de Rasch et dans le modèle à deux paramètres, cette conversion est la même que dans l'estimation de la probabilité maximale. L'estimation Warm est proposée dans le logiciel OPLM, disponible sur simple demande à norman.verhelst@cito.nl

0.693(=ln(2)) entre les deux groupes. Plus généralement, cela indique que les points de césure définis sont arbitraires pour une large partie, et qu'ils dépendent des valeurs que l'on aura décidé d'utiliser pour les niveaux de difficultés.

6.9. Variante de la méthode du marque-page selon le Cito

La méthode du marque-page peut être plus compliquée si les items ne discriminent pas de façon identique (ce qui est souvent le cas). Un exemple, avec deux items, est proposé dans la figure 6.4 pour illustrer ce propos. La courbe en pointillé représente le meilleur taux de discrimination de l'item. Les courbes en trait plein représentent la fonction de réponse de l'item : elles relient l'échelle latente (axe des abscisses) à la probabilité d'obtenir une réponse correcte (axe des ordonnées).

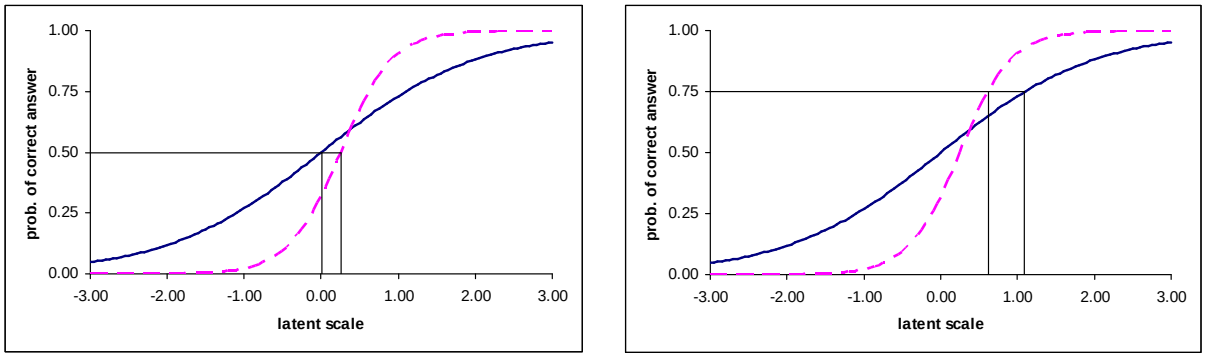


Figure 6.4: items de discrimination différente

Si l'on utilise la méthode du marque-page avec une probabilité de réponse fixée à .5 (partie gauche du graphique), la courbe en pointillé aura un numéro de page plus élevé (car correspondant à un niveau de difficulté plus important) dans le livret que l'autre item, alors qu'avec une probabilité de réponse à .75 (partie droite du graphique), l'inverse se produit. En l'occurrence, la courbe en pointillé fait maintenant apparaitre un item plus facile. Cette remarque illustre le fait que la « difficulté d'un item » n'est pas un concept trivial. En fait la présentation de l'ordre des difficultés aux panélistes par une simple valeur numérique pourrait provoquer une certaine confusion.

La méthode développée au Cito (Van der Schoot 2001) vise à présenter graphiquement les valeurs de difficulté et de discrimination de tous les items dans un seul et même graphique. Considérons l'item moins discriminant de la figure 6.4. :quand la probabilité de réponse est de .5, le niveau exigé d'habileté est de 0 (partie gauche du graphique), alors qu'il est d'environ 1.1 pour une probabilité de réponse fixée à .75 (partie droite du graphique). On pourrait fixer la probabilité d'avoir une réponse correcte à 50% pour désigner la « maîtrise limite » alors qu'une probabilité à 75% désignerait la « pleine maîtrise ». Pour aller de la « maîtrise limite » à la « pleine maîtrise » l'habileté doit croitre de 0 à 1.1. Il est possible d'en faire une représentation graphique comme dans la figure 6.5. où est proposée une cartographie d'items pour 16 items et qui comprend des informations relatives à la difficulté et à la discrimination de chacun des items. Chaque item est représenté par un segment de droite horizontal. L'extrémité gauche du segment correspond au paramètre de difficulté de l'item (probabilité de réponse de 50%) et la longueur du segment indique la valeur discriminative : plus la ligne est longue moins l'item est discriminant. L'extrémité droite du segment correspond à une probabilité de réponse plus élevée, en l'occurrence 0.75 ou 0.80. La représentation est construite de telle sorte que l'extrémité gauche des segments augmente au fur et à mesure qu'on se déplace du bas vers le haut du graphique. On doit rester vigilant à l'identification des segments pour que les panelistes puissent associer clairement chaque segment à un item du test.

Le trait vertical représente le point de césure provisoire d'un membre du panel. En apposant ce trait, le panéliste peut bénéficier rapidement d'un aperçu des conséquences de sa décision. Dans l'exemple proposé, le point de césure implique une « pleine maîtrise » des items 1 à 8 et de l'item 11. Pour les items 9 et 10, la « maîtrise totale» est quasiment

atteinte. Pour l'item 12, la « maîtrise limite » a été atteinte, et pour les items 13 à 16 la « maîtrise limite » n'est pas du tout atteinte.

Pour mettre en œuvre cette méthode, on peut demander aux experts du panel de représenter un trait vertical, ou de donner une valeur numérique qui correspond à l'intersection entre le trait vertical et l'axe horizontal dans la figure (dans l'exemple il s'agit de la valeur 0.6).

Notons que l'on ne peut pas déduire à partir de la figure 6.5., la forme de la distribution des habiletés latentes dans la population cible. Pour éviter des associations, par exemple avec une distribution normale, il est préférable de changer l'échelle des valeurs qui sont représentées le long de l'axe horizontal dans la figure en une échelle n'incluant aucune valeur négative et dont l'unité est facilement compréhensible. Par exemple, en ajoutant 8 à tous les nombres représentés le long de l'axe dans la figure, et en multipliant par 10, on obtiendra des nombres de 50 à 110, évitant ainsi des interprétations en termes de pourcentage, et dont le pas de l'échelle permettra d'obtenir des nombres entiers pour les points de césure provisoires[21]. Une fois la définition du score de césure effectuée, le score résultant peut être facilement reconverti dans son échelle originelle, et les points de césure sont alors déterminés de la même façon que dans la méthode du marque-page (voir section 6.8.3.)

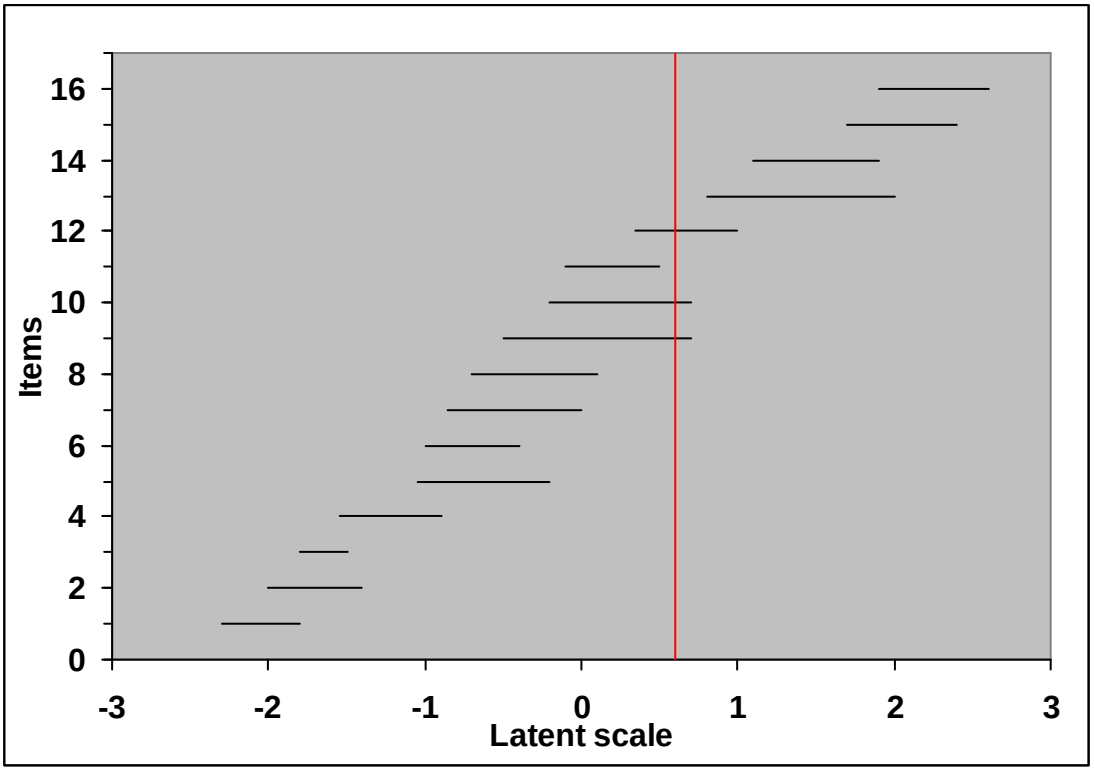


Figure 6.5: cartographie d'items, indiquant la difficulté et la discrimination

[21] Une approche alternative ou supplémentaire serait d'inclure les descripteurs pertinents du CECRL pour les tâches du test dans le pilotage ou le pré-testing comme des items pour l'évaluation d'enseignant ou l'auto-évaluation et ainsi de montrer aux panélistes, à un moment approprié dans les tours de détermination du point de césure, où ils apparaissent calibrés sur l'échelle du trait latent comme indiquée dans la figure 6.5. (Voir la section 7.5.4.2.)

6.10. Déclinaisons particulières

Dans cette conclusion, certains aspects particuliers seront brièvement abordés. Ces points sont :

- La détermination des scores de césure avec des tests hétérogènes (sur plusieurs capacités) ;
- La détermination des scores de césure et ajustement des tests (à travers les administrations) ;
- La détermination des scores de césure sur plusieurs langues.

6.10.1. Définition des scores de césure sur plusieurs capacités langagières

Dans certains cas, il peut être exigé de définir un seul point de césure, le résultat global comme le niveau du CECRL d'un candidat, alors que le test lui-même peut comprendre trois voir plus de trois parties, où chacune permet de tester les performances dans des capacités différentes.

Il y a alors plusieurs façons de voir les choses. Deux points de vue seront ici discutés, une approche compensatoire et une approche conjonctive. Lorsque ces deux approches sont appliquées de façon stricte, elles peuvent conduire à des résultats inacceptables ; une solution raisonnable est donc également discutée.

Approche compensatoire : d'une part, selon une position extrême, on peut considérer toutes les tâches et tous les items comme un mélange des capacités et alors appliquer l'une des méthodes abordées précédemment sur l'ensemble des items et des tâches simultanément. En procédant de la sorte, on doit concevoir que les scores au test sont, par définition, compensatoires puisqu'ils sont les sommes des scores aux items et aux tâches. Echouer à certaines tâches peut être compensé par une bonne performance à d'autres tâches. Dans la mesure où le test est homogène du point de vue de la nature des tâches, un tel mécanisme compensatoire est légitime et ne doit concerner précisément que les items et les tâches qui sont échoués ou non.

Néanmoins, avec un test plus hétérogène, ce point de vue compensatoire pourrait être inadéquat. Par exemple, supposons qu'un examen national pour l'anglais, composé d'une épreuve de réception écrite, d'une épreuve de réception orale, d'une épreuve de production orale et d'une épreuve de production écrite, ait un score maximal de 100 points sur l'ensemble des quatre parties. En outre, supposons que la méthode du corpus de productions soit appliquée pour fixer les scores de césure et qu'on ait pris le soin de collecter des échantillons de production en provenance de différentes régions dans le pays. Si les régions diffèrent significativement dans leurs méthodes et du point du vue de leur expertise pour une ou plusieurs compétences, des profils typiques de compétences devraient révéler les différentes structures selon les régions. Si dans certaines régions une faible attention est allouée à la production orale, même les meilleurs étudiants de cette région pourraient être qualifiés comme faibles dans cette compétence et réussir au même niveau que l'étudiant moyen des régions où une attention plus importante aura été allouée à cette compétence. Prendre en compte l'ensemble des compétences pourrait masquer des différences importantes dans les profils.

Par conséquent, il est important qu'une étude minutieuse soit entreprise pour examiner dans quelle mesure une approche unidimensionnelle est appropriée. En plus de l'étude de la structure des différentes compétences, des différences structurelles possibles entre écoles, régions ou méthodes utilisées qui pourraient révéler des items à fonctionnement différentiel (DIF) devraient être examinées avant que l'approche unidimensionnelle puisse être justifiée. Si des différences marquées ou des corrélations moyennes entre compétence étaient avérées on devrait être confronté à plusieurs problèmes. Deux d'entre eux sont discutés ci-après :

1. Une décision rationnelle doit être prise sur la pondération qui sera attribuée à chaque capacité pour le score total. S'il y a une disposition légale qui stipule que chacune des capacités doit avoir le même poids, le problème est alors soldé.
2. Toutefois, même avec une pondération imposée, nous ne disposons d'aucune garantie, pour les méthodes centrées sur les candidats, comme avec la méthode du corpus de productions par exemple, que les panélistes utiliseront effectivement cette pondération définie a priori pour fournir un jugement holistique du niveau de l'étudiant.

Approche conjonctive : l'alternative est une approche qui prend en compte séparément chacune des capacités et qui implique que chaque point de césure soit défini indépendamment. La règle de décision conjonctive stipule que l'on a globalement atteint un niveau donné si l'on a atteint ce niveau pour chacune des capacités. L'application stricte de cette règle pourrait conduire à des résultats inacceptables. Par exemple, un étudiant pourrait se voir refuser le niveau B1, même si il a atteint le niveau B2 dans trois des quatre compétences et pas le point de césure A2/B1 dans la quatrième.

Dans ce cadre, un compromis entre les règles de compensation et les règles conjonctives semble raisonnable. Une règle conjonctive générale pourrait être fixée à laquelle on pourrait adjoindre des exceptions compensatrices, comme pour l'exemple ci-dessus où il apparaitrait raisonnable d'attribuer le niveau B1 à cet étudiant. La nature exacte des exceptions compensatrices doit être considérée avec une grande vigilance. Une bonne façon de les appréhender serait d'en discuter avec les experts du panel après qu'ils ont statué sur les points de césure pour chacune des capacités séparément.

6.10.2. Définition des scores de césure et ajustement de tests

Etant donné que la procédure pour déterminer les scores de césure est un dispositif onéreux, cela vaut la peine de chercher comment éviter une somme importante de travail, en particulier pour les examens cycliques où les spécifications des tests se répètent généralement d'année en année sans modifications majeures.
Si une procédure de définition des points de césure a été effectuée selon les règles de l'art pour un examen annuel, les résultats de la détermination du score de césure pourraient être transférés tels quels à une même forme d'examen (par exemple de l'année suivante) en appliquant la technique dite de l'ajustement de tests[22]. L'ajustement de tests désigne un ensemble de techniques permettant d'avoir pour chaque score dans un test un score équivalent dans un autre test. Supposons que le point de césure A2/B1 ait été fixé pour la première année de l'examen à 35 points. Si le score équivalent à 35 est 37 pour l'examen de la deuxième année, cela implique nécessairement que le score de césure soit alors de 37 points.

La mise en œuvre des ajustements de tests présente deux aspects auxquels il faut accorder une attention toute particulière. Le premier est quasi-exclusivement de nature technique, le second est conceptuel.

Pour appliquer les techniques d'ajustement de tests, il est essentiel que les deux échantillons de candidats ayant pris part aux examens soient comparables. Une telle comparabilité peut être garantie soit par l'utilisation d'items en commun dans les deux examens ou en prenant des dispositions de telle sorte que les deux échantillons soient statistiquement équivalents. Aucune approche ne peut être mise en œuvre aisément dans un contexte d'examen : généralement il n'est pas possible de répéter l'examen de l'année précédente lors de l'année en cours pour des raisons de confidentialité, et l'équivalence des échantillons n'est pas simple à obtenir étant donné que les étudiants ne peuvent être assignés à un examen de façon aléatoire. Une population légèrement plus compétente qu'une autre (que la précédente ou que la suivante) donnera le sentiment que l'examen était plus facile qu'il ne l'est en réalité. Si cela n'est pas clairement identifié et si les populations

[22] Pour une bonne introduction, on peut consulter l'ouvrage de Kolen & Brennan (2004).

sont considérées comme identiques au regard de leurs compétences, cela conduirait à des points de césure absolus.

L'utilisation de techniques des MRI exige que les deux examens soient ancrés d'une certaine façon, ce qui signifie que les parties des deux examens ont été administrés à un échantillon d'étudiants. (Voir la section G au Supplément du Manuel pour plus de détails ; voir aussi la section 7.2.3.)

L'aspect conceptuel est lié à la validité de construit des deux examens. Bien que le recours aux mêmes spécifications soit nécessaire pour obtenir des construits équivalents, il pourrait se révéler non suffisant étant donné que personne n'a une compréhension complète de la composition des construits mesurés par les examens de langue. Les techniques pour examiner la dimensionnalité d'un test complexe comme l'analyse factorielle (section F du Supplément au manuel) pourraient permettre de bénéficier ici de réponses.

Toutefois, la façon la plus prudente de garantir la validité du transfert des points de césure par ajustement est d'accomplir une définition des points de césure sur le nouvel examen, pour vérifier si les points de césure obtenus par application de l'équation d'ajustement correspondent effectivement aux points de césure fixés par un panel indépendant de juges experts.

6.10.3. Définition des scores de césure sur plusieurs langues

L'aspect probablement le plus stimulant dans le fait de relier les examens au CECRL est de trouver des méthodes qui montrent que les examens dans différentes langues sont liés d'une manière comparable à des standards communs.

Bien qu'il soit théoriquement possible d'administrer deux examens dans des langues différentes à un même échantillon de candidats, cela supposerait que chaque candidat de l'échantillon a le même niveau de compétences dans chacune des langues, ce qui est clairement impossible. Par conséquent, les méthodes qui doivent être recherchées doivent considérer que chaque candidat n'a participé qu'à un seul des deux examens et que les performances de chaque candidat dans les différentes langues seront traitées comme celles de candidats sans lien particulier.

Pour relier ces deux examens aux CECRL, on peut faire appel à des panélistes plurilingues, qui pourront offrir des jugements sérieux et dignes de confiance à la fois sur les items (pour les méthodes centrées sur les tests) et sur le travail des candidats dans les deux langues. La méthode du corpus de productions pourrait être l'une des méthodes à privilégier pour ce dernier cas. Pour les méthodes centrées sur le test, toute méthode qui ne présuppose pas une définition des points de césure selon les modèles de réponse à l'item (MRI) peut être en principe utilisée. Les méthodes reposant sur les MRI ne fonctionneront pas parce qu'il est impossible de rapporter les deux examens sur la même échelle, et ce parce que la conception ne sera pas liée par des personnes communes (voir ci-avant) ou des items en communs.

Etant donné que les procédures de détermination des points de césure sur plusieurs langues sont relativement récentes[23], une attention particulière doit être attribuée au risque de non validité de la procédure. En particulier, il conviendra d'être vigilant à l'égard des points suivants :

- Parce qu'il est impossible de masquer la langue du test aux experts du panel, excluant ainsi les jugements « aveugles », il est important qu'il n'y ait pas trop de différences

[23] Un séminaire de définition des points de césure sur plusieurs langues a été organise par le CIEP à Sèvres du 23 au 25 Juin 2008. Au cours de cette manifestation, des échantillons d'adolescents français parlant anglais, allemand, français, italien et espagnol ont été évalués sur les niveaux du cadre par des équipes multilingues. Un rapport de ce séminaire est disponible sur le site du conseil de l'Europe (www.coe.int/lang)

systématiques dans le construit entre les tests des deux langues. Par conséquent, une attention particulière doit être accordée au fait que les deux examens ou tests ont des spécifications identiques, voire similaires.

- On doit être vigilant à la composition du panel pour avoir un « équilibre d'expertise » dans les deux langues. Si les deux tests sont en anglais et en français, l'attention doit être orientée vers la langue et la formation des panélistes. Par exemple, la moitié d'entre eux pourraient être de langue maternelle anglaise, l'autre moitié de langue maternelle française, ou bien un équilibre doit être recherché pour la tâche d'évaluation principale : la moitié des membres du panel seront des enseignants de français avec une certaine aisance en anglais et vice versa pour l'autre moitié.
- Cet équilibre doit être maintenu constant dans les sous-groupes du panel qui seront formés pour les discussions.
- De façon similaire, le matériel qui va être évalué (soit des échantillons de production, soit des items) devrait être présenté de manière équilibrée du point de vue des séquences de présentation et du point de vue du contenu.
- Les étapes doivent être respectées au cours de la phase d'entraînement de définition des scores de césure pour s'assurer que tous les membres du panel appliqueront le même standard à chacune des langues. Les usages des panélistes peuvent présenter des risques, des distorsions pouvant se produire, en lien avec les publications de référence et les différences terminologiques associées à des cultures pédagogiques différentes. Il est primordial que les membres experts du panel utilisent et se réfèrent aux critères officiels et non aux standards internes.
- Des enregistrements détaillés de la procédure doivent être conservés, et autant que faire se peut, les résultats de la procédure de détermination des points de césure sur les deux langues (approche bi langue) devraient être comparés aux résultats obtenus lors de la détermination des points de césure sur chacune des deux langues (approche mono langue) avec des panélistes indépendants.

6.11. Conclusion

Ce chapitre a passé en revue un certain nombre de procédures de définition des scores de césure, mais ne prétend pas l'avoir fait de façon exhaustive. Une présentation accessible peut être consultée dans la section B du Supplément au manuel et des procédures additionnelles exploitant les jugements des enseignants et la TRI pour inclure un critère externe dans l'étude de l'appariement sont présentées dans les annexes fournies par Brian North et Neil Jones. Dans ce chapitre, l'accent a porté sur la faisabilité et l'adéquation des méthodes sélectionnées aux tests langagiers et pour relier les examens au CECRL en soulignant l'importance d'une bonne compréhension des notions de base.

Bien entendu, au cours et après la mise en œuvre de ces procédures, il sera nécessaire d'en suivre la qualité en se centrant sur plusieurs questions :

- Est-ce que la procédure de détermination des points de césure a eu les effets attendus ? La formation a-t-elle été efficace ? Est-ce que les panélistes se sont sentis libres de suivre leurs propres intuitions ? Des questions similaires sont ici bienvenues. Ce sont les questions liées à la validité procédurale.
- Est-ce que les évaluations des experts du panel sont fiables : est-ce que chaque membre du panel a été régulier au cours des différentes tâches qu'il a réalisées ? Est-ce que les membres du panel ont été en accord avec les autres dans leurs jugements et dans quelle mesure un consensus a permis de considérer le point de césure comme définitif ? Est-ce des erreurs ont été commises dans les scores au test ? Ces questions, et leurs réponses, constituent la validité interne de la procédure de définition des scores de césure.
- La question la plus importante est de savoir si les résultats de la procédure conduisant aux points de césure – qui attribuent un niveau du CECRL aux étudiants sur la base de leur score au test – sont dignes de confiance. La réponse à cette question vient de la preuve indépendante qui corrobore les résultats d'une procédure particulière de définition des scores de césure. C'est la tâche de tout un chacun que d'appliquer une telle

procédure pour fournir une réponse à cette question ; c'est précisément ce qui est signifié par le terme validation. Une telle preuve peut provenir de sources différentes, comme :

- **la validation croisée :** la répétition des procédures de détermination des scores de césure avec des groupes indépendants de panélistes ;
- **la détermination complémentaire des scores de césure :** mettre en place des méthodes indépendantes de détermination des points de césure en utilisant une procédure différente et appropriée au contexte ;
- **La validation externe :** en conduisant une étude indépendante pour vérifier les résultats de la procédure de détermination des scores de césure en les rapprochant d'un critère externe. Ce critère externe peut être un test pour la (les) même(s) compétence(s), connue(s) pour être fidèlement calibrée(s) au CECRL. Cela pourrait être également les jugements d'enseignants ou d'apprenants formés aux descripteurs du cadre.

Toutes ces questions sont traitées dans la section 7.5.

Les utilisateurs du Manuel devraient considérer :

- *La nécessité de lectures supplémentaires sur les procédures de détermination des scores de césure.*
- *Quelle(s) méthode(s) est (sont) la(les) adaptée(s) au contexte.*
- *S'il faut opter pour une méthode évaluant la difficulté des items (ex : l'appariement au descripteur ou la méthode du panier) ou pour une méthode évaluant le score de césure sur l'échelle du pré-test (ex : méthode du marque-page, méthode du corpus de productions).*
- *Si deux méthodes devraient être utilisées pour la validation de leurs résultats respectifs.*
- *Comment les panelistes proposeront leurs évaluations sur les points de césure après le premier tour ; est-ce que le vote électronique[24] est réalisable ?*
- *Si les paramètres de difficulté de la TRI seront disponibles pour renseigner le procédé permettant la précision des points de césure ou si les valeurs de probabilité devront être utilisées.*
- *Quels types de données d'impact sur les effets provisoires du point de césure devraient être disponibles pour enrichir les derniers tours de discussion.*
- *Quel(s) type(s) de moyen(s) devrai(en)t être nécessaire(s) pour appliquer la(les) méthode(s) retenue(s).*

[24] Pour information sur la mise en œuvre du vote électronique, voir Lepage and North (2005).

Chapitre 7 : Validation

7.1. Introduction

7.2. Pré-requis : la qualité de l'examen

- **7.2.1. Validité de contenu**
- **7.2.2. Aspects opérationnels : le test pilote**
- **7.2.3. Aspects opérationnels : le pré-test**
- **7.2.4. Considérations psychométriques**
- **7.2.5. Le bon moment pour déterminer les scores de césure**

7.3. Validité procédurale de la formation à la standardisation et à la détermination des scores de césure

7.4. Validité interne de la détermination des scores de césure

- **7.4.1. Consistance intra-juge**
- **7.4.2. Consistance inter-juges**
  - **7.4.2.1. Accord et consistance**
  - **7.4.2.2. Trois mesures d'accord**
  - **7.4.2.3. Evaluation des indices d'accord**
  - **7.4.2.4. Repérer les items problématiques**
  - **7.4.2.5. Indices de consistance**
- **7.4.3. Exactitude et consistance de la méthode de détermination des scores de césure**
  - **7.4.3.1. Erreur standard du score de césure**
  - **7.4.3.2. Une situation paradoxale**
  - **7.4.3.3. Exactitude et consistance des décisions**

7.5. Validation externe

- **7.5.1. Validation croisée**
- **7.5.2. Comparaison des distributions marginales**
- **7.5.3. Tables de décision**
- **7.5.4. Quelques scénarii**
  - **7.5.4.1. Tirer parti du calibrage de la TRI**
  - **7.5.4.2. Utilisation des "Etre capable de"**
  - **7.5.4.3. Détermination des scores de césure sur plusieurs langues**

7.6. Conclusion

7.1. Introduction

Relier un examen au CECRL est un processus complexe qui implique plusieurs étapes, qui toutes exigent du professionnalisme. La validation a trait au corpus de preuves proposé pour convaincre les utilisateurs du test que le processus, dans sa globalité, et ses résultats sont dignes de confiance. Les utilisateurs du test doivent ici être compris dans un sens très large ; ils comprennent les élèves (ou leurs représentants légaux, comme les parents) qui passent le test, les autorités éducatives et politiques qui utilisent les résultats du test pour prendre des décisions politiques, les éditeurs de manuel et les enseignants, les organismes certificateurs, les employeurs et les formations syndicales, la communauté scientifique impliquée dans les tests de langue, et si les enjeux sont véritablement forts, également les autorités légales. Bien que le présent Manuel se concentre sur le procédé pour relier les examens au Cadre, dans un sens plutôt strict, en mettant l'accent sur l'application d'une ou plusieurs procédures de détermination des scores de césure, il serait erroné de considérer que le processus de validation peut être totalement restreint aux activités et résultats décrits au cours des chapitres 3 à 6. Dans ce présent chapitre, la plupart des procédures et techniques discutées se centrent sur l'adaptabilité du procédé qui permet de relier les examens au Cadre. Néanmoins, une section indépendante (7.2.) est consacrée au prérequis généraux se rapportant à la qualité de l'examen, le test pilote, le pré-test, les considérations psychométriques et au choix du moment approprié pour conduire une procédure de détermination des scores de césure.

La discussion autour de la validation est organisée en trois sections ; deux d'entre elles traitent de la validité ou de la fiabilité de la procédure elle-même et de ses résultats élémentaires. Dans la section 7.3., la **validité procédurale** est discutée et dans la section 7.4. une attention particulière est portée à la **validité interne**, au sens de la consistance interne. Dans la section 7.5., **la validité externe** est traitée, partie la plus importante et la plus délicate du processus de validation. D'une manière générale, la validité externe se réfère à l'ensemble des preuves *indépendantes* en provenance d'autres méthodes conduisant essentiellement aux mêmes conclusions que les méthodes et procédures de l'étude en cours.

La validité n'obéit pas à une loi de type tout ou rien, mais s'établit plutôt sur un continuum. Pour un rapport sur la validité, il faudra être attentif aux nombreuses facettes impliquées, en mettant en avant de solides arguments et des preuves empiriques pour faire face aux critiques relatives à la généralisabilité. Il est ainsi indispensable, pour une bonne étude de la validation, de disposer d'une documentation conséquente sur l'ensemble des activités entreprises.

Ce chapitre conclura le Manuel et se terminera par quelques réflexions sur l'état de l'art relatif aux procédures de détermination des scores de césure. Il proposera également un bref regard orienté sur l'avenir.

7.2. Pré-requis : la qualité de l'examen

Relier au Cadre un examen qualitativement pauvre est une entreprise vouée à l'échec et qui ne peut être sauvée même par une détermination attentive des scores de césure. Dans cette section, un nombre important d'aspects de l'examen lui même seront discutés brièvement, en gardant un seul objectif en tête, celui de relier correctement l'examen au cadre. Ces aspects se réfèrent au contenu de l'examen, à ses aspects opérationnels et psychométriques.

7.2.1. Validité de contenu

D'une manière générale, le contenu d'un examen est dicté par des prescriptions curriculaires qui laissent peu de marges de liberté. Bien que les descripteurs de compétences en termes de savoir-faire du CECRL soient formulés de façon abstraite, il est possible d'entrevoir des zones de conflits entre les exigences curriculaires et la façon dont le CECRL est articulé. Il

se pourrait que certains items de l'examen soient si complexes qu'une correspondance univoque à l'un des niveaux du CECRL soit impossible ; toutefois, ne pas prendre en compte le caractère équivoque pourrait également introduire des conflits avec les exigences curriculaires.

Pour solder ce problème, considérons différents points :

- La position la plus extrême est de s'abstenir totalement de lien avec le CECRL. Bien que ça ne puisse probablement pas solder le problème à court terme, une publication à cet égard pourrait s'avérer utile pour une révision (ou une extension) du CECRL, ou pour une révision des exigences curriculaires pour les rendre plus compatibles avec le CECRL.

- Une approche plus nuancée pourrait être de rechercher un compromis et de relier l'examen au Cadre sur une seule partie de l'examen, en laissant de côté par exemple 25% des tâches et des items, parce qu'ils sont trop difficiles à apparier avec les catégories ou niveaux du CECRL.

- Une autre alternative serait de sélectionner une méthode de détermination des scores de césure moins analytique, pour laquelle aucune référence spécifique aux descripteurs du CECRL n'est nécessaire. Quelques méthodes de détermination des scores de césure reposent sur des jugements globaux, holistiques, (par exemple la méthode du corpus de productions, voir section 6.6.) alors que d'autres impliquent des jugements globaux sur la localisation du point de césure entre les niveaux d'un test, renseignés par une somme notable d'informations psychométriques (par exemple : la méthode du marque-page ou sa variante selon le Cito : voir section 6.8.-6.9.).

Un autre aspect de ce problème est de savoir dans quelle mesure les activités pertinentes et les compétences décrites dans le CECRL sont couvertes par l'examen. Les spécifications de l'examen (Chapitre 4) détaillent ce qui est inclus dans l'examen, mais pas ce qui a été laissé de côté. L'omission de parties et d'aspects importants du construit du CECRL peut conduire à un caractère unilatéral et engendrer des critiques quant à la généralisation d'un adossement injustifié de l'examen au Cadre. Il existe des méthodes pour quantifier la validité de contenu d'un examen et Kaftandjieva (2007) en a proposé un exemple pratique. Pour éviter tout danger d'une « sur-généralisation », il est préférable de mentionner explicitement le contenu couvert par l'examen (représentativité du contenu).

7.2.2. Aspects opérationnels : le test pilote

En amont de l'administration d'un examen en contexte réel, les données peuvent être collectées au cours de plusieurs étapes. D'une façon générale, on distingue la phase pilote et la phase de pré-test.

Le plus souvent, on entend par test pilote l'expérimentation du matériel de test de manière à éliminer les ambigüités, à vérifier la clarté et la compréhension des questions et de leurs consignes, à disposer d'une première estimation de la difficulté des tâches et des items et pour estimer la durée nécessaire à la passation. Un test pilote peut être conduit sur un petit échantillon (une ou deux classes suffisent généralement) ; il est cependant utile de ne pas présenter le matériel exclusivement comme un test, mais d'essayer de disposer d'un maximum de retour d'informations sur la qualité du matériel de test. Des méthodes qualitatives, comme les interviews et les « labos cognitifs[25] », peuvent révéler de nombreuses informations intéressantes ; les participants au test pilote peuvent être des élèves et des enseignants. Un bon pilotage permet d'éviter les mauvaises surprises lors de la phase de pré-test et de l'examen réel.

[25] Un "labo cognitif" est une procédure au cours de laquelle les participants sont invités à faire le test tout en réfléchissant à voix haute, en explicitant la façon dont ils comprennent les questions, leurs stratégies de réponses et les différentes étapes par lesquelles ils passent.

La dépendance entre les items est un aspect qui est facilement maitrisé dans la construction de tests à items. Un test fournit un maximum d'informations sur le construit qui doit être mesuré si chaque item est une nouvelle opportunité pour le candidat de montrer sa compétence et son niveau. Un item *i* qui peut être correctement traité seulement si un autre item *j* a été correctement traité, ou une mauvaise réponse à l'item *i* qui entraine une mauvaise réponse à l'item *j* sont des exemples caractéristiques de dépendance ; on parle alors de *dépendance fonctionnelle.* Cependant, des formes plus subtiles de dépendance peuvent apparaitre ; par exemple, traiter un item *i* peut fournir de l'information sur l'exactitude de la réponse à l'item *j,* même si l'information n'est pas complète. Plus encore, cette information peut être sélective de telle sorte qu'elle devienne facilitante si la réponse à l'item *i* est correcte. Ce type de dépendance est nommé la *dépendance statistique*. Ignorer la dépendance peut avoir des conséquences graves sur les caractéristiques psychométriques d'un test (par exemple conduire à une surestimation du coefficient de fidélité) ainsi que sur la détermination des scores de césure. Lors de projets ambitieux pour lesquels une banque d'items calibrés est construite permettant d'élaborer les examens par une sélection d'items issus de la banque, la dépendance peut avoir des conséquences fâcheuses. Si des items *i* et *j* ont été administrés de façon conjointe pour recueillir les données pour le calibrage de la banque et s'il y a une dépendance statistique entre eux, alors les paramètres psychométriques de l'un ou de l'autre, si l'un deux est utilisé de façon isolée dans l'examen deviennent imprévisibles.

Comme la démonstration de l'indépendance statistique n'est pas simple, il est préférable d'essayer, pendant la phase pilote, de détecter les stratégies subtiles auxquelles les candidats ont recours pour relier les items entre eux. Une collecte d'informations de la part des candidats, bien élaborée, pendant la phase de pilotage est une bonne façon d'identifier de tels problèmes[26].

7.2.3. Aspects opérationnels : le pré-test

Un pré-test est généralement conçu pour obtenir de l'information sur les principales caractéristiques de l'examen. En plus des paramètres psychométriques (qui seront discutés par la suite), les caractéristiques opérationnelles doivent aussi être observées. Le temps attribué et le temps nécessaire pour le pré-test est une source majeure d'information qui doit être collectée. Même si le nombre d'items non traités en fin de test par les candidats peut fournir une information utile, au moins deux aspects ne sont habituellement pas détectés :

- Les candidats qui manquent de temps pourraient être attirés par les items paraissant faciles. En particulier si l'examen est un mélange de questions à choix multiples et de questions à réponses construites, les candidats pourraient avoir tendance à traiter les QCM pour viser le score le plus haut possible. Dans une telle situation, une absence de réponse est difficile à interpréter : elle pourrait provenir de la difficulté intrinsèque de l'item ou d'une stratégie liée à la pression temporelle. Un court questionnaire administré aux candidats (ou à un échantillon d'entre eux) après le pré-test pourrait se révéler utile pour proposer une explication raisonnable en ce qui concerne les absences de réponses.

- Il est possible que le temps total alloué pour le test ait été surestimé, ce qui entraine une perte d'information. Pour mettre à jour simplement cette éventualité, il suffit de demander aux enseignants de noter pour chaque candidat le temps exact nécessaire pour faire l'examen dans son ensemble.

En dehors du fait d'être une répétition de l'examen à venir, le pré-test permet également la réalisation d'une fonction centrale, en l'occurrence celle de relier les examens entre eux. Etant donné que les examens tendent à être uniques du point de vue de leur composition d'une année sur l'autre et que les populations cibles n'ont pas d'élèves en commun[27], les

[26] Pour un traitement statistique et psychométrique de la dépendance par la TRI, voir Verhelst & Verstralen (2008).

données recueillies sur les deux examens ne peuvent être comparées ; les différences au niveau du score moyen pourraient être dues à des différences systématiques entre les deux groupes de candidats ou à une différence en termes de difficulté des contenus des deux examens ou encore par un mixte de ces deux raisons. Il n'y a aucune manière de savoir dans quelle mesure l'une et/ou l'autre de ces deux causes sont avérées, sauf si les données soient liées d'une certaine façon.

Parce que présenter des items aux mêmes candidats dans un pré-test que dans un examen a des conséquences imprévisibles en regard des effets mnésiques, les bonnes pratiques exigent que le pré-test soit conduit deux ans avant le test (ou sur une période de deux rotations d'examens). Si les examens des années 1 et 2 doivent être liés, le pré-test qui les lie devra alors être organisé deux ans avant l'examen 2, en l'occurrence en l'année 0.

Il est recommandé de planifier le pré-test selon un dispositif qu'on nomme «bloc incomplet équilibré». Les items des deux examens sont alors séparés en un nombre de sous-ensembles. Chaque candidat participant au pré-test se voit administrer le même nombre de sous-ensembles, mais aucun d'entre eux ne se voit administrer l'ensemble des items. Un dispositif de bloc incomplet équilibré présente ainsi les caractéristiques suivantes :

- chaque bloc est présenté à un nombre identique de candidats ;
- chaque paire de bloc est présentée à un nombre identique de candidats ;
- chaque bloc d'items est présenté dans chaque position, de façon sérielle.

Pour parvenir à cela, des restrictions doivent être introduites au niveau du nombre de blocs. Les dispositifs incomplets équilibrés sont possibles pour 2, 3, 7 et 13 blocs, mais pas pour d'autres nombres inférieurs à 13. Pour chacun des nombres ici mentionné, le nombre de formes différentes du test qui doivent être préparées équivaut au nombre de blocs. Le tableau 7.1. montre le dispositif qui doit être mis en œuvre pour trois blocs et le tableau 7.2. celui pour sept blocs. Dans le tableau 7.1., chaque candidat reçoit l'une des trois formes différentes de test. Les nombres en ligne indiquent pour une forme donnée du test le contenu de ce dernier mais également la séquence ordonnée de chacun des blocs. Il est aisé de vérifier que les trois exigences mentionnées ci-dessus en ce qui concerne un dispositif de bloc incomplet équilibré sont respectées ; c'est également le cas pour le dispositif en sept blocs.

Tableau 7.1: dispositif de blocs incomplets équilibrés avec trois blocs

Test	Blocs d'items	
1	1	2
2	2	3
3	3	1

[27] Même si un étudiant passe deux formes d'un examen (suite à un redoublement par exemple), on ne peut pas considérer que sa compétence est la même lors des deux examens, et dans toutes les analyses psychométriques un pareil étudiant sera analysé comme représentant deux individus (statistiques) distincts.

Tableau 7.2: dispositif de blocs incomplets équilibrés avec sept blocs [28]

Test	Blocs d'items		
1	1	2	4
2	2	3	5
3	3	4	6
4	4	5	7
5	5	6	1
6	6	7	2
7	7	1	3

On doit prendre garde à ne pas administrer la même forme de test à l'ensemble des étudiants d'une classe ou d'une école, parce que des différences systématiques entre les classes et les écoles pourraient biaiser les estimations des p_{obs} des items. En principe, toutes les formes de test devraient être administrées en nombre identique dans chaque classe. Pour mettre en œuvre ce principe on peut avoir recours au dispositif en spirale. Les différentes formes du test sont distribuées dans la classe en une séquence fixe : si le premier candidat reçoit la forme 4, le suivant recevra la forme 5, puis la 6, 7, 1, 2, 3 et enfin la séquence sera répétée. Il convient de commencer dans chaque classe par une séquence différente. La forme du test du début de la séquence devrait être choisie de façon aléatoire, ou devrait être d'un rang supérieur à celle avec laquelle la classe précédente s'est terminée. Toutes ces règles de planification sont un gage pour éviter les biais imprévus, difficiles à repérer.

Recourir à un dispositif de blocs incomplets équilibrés présente des avantages notables pour la construction de l'examen. Quel que soit le sous-ensemble d'items sélectionnés pour figurer dans l'examen de l'année 1, chaque item aura été observé en conjonction avec tous les autres items. Pour les items de l'examen de la deuxième année la même remarque s'applique, aussi bien d'ailleurs que pour les items non utilisés. Ainsi, chaque item de l'examen 1 est lié à chaque item de l'examen 2. Pour obtenir des contenus équilibrés dans chaque forme de test utilisée, il est primordial de rendre chaque bloc aussi hétérogène que possible, en regard du contenu et de la difficulté.

Examinons ce qu'il advient en année 1. L'examen de l'année 1 est administré, et au cours de cette même année un pré-test est nécessaire pour les deux années à venir. En appliquant le même principe que celui évoqué précédemment, pendant l'année 1, les items des années deux et trois doivent être pré-testés pour garantir le liage entre les examens des années 2 et 3. Ainsi, le matériel de l'année 2 doit de nouveau être pré-testé. Ici est illustré le principe de base : de manière à avoir un liage fort entre les examens d'une année sur l'autre, les items doivent être pré-testés deux fois.

Ensuite, il est important qu'un nombre suffisant de candidats fournisse des réponses pour chaque item. La théorie classique des tests est mal appropriée pour traiter les données recueillies dans le cadre d'un dispositif incomplet, de telle sorte qu'il faudra probablement avoir recours à la TRI. Or, pour utiliser de façon efficace la TRI des échantillons conséquents sont exigés ; un minimum de 200 réponses[29] doit être pris en compte pour obtenir des estimations relativement stables.

[28] Si l'on considère les trois colonnes, on peut remarquer qu'elles démarrent à une certaine valeur en haut de la colonne, jusqu'à 7 pour redémarrer à 1. Les premières valeurs sont respectivement 1, 2 et 4. Pour 13 blocs, on peut appliquer le même principe : les premières valeurs seront respectivement 1, 2, 4 et 10. Bien entendu, le cas échéant, le tableau présentera 13 lignes et on disposera de 4 blocs d'items. Pour 5 blocs d'items, il faudra 21 livrets différents, mais en pratique c'est rarement réalisable.

[29] Il est fortement recommandé de ne pas considérer ce nombre comme une règle absolue. Il constitue une indication de l'ordre de grandeur de la taille de l'échantillon. Dans les situations à forts enjeux, il faudra se faire conseiller par un psychométricien expérimenté pour qu'il puisse évaluer, probablement à l'aide de simulations computationnelles, la taille appropriée de l'échantillon.

7.2.4. Considérations psychométriques

Il est primordial que le pré-test fournisse suffisamment de données pour que suffisamment d'aspects psychométriques de l'examen puissent être fournis. Le premier aspect concerne les paramètres de l'item, comme la difficulté (valeur p) et le pouvoir discriminant. Si l'on s'en tient aux indices de la Théorie Classique des Tests, on doit considérer que ces indices sont dépendants de la population et que leurs valeurs sont simplement une indication des valeurs qu'ils ont au niveau de la population parente, à condition que l'échantillon du pré-test soit représentatif de la population cible. Organiser un pré-test uniquement dans un nombre restreint de centres pour des raisons de commodités (par exemple les centres où les enseignants sont membres de l'équipe d'élaboration du test) pourrait conduire à de sérieux biais au niveau des estimations.
La fidélité de l'examen est également un aspect important si l'on souhaite le relier correctement au CECRL. En effet, elle a un impact sur la précision et la consistance de la classification en termes de niveaux du CECRL, comme ce sera démontré ci-après. En estimant la fidélité, deux aspects doivent être gardés à l'esprit :

- Le KR20 (ou l'alpha de Cronbach) est souvent mentionné comme un indice de fidélité. En fait, il ne l'est pas exactement. Il permet faire une estimation par défaut de la fidélité. Ainsi avec des tests hétérogènes, il sous-estime substantiellement la fidélité. Le GLB (pour « greatest lower bound ») est un bien meilleur indicateur de la fidélité ; on peut en trouver une explication dans la section C du Supplément au Manuel.

- Si un dispositif de blocs incomplets a été utilisé pour le pré-test, le GLB sera uniquement disponible par livret de tests. Pour obtenir une estimation raisonnable de la fidélité sur l'ensemble de l'examen, cet indice devra être calculé uniquement pour les items qui seront sélectionnés pour l'examen. Sur ces estimations, la formule de Spearman-Brown peut être appliquée pour estimer la fidélité de l'examen dans son ensemble. Prendre en compte la moyenne de toutes ces estimations offrira une approximation raisonnable de la fidélité si une attention suffisante a été allouée à l'hétérogénéité et la représentativité des blocs par rapport à l'examen final.

7.2.5. Le bon moment pour déterminer les scores de césure

Si l'adossement au CECRL est lié à une situation à forts enjeux, le temps est généralement insuffisant pour collecter les données de l'administration de l'examen, remettre les résultats, organiser de façon complète une procédure de détermination des scores de césure et évaluer la validité de cette procédure.

Comme l'utilisation de données réelles de candidats est conseillée, y compris pour les méthodes de détermination des scores de césure centrées sur les tests (étude d'impact, retour d'informations réaliste ; voir chapitre 6), le laps de temps entre le pré-test et l'administration finale de l'examen sera probablement le plus adapté pour déterminer les scores de césure. Une planification sur une période de deux années, comme décrit précédemment, peut même offrir la possibilité, de façon croisée, de deux procédures de détermination des scores de césure. Cette éventualité sera envisagée en détail dans la section 7.4.

Dans cette section, la discussion sera focalisée sur les conséquences de ce qui est parfois nommé « l'effet pré-test ». Cette appellation fait référence à toutes les différences systématiques entre le pré-test et le véritable examen, différences qui pourraient moduler les performances des candidats. L'influence principale provient d'une différence en termes de motivation et de l'ensemble des facteurs directement liés à la motivation, comme le sérieux de la préparation et l'anxiété. S'il s'agit d'un examen à fort enjeu et d'un pré-test à faible enjeu, tous ces facteurs pourraient suivre la même tendance, en l'occurrence diminuer la performance dans le pré-test comparativement à la situation d'examen. Le cas échéant, la mesure d'impact présentée aux panélistes au cours de la procédure de détermination des scores de césure sera biaisée et pourrait avoir un effet systématique sur les scores de

césure proposés ; suite à cette information biaisée, si les panélistes se considèrent eux-mêmes trop stricts cela pourrait conduire à minimiser les scores de césure.

Nous suggérons ici quelques pistes qui pourraient permettre d'éviter -ou du moins contrôler- l'effet pré-test :

- Essayer d'organiser le pré-test, autant que faire se peut, dans des conditions semblables aux conditions de l'examen réel. Présenter un pré-test comme une sorte de répétition générale de l'examen, le plus proche possible dans le temps et avec des enjeux forts pourrait permettre d'accroitre la motivation et la préparation, de sorte qu'elles soient semblables au cours des deux sessions.

- Ajouter un court questionnaire à la suite du pré-test peut être un atout. Par exemple, les candidats qui montrent peu d'intérêt pour le pré-test ou qui prétendent qu'ils n'ont pas eu le temps ou l'opportunité pour se préparer sérieusement devraient être exclus des données à analyser.

- Si l'on parvient à conduire les pré-tests de la même manière depuis un longue période, les données des pré-tests et les données des examens réels devraient être comparées pour obtenir une estimation de l'effet pré-test. Si l'on obtient une estimation constante longitudinalement, l'effet pré-test pourrait être expliqué par les panélistes. Ainsi, une sorte de données d'impact corrigée devrait être présentée pendant les sessions de discussion. Par exemple, si l'effet du pré-test est estimé à deux points (la moyenne étant deux points au-dessus dans un examen réel que lors du pré-test), on pourrait ajouter cet effet à chaque score obtenu lors du pré-test pour calculer la proportion des candidats de chaque niveau en utilisant les scores de césure provisoires. Bien entendu, on se doit d'informer les panélistes de cette correction (et de sa justification) ; il n'y aurait rien à gagner en omettant l'information, au contraire cela pourrait avoir de graves conséquences.

7.3. Validité procédurale de la formation à la standardisation et à la détermination des scores de césure

Au cours du chapitre précédent, nous avons décrit plusieurs procédures pour familiariser les panélistes au CECRL, pour comprendre les spécifications d'un examen, pour déterminer des critères pertinents et pour définir les scores de césure. Les sessions de détermination des scores de césure exigent de débuter avec de telles explications et instructions ; les panélistes doivent se sentir en confiance pour réaliser leurs tâches. L'ensemble de ces procédures peut être considéré comme une étape, un pas supplémentaire vers les « bonnes pratiques » ; si on les ignorait, on se dirigerait vers des situations risquées. Le respect de telles procédures est une garantie ***nécessaire*** pour obtenir de bons résultats, en un mot : instructions correctes, résultats corrects.

Le problème de la validité est à mettre en relation avec le caractère nécessaire des procédures. Par exemple, en ce qui concerne la familiarisation (chapitre 3) et la formation à la standardisation (chapitre 5), s'il n'y a aucune phase préparatoire relative à la compréhension du CECRL, on ne peut pas espérer aboutir à un résultat valide. Par ailleurs, même si la procédure de formation suggérée est mise en œuvre, rien ne garanti que le résultat obtenu soit un succès ; la phase d'entraînement est nécessaire, mais est-elle suffisante ? La validation de cet aspect exige que la formation ait été efficace : si l'on forme les gens pour comprendre quelque chose, on doit aussi s'assurer qu'ils l'ont réellement compris à l'issue de la formation.

Plusieurs aspects relatifs à cette validité procédurale seront exposés ci-après. Il s'agit du caractère explicite, du caractère pratique, de la mise en œuvre, du retour d'information et de la documentation.
Le caractère explicite : il s'agit du degré selon lequel l'objectif de la procédure de détermination des scores de césure et la procédure elle-même sont clairement et

explicitement articulés. En d'autres termes, le processus est défini dans son intégralité avant qu'il soit conduit ; les étapes sont clairement décrites, les conditions de déroulement et les résultats attendus après chaque étape sont décrits comme un scénario immuable.
Si l'échéancier est suffisamment précis pour qu'il puisse constituer un guide pour une véritable réplication de l'intégralité de la procédure, on dispose d'un bon critère pour juger du caractère explicite. Une autre manière de vérifier si le caractère explicite est satisfait consiste à demander aux participants s'ils ont clairement compris l'objectif de la réunion permettant de déterminer les scores de césure et si les tâches relatives à cette détermination ont été clairement explicitées.

Le caractère pratique : même si certaines procédures sont compliquées, la préparation doit être pratique (voir Berk 1986), ainsi :

- La méthode de détermination des scores de césure doit pouvoir être mise en œuvre sans grande difficulté.
- L'analyse des données doit pouvoir être réalisée sans calculs laborieux. Cela ne signifie pas pour autant que les calculs ne sont pas compliqués, mais que le travail de préparation (comme par exemple la préparation de feuille de calculs Excel avec les formules appropriées) doit être accompli bien en amont de la session.
- Les procédures doivent être crédibles et interprétables par des non-techniciens.

Une manière de vérifier que le caractère pratique est satisfait consiste à demander aux panélistes si la formation a véritablement facilité la compréhension des tâches à accomplir.

La mise en œuvre : cet aspect fait référence à la manière, du point de vue de la rigueur, dont le panel est sélectionné et formé, à la manière dont les niveaux du CECRL sont intégrés et à celle dont les données de jugement sont effectivement traitées et analysées. Des informations relatives à ces points doivent être fournies.

Le retour d'information : cet aspect se réfère au niveau de confiance des panélistes à l'égard de la procédure de détermination des scores de césure et aux résultats qui y sont liés. Est-ce que les panélistes estiment qu'ils ont trouvé les bons résultats ? Des informations relatives à ces points doivent être collectées et rapportées.

Documentation : cet aspect se réfère à la manière dont la procédure de détermination des scores de césure est documentée, en particulier à l'égard des objectifs d'évaluation et de communication.

7.4. Validité interne de la détermination des scores de césure

Les questions relatives à la validité interne doivent permettre de se prononcer sur la *précision*, au sens de l'exactitude, et sur la *consistance* du résultat de la procédure de détermination des scores de césure. Un défaut de consistance peut provenir d'une faiblesse générale de la méthodologie mise en œuvre ou avoir une origine plus locale en reposant sur un ou deux juges ou quelques items. Le cas échéant, on pourrait : i) pour ce qui concerne les panélistes, supprimer certains d'entre eux (ou de l'analyse faisant suite à la procédure de détermination) ou ii) pour ce qui concerne les items, ne retenir qu'un sous-ensemble d'items et de tâches dans le test, en excluant ceux qui posent problème.

- En supprimant des panélistes, on doit prendre garde de ne pas influencer le résultat relatif aux scores de césure dans une direction souhaitée par l'organisateur. S'il l'on dispose de preuves quant à l'incompréhension des instructions à suivre par un panéliste, ou s'il les ignore volontairement, on dispose alors d'une raison valide pour le retirer des données à analyser. Des entretiens à l'issue de la session et un questionnaire bien conçu peuvent fournir les preuves recherchées. Une telle suppression doit être dûment documentée et le rapport final doit mentionner le nombre de panélistes retirés de l'analyse ainsi qu'expliquer les raisons du retrait.

- Supprimer des items ou des tâches est un problème bien plus délicat. Lorsque le premier souhait est d'adosser son examen au CECRL (par exemple en appliquant une règle qui

associe un échec à l'examen au fait ne pas avoir atteint le niveau B1/B2), retirer certains items pourrait sérieusement biaiser la validité de contenu du test. En outre, cela pourrait introduire sur le plan éthique des problèmes en ayant des candidats qui font des efforts inutiles pour se préparer à un examen. D'un autre côté, si le fait d'adosser son examen au CECRL est envisagé comme un aspect constituant de l'examen, on pourrait alors retirer les items problématiques de l'étude d'adossement tout en conservant ces mêmes items pour l'analyse permettant de remettre les résultats aux candidats.

La suite de cette section traitera des points relatifs à la consistance et à la précision :

- La consistance intra-juge consiste à rechercher les informations qui montrent qu'un juge est cohérent dans son jugement.
- La consistance inter-juges consiste à rechercher dans quelle mesure les panélistes s'accordent les uns avec les autres dans leurs jugements.
- La stabilité des résultats est exprimée par l'erreur standard des points de césure.
- La précision et la consistance de la classification reposent sur la procédure de détermination des scores de césure.

Parmi les méthodes proposées pour vérifier la consistance, toutes ne sont pas applicables à l'ensemble des méthodes de détermination des scores de césure discutées dans le chapitre précédent. Ainsi, nous utiliserons la méthode révisée de Tucker-Angoff pour illustrer le travail et nous ferons mention de commentaires supplémentaires pour ce qui concerne les autres méthodes quand cela s'avérera nécessaire.

7.4.1. Consistance intra-juge

A ce propos, deux questions sensibles peuvent être posées : est-ce que le juge (le panéliste) est consistant avec lui-même et est ce que sa réponse est consistante avec celle fournie pour les autres informations relatives au test ?

Pour répondre à la première question, il est nécessaire que le panéliste donne sa réponse à la même question à deux reprises (ou bien à deux questions très similaires). Cela pourrait être réalisé pendant une procédure de détermination des scores de césure avec un dispositif particulier de mesures répétées, dans lequel le tour final serait une nouvelle présentation (partielle) des items des premiers tours. Lorsque l'on travaille à déterminer plusieurs scores de césure (pour plusieurs niveaux) avec la méthode de Tucker-Angoff, on peut interroger chaque juge pour qu'il donne une seconde fois son estimation de probabilité pour l'un des scores de césure. Les probabilités estimées étant des nombres fractionnels, un diagramme de dispersion et un coefficient de corrélation peuvent offrir un éclairage sur la consistance interne des jugements. La corrélation peut être directement interprétée comme la fidélité des jugements. Les comparaisons de ces fidélités sur les juges peuvent fournir des informations utiles en ce qui concerne les juges atypiques, et ces indications pourraient conduire à exclure des analyses supplémentaires des données un ou deux panélistes.

En fournissant les probabilités pour une personne se situant à la limite des niveaux, les panélistes donnent implicitement une indication de la difficulté des items. En estimant que cette personne a une probabilité de 0.6 de donner une réponse correcte à l'item *i* et de 0.4 à l'item *j* cela signifie que le panéliste juge que l'item *i* est plus facile (des valeurs plus élevées correspondent à des items plus faciles). Ces probabilités estimées pourraient également être corrélées avec des indices empiriques de difficulté, comme les valeurs de p (on s'attend là à des corrélations positives) ou à des paramètres de difficulté de la TRI (on s'attend là à des corrélations négatives). Ce type d'indicateurs peut être considéré comme un coefficient de validité puisqu'il exprime la relation entre les jugements sur un ensemble d'items à l'aide d'un critère externe, les difficultés étant empiriquement déterminées à partir des réponses des candidats.
Utiliser des règles empiriques pour la corrélation est délicat et on se doit d'être vigilant avec de telles règles. Les valeurs des corrélations dépendront fortement de l'écart-type de la difficulté des items. Des faibles valeurs de variation conduiront à de faibles corrélations (effet

d'étendue). Mais comme pour la fidélité, comparer les corrélations sur les panélistes pourrait fournir des informations pertinentes concernant les valeurs atypiques.

Calculer ces indicateurs à la suite de l'ensemble de la procédure s'avèrera utile pour le rapport et la publication des objectifs, mais cela peut également être très utile au cours des sessions. Après chaque tour de jugement, ces corrélations, et les graphiques de dispersion qui y sont associés, peuvent être aisément produits pour pointer les incompréhensions ou les désaccords qu'on souhaite résoudre.

Des techniques similaires peuvent être utilisées avec d'autres méthodes de détermination des scores de césure. Nous discuterons maintenant de deux cas, la méthode du corpus de productions et la méthode du panier.

- Dans la méthode du corpus de productions, les candidats sont assignés à un niveau du CECRL sur la base d'un jugement holistique qui porte sur un dossier comprenant leur travail. On peut considérer ces niveaux du CECRL comme des modalités d'une variable ordinale, A2 étant supérieur à A1, B1 supérieur à A2, etc. Pour tous les candidats considérés, le score au test est connu et la corrélation entre le test au score et le niveau assigné peut être calculée ; les données (le score et le niveau assigné) peuvent aussi être représentées graphiquement dans un diagramme de dispersion. Pour le calcul de la corrélation, il est conseillé d'utiliser un coefficient de corrélation sur les rangs, le taux de Kendall[30], qui permet d'effectuer une correction.

- Dans la méthode du panier, la même approche peut être utilisée pour relier les niveaux assignés aux items et leur difficulté empirique.

Nous terminerons cette discussion par deux mises en garde :

- Dans la méthode du corpus de productions décrite au cours du chapitre précédent, les dossiers des candidats sont présentés par ordre croissant de score. Soit l'information de rang est communiquée aux panélistes, soit elle ne l'est pas, auquel cas ils trouveront rapidement qu'il y a un ordre. En présentant les dossiers de façon ordonnée, la consistance interne, dans une certaine mesure, est induite par la méthode elle-même : un panéliste réalisera très vite que plus le rang d'un dossier qu'il doit juger est élevé plus le niveau qui devrait être assigné sera élevé. Cela devrait induire une tendance (le panéliste n'osant pas affecter un niveau élevé à un dossier lui arrivant tôt ou un niveau faible à un dossier arrivant tardivement dans la séquence à évaluer). Cette tendance pourrait moduler, en partie, ce que le panéliste pense véritablement (et cela pourrait avoir des conséquences surprenantes au niveau des résultats de la procédure), et en même temps conduire à une augmentation des corrélations mentionnées précédemment.

- Certaines méthodes fournissent tellement d'information aux panélistes qu'il est virtuellement impossible de montrer un comportement inconsistant. Des exemples typiques sont trouvés à travers la méthode du marque-page et sa variation proposée par le Cito, où pour chaque point de césure un unique jugement holistique doit être fourni. Dans la méthode du marque-page, il est même impossible, de par la manière dont est définie la procédure, de générer un score de césure A2/B1 inférieur à celui pour A1/A2. Pour autant, cela ne signifie pas que la consistance intra-juge n'est pas importante dans ces procédures. Dans la variation de la méthode du marque-page proposée par le Cito, la tâche opérationnelle des panélistes est si simple (dessiner une ligne ou noter un nombre, voir section 6.9.) qu'un score de césure proposé arbitrairement par un panéliste désintéressé pourrait passer inaperçu. Aussi, il est conseillé, dans cette procédure, de vérifier la consistance intra-juge par une tâche supplémentaire qui pourrait se dérouler comme il est indiqué ci-dessous. Une fois que le score de césure est déterminé, on peut dériver pour chaque item la valeur normée correspondante à « pas de maîtrise », « à la limite de la maîtrise » ou à « maîtrise totale ». On aboutit ainsi à une classification des

[30] Voir, par exemple, Siegel & Castellan (1988).

items en trois classes. Dans une tâche indépendante, les panélistes pourraient être interrogés pour classer tous les items dans l'une de ces trois catégories sans information psychométrique disponible (voir figure 6.5.). Ces deux classifications, l'une dérivée des scores de césure provisoires et l'une collectée par l'assignation « en aveugle », peuvent être représentées (par panéliste) dans une table de fréquence 3 x 3, et un indicateur d'accord peut être calculé.

7.4.2. Consistance inter-juges

Pour évaluer la consistance inter-juges, on peut essayer de déterminer dans quelle mesure les panélistes s'accordent les uns avec les autres, ou dit autrement, dans quelle mesure ils donnent des jugements similaires. C'est ce dernier point qui est généralement appelé la consistance. Il est important de faire une distinction claire entre ces deux concepts. Nous proposons un petit exemple pour expliquer les différences.

7.4.2.1. Accord et consistance

Supposons que 30 items doivent être assignés à l'un des niveaux du CECRL, comme dans la méthode du panier, et que les jugements de deux panélistes sont résumés dans un tableau de fréquences à deux dimensions (voir figure 7.3.). On peut y voir que le panéliste 1 a assigné 7 items au niveau A1 alors que le panéliste 2 a étiqueté ces 7 mêmes items au niveau A2. Ainsi, pour ces 7 items, les deux panélistes sont en complet désaccord quant au niveau des items. La même chose se produit pour les autres items, comme on peut le voir aisément dans le tableau, parce que toutes les fréquences sur la diagonale principale (cf. les nombres soulignés) sont égales à zéro. Mais en dépit de ce désaccord total, on ne peut pas dire qu'il n'y a pas de similarités systématiques entre les décisions des deux panélistes : le panéliste 2 place tous les items à un niveau au-dessus du panéliste 1, ce qui signifie que le panéliste 2 est plus indulgent dans son évaluation que le panéliste 1.

Tableau 7.3: exemple de consistance forte et de désaccord complet

		Panéliste 2				
		A1	**A2**	**B1**	**B2**	**Total**
Panéliste 1	**A1**	0	7	0	0	7
	A2	0	0	11	0	11
	B1	0	0	0	12	12
	B2	0	0	0	0	0
	Total	0	7	11	12	30

Parce que les quatre niveaux du CECRL sont clairement ordonnés, on peut calculer un coefficient de corrélation sur les rangs entre les évaluations des deux panélistes. Le taux de Kendall dans ce cas est égal à 1, relatant la consistance totale entre les deux panélistes. En général, nous pouvons dire que les mesures de consistance, usuellement exprimées par un coefficient de corrélation, ne sont pas sensibles aux décalages systématiques des évaluations qui peuvent être rapportés au caractère indulgent ou sévère des jugements. Ainsi, il est utile d'être vigilant à la fois au degré d'accord mais aussi à la consistance lorsqu'on évalue le travail des panélistes[31].

[31] Une analyse multi facettes (TRI) des données relatives aux jugements à l'aide du programme FACETS est une façon pour y parvenir.

7.4.2.2. Trois mesures d'accord

Pour illustrer ces mesures, nous utilisons un résultat plus réaliste que les données artificielles du tableau 7.3. Supposons que 50 items doivent être assignés à quatre niveaux et que pour deux panélistes on dispose des fréquences représentées dans le tableau 7.4.

Tableau 7.4: tableau de fréquence pour quatre niveaux et deux panélistes

		Panéliste 2				
		A1	**A2**	**B1**	**B2**	**Total**
Panéliste 1	**A1**	7	2	1	1	11
	A2	1	10	2	1	14
	B1	1	2	12	2	17
	B2	0	1	0	7	8
	Total	9	15	15	11	50

L'indice d'accord exact est la proportion des cas (ou d'items) où les deux panélistes donnent exactement le même jugement. Les fréquences d'accord exact sont données par les cellules de la diagonale principale (en gris foncé) du tableau. Ainsi, dans cet exemple :

$$p_{exact} = \frac{7+10+12+7}{50} = \frac{36}{50} = 0.72$$

Cette valeur n'est pas particulièrement élevée ici. Bien entendu, pour ce qui concerne les items pour lesquels les deux panélistes sont en désaccord, le désaccord pourrait varier en degré : un résultat où un item est déplacé de trois niveaux est plus inquiétant qu'une situation où les niveaux donnés aux items par les panélistes sont adjacents. Ces dernier cas sont représentés dans le tableau 7.4., par les cellules en gris clair. Au total, il y a 2+2+2+1+2 = 9 items pour lequel c'est le cas. L'indice d'accord adjacent est la proportion des items conduisant à un accord parfait ou à une différence d'un niveau. Dans l'exemple en cours, on trouve :

$$p_{adj} = \frac{36+9}{50} = \frac{45}{50} = 0.90\,.$$

Même si les deux panélistes donnaient leurs jugements au hasard, les indicateurs d'accord ne seraient pas égaux à zéro. Ils prendraient une valeur positive dont l'amplitude dépendra des fréquences marginales (la ligne du bas et la colonne la plus à droite du tableau 7.4.). Le nombre attendu dans chaque cellule, selon l'hypothèse de réponses aléatoires mais avec des marges fixes, est donné par le produit des lignes multiplié par les colonnes, le tout divisé par le total. Pour la cellule (A1, A1) du tableau 7.4., on a 11 x 9 / 50 = 1.98. Pour les trois autres cellules de la diagonale principale les fréquences attendues sont 4.20, 5.10 and 1.76, et la somme des fréquences attendues pour l'ensemble des cellules de la diagonale principale est 13.04. Ainsi, si les panélistes répondent aléatoirement, on s'attend à un indice d'accord exact égal à :

$$E(p_{exact}) = \frac{13.04}{50} = 0.26\,.$$

Le coefficient kappa de Cohen est un indice d'accord bien plus utilisé qui prend en compte l'accord obtenu par chance. Il est défini (pour l'accord parfait) par :

$$\kappa = \frac{p_{exact} - E(p_{exact})}{1 - E(p_{exact})}\,.$$

Au numérateur de cette formule, la proportion empirique d'accord trouvée est comparée à ce qui devrait être attendu sous des conditions de réponses aléatoires. La fonction du dénominateur est de maintenir la valeur maximale du kappa à 1. Notez que le kappa peut

être négatif dans le cas où l'accord trouvé est plus faible que ce qui pourrait être attendu sous des conditions de réponses aléatoires.

7.4.2.3. Evaluation des indices d'accord

Comme c'est le cas pour de nombreux indicateurs psychométriques, il est difficile d'évaluer les résultats d'une étude de façon absolue, c'est avant tout peu réalisable et surtout potentiellement risqué.

- Considérons l'indice d'accord absolu. Si les items qui doivent être évalués constituent un sous-ensemble pratiquement homogène, pour des exemples relatifs aux niveaux A2+ et B1, un indice d'accord moyen de 0.8 pourrait être exceptionnellement élevé. D'un autre côté, pour une situation très hétérogène du point de vue de la collection des items couvrant une large étendue de niveau, la même valeur d'indice pourrait être insatisfaisante, indiquant même une attitude peu sérieuse d'un ou plusieurs panélistes.

- Il faut accorder une attention particulière au dispositif de l'étude permettant la détermination des scores de césure et garder à l'esprit que la méthode utilisée peut induire des valeurs élevées ou faibles en ce qui concerne l'accord entre panélistes. La méthode dite du corpus de productions offre un bel exemple. Dans cette méthode, les candidats sont assignés à un niveau, mais le matériel sélectionné doit être très hétérogène et l'étendue des scores couverts doit être totale. C'est cette hétérogénéité qui facilitera un haut niveau d'accord. Si l'on travaille avec un critère absolu (disons 0.8) pour l'indice moyen d'accord, atteindre cette valeur pourrait créer un sentiment de satisfaction. Néanmoins, il se pourrait que cet indice en apparence élevé masque en fait l'incompréhension des instructions, pour un ou deux panélistes, qui auraient influencé le score de césure définitif dans une direction non souhaitée.

Une approche plus intéressante consiste à adopter un point de vue relatif. Les indices discutés ci-dessus sont définis pour des paires de panélistes. Avec 12 panélistes, cela signifie qu'il y a (12x11)/2=66 paires et un ou plusieurs indices qui peuvent être calculés pour chaque paire. Bien entendu, ces indices montreront une certaine variabilité entre eux, et la question qui reste à solder est de savoir si l'on peut étudier cette variabilité pour améliorer les résultats (dans un tour suivant au cours de discussions centrées sur les zones à problèmes) ou identifier et retirer quelques panélistes dont la performance est mauvaise ou encore des items de manière à améliorer la qualité globale de la détermination des scores de césure.

Bien qu'il y ait certaines méthodes pour généraliser les indices comme le kappa de Cohen à plus de deux panélistes, de tels résumés pourraient masquer des réponses isolées et sont rarement utiles pour se centrer sur les points faibles d'une étude comprenant plusieurs évaluateurs. Ici nous esquisserons une manière simple pour évaluer les forces et les faiblesses de l'accord inter-juges. Nous utiliserons le kappa de Cohen comme exemple, mais la même procédure peut être appliquée avec l'indice d'accord exact ou d'accord adjacent.

- Il convient de disposer les indices dans une matrice. La valeur dans la cellule (*i,j*) est le coefficient kappa calculé pour les panélistes *i* et *j*. Le tableau est symétrique et les valeurs de la diagonale principale sont laissées indéfinies. Elles n'entrent dans aucun calcul par la suite.

- On peut maintenant extraire une information pertinente en calculant deux indices pour chaque colonne du tableau :

    - La moyenne de chaque colonne offre un indicateur pour chaque juge exprimant le niveau d'accord général avec l'ensemble des autres juges. Un graphique représentant ces valeurs moyennes des colonnes indiquera immédiatement les panélistes qui sont le plus en désaccord avec les autres, puisqu'ils auront les plus faibles valeurs

- L'écart-type de chaque colonne. L'évaluation conjointe de la moyenne et de l'écart-type offre une information supplémentaire. Si la moyenne est faible et que l'écart-type est petit, cela signifie que le panéliste est en désaccord avec les autres et qu'il le fait de façon systématique. Cela peut se produire dans une situation où le panéliste a systématiquement une idée déviante du CECRL ou de la signification des items. Au contraire, un écart-type élevé révèle un comportement erratique. Un graphique de dispersion des moyennes et des écarts-type pourrait aider à diagnostiquer les problèmes d'un ou plusieurs panélistes.

La technique expliquée ci-dessus est utile dans les cas où seulement quelques panélistes montrent un comportement déviant par rapport à la majorité des autres panélistes. Pour les situations où par exemple les panélistes sont en deux sous-groupes, que chaque panéliste est en fort accord avec l'ensemble des panélistes du sous-groupe auquel il appartient et en fort désaccord avec les membres de l'autre sous-groupe, cette technique fait défaut. Dans une telle situation, il est conseillé de recourir à des techniques qui peuvent révéler une structure complexe dans la matrice des accords. Une analyse par groupe et une approche multidimensionnelle pourraient être appropriées.

7.4.2.4. Repérer les items problématiques

Dans les procédures de détermination des scores de césure où les panélistes attribuent un niveau aux items ou à des tâches (comme dans la méthode du panier ou celle de l'appariement au descripteur), il y a deux façons simples de repérer si un défaut d'accord peut être attribué à quelques items.

La première est de construire un tableau ou une représentation graphique par item (un histogramme) qui indique les fréquences (absolues ou relatives) pour chaque niveau. Un exemple d'item problématique est proposé dans le tableau 7.5[32]. Dans la figure 7.1. la courbe des caractéristiques empiriques de l'item est représentée. Les candidats ont été classés par niveau (représenté sur l'axe horizontal) en utilisant les points de césure tels que définis par le panel d'experts. Pour chaque groupe, le pourcentage de réponses correctes à cet item est représenté.

Tableau 7.5: fréquence d'attribution des niveaux du CECRL pour un item

Niveau	**A1**	**A2**	**B1**	**B2**	**C1**	**C2**
Fréquence	0	17	11	5	0	1

Deux propriétés importantes de l'item peuvent être déduites de la figure : (a) il s'agit d'un item particulièrement difficile, que les candidats de niveau A ne peuvent résoudre, et (b) la proportion de réponse correcte est inférieure à 0.6., pour les candidats de niveau C. De plus, la courbe croit très rapidement, ce qui indique un fort pouvoir discriminant de l'item. En combinant ces informations aux jugements des panélistes, une question apparait : comment peut-on expliquer le fait qu'une majorité de panélistes attribue le niveau A2 à cet item ? De plus, on peut voir qu'un seul panéliste localise cet item au niveau C tandis qu'une analyse simple de la figure 7.1., semble démontrer qu'il ou elle a en fait raison ! Ceci nous enseigne qu'appliquer une simple règle de majorité et supprimer les désaccords par un consensus n'est pas toujours une bonne décision. Il est clair que le tableau 7.5., et la figure 7.1., seraient des informations de valeur à prendre en considération pour un futur tour de discussion.

[32] Il s'agit d'un véritable exemple issu d'un séminaire récent sur la détermination des scores de césure.

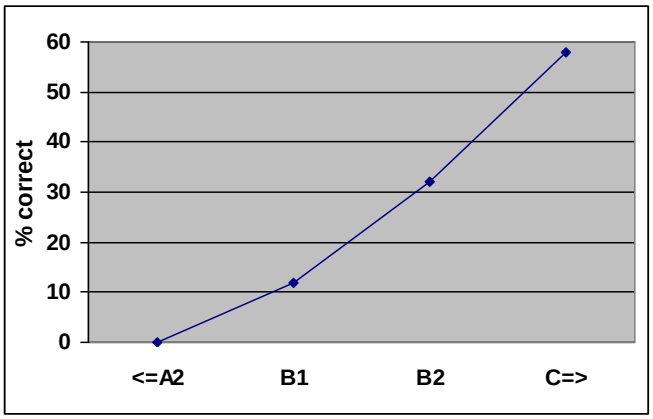


Figure 7.1: courbe caractéristique empirique de l'item pour un item problématique

Une seconde méthode pour proposer une vue d'ensemble des items problématiques est d'utiliser l'information des tableaux de fréquences comme le montre le tableau 7.4. Dans ce tableau on trouve cinq items pour lesquels les niveaux assignés par les deux panélistes sont différents d'au moins deux niveaux. Si l'on identifie ces items, et qu'on le reproduit pour chaque paire de panélistes, on peut construire une table de fréquence comme celle montrée dans le tableau 7.6.

Les lignes correspondent aux items et les valeurs des cellules correspondent au nombre de fois où l'item a été assigné à des niveaux différents. La valeur 3 à l'intersection de la première ligne (premier item) et de la première colonne indique que 3 paires de panélistes ont positionné cet item à deux niveaux d'écart. Les items présentant les fréquences les plus élevées dans la colonne la plus à droite sont probablement les items les plus problématiques et ceux qui méritent le plus d'attention au cours des discussions. Le tableau ci-dessous indique clairement que l'item 3 est celui qui mérite l'attention la plus soutenue.

Tableau 7.6: résumé des désaccords par item

Item ID	Deux niveaux d'écart	Trois niveaux d'écart
1	3	1
2	2	0
3	3	7
4	0	0
5	2	0
⋮	⋮	⋮

7.4.2.5. Indices de consistance

Trois méthodes différentes pour évaluer la consistance ou le manque de consistance au niveau des correcteurs seront discutées : la corrélation intra-classe, une méthode qui est une application directe de la Théorie Classique des Tests, et, très brièvement, une mesure de consistance appropriée aux jugements sur une échelle ordinale.

La corrélation intra-classe : considérons la méthode révisée de Tucker-Angoff. Les résultats principaux de cette procédure sont appelés les taux d'Angoff, en l'occurrence les déclarations de probabilité d'une réponse correcte pour une personne à la limite des niveaux. Ces données peuvent être disposées dans une matrice dont les lignes indiquent les items et les colonnes les panélistes.

Dans la situation idéale, où tous les juges seraient en accord parfait, toutes les colonnes de ce tableau seraient identiques. Cela signifie que toute variation entre les nombres de ce tableau peut être attribuée aux items. Si des variations sont dues aux juges, il s'agit d'une entorse à la situation idéale, qui précisément est nommée inconsistance. Une façon d'exprimer le manque de consistance est de considérer la proportion de variance due à la

variance liée aux items. Cette proportion est appelée la corrélation intra-classe et varie entre zéro et un, un correspondant à la situation idéale. Voilà comment calculer cette corrélation intra-classe :

- Calculer la variance de l'ensemble des nombres du tableau. Cette dernière est appelée la *variance totale*.
- Calculer pour chaque ligne du tableau la valeur moyenne. Puis calculer la variance de ces valeurs moyennes. Cette variance est celle liée aux items.
- Le rapport entre ces deux variances correspond à la corrélation intra-classe, symbolisée par ρ_{ic}

La différence 1 - ρ_{ic} est la proportion de variance qui n'est pas due aux différences entre les items. Cette variance serait due aux différences systématiques entre juges ou aux interactions entre items et juges et à un bruit de fond. Pour distinguer les sources de variation on peut facilement calculer la variance sur les juges (en colonne), en calculant la moyenne pour chaque colonne, puis en calculant la variance sur ces valeurs moyennes.

Tableau 7.7: résultat d'une procédure de Tucker-Angoff

Items/juges	**1**	**2**	**3**	Moyenne
1	38	32	24	31.3
2	27	31	38	32.0
3	42	33	50	41.7
4	51	49	47	49.0
5	52	60	62	58.0
6	63	58	71	64.0
7	71	68	75	71.3
8	82	77	92	83.7
Moyenne	53.3	51.0	57.4	

Dans le tableau 7.7., un exemple factice est donné pour huit items et trois juges. Les nombres dans ce tableau représentent le nombre sur 100 de personnes à limite des niveaux, qui selon les juges répondraient correctement à chacun des items. La colonne la plus à droite contient la moyenne des lignes et la ligne inférieure la moyenne des colonnes.

Dans le tableau 7.8., la décomposition de la variance totale en trois composantes est représentée. La variance résiduelle (interaction ou erreur) est obtenue en soustrayant les composantes items et juges à la variance totale.

Tableau 7.8: décomposition de la variance

Source	
Items	308.91
Juges	6.97
Résiduelle	17.89
Total	333.78

De ce tableau, nous apprenons que :

- La corrélation intra-classe est de 308.91/333.78 = 0.926, ce qui signifie que seulement 7.5% de la variance totale est due aux manières différentes des juges de traiter les items.
- La variance imputable aux différences systématiques entre juges est de 6.97, ce qui représente 2.1% de la variance totale.
- La proportion restante (5.4%) est véritablement ce que l'ont pourrait appeler l'inconsistance.
- Dans cet exemple factice, la corrélation intra-classe est très élevée, mais ce n'est pas nécessairement attribuable à la qualité des juges ou au processus de détermination des scores de césure de façon absolue. Les items (les moyennes en ligne dans le tableau 7.7.) indiquent une forte source de variation, et ce qu'indique véritablement le tableau 7.8. est que l'inconsistance des juges est relativement faible comparée à celle au niveau des items.

La décomposition de la variance totale peut être aisément réalisée (par exemple dans une feuille de calcul Excel). Elle est utile pour guider les discussions suivantes mais aussi pour le rapport sur la validité interne de la détermination des scores de césure.

Utilisation de la Théorie Classique des Tests : La Théorie Classique des Tests offre un indice de consistance avec l'alpha de Cronbach. Pour mettre en œuvre cette procédure, nous utilisons les taux d'Angoff du tableau 7.7., où les items (en ligne) vont prendre le rôle des candidats et les juges le rôle des items. Ainsi, pour le tableau 7.7., cela signifierait que l'on dispose de 8 étudiants et trois items. La valeur de l'alpha dans cet exemple est égale à 0.97.

Notez que la valeur de l'alpha ne varie pas si l'unité de mesure est changée. Concrètement, le résultat restera le même si les données du tableau 7.7., expriment des pourcentages ou des proportions[33]. Plus de détails sur l'alpha de Cronbach sont proposés dans la section C du Supplément au Manuel.

Utiliser la Théorie Classique des Test offre également un avantage supplémentaire. La corrélation item-total, dans ce contexte, fournit une indication de la façon dont chaque juge (qui a pris le rôle des items) s'accorde avec la moyenne. Ainsi, l'on dispose d'une belle façon de détection des panélistes atypiques. Dans l'exemple du tableau 7.7., les trois corrélations valent 0.98.

Mesures ordinales : les méthodes discutées au cours des sections précédentes sont applicables dès lors que les observations sont transposables dans un tableau à deux entrées, principalement du type items/juges pour les méthodes de détermination des scores de césure centrées sur le test ou candidats/juges pour les méthodes centrées sur le candidat comme avec la méthode du corpus de productions. On peut toutefois rencontrer un problème quand on doit décider de ce que l'on doit reporter dans le tableau à deux entrées et sur la façon dont on doit interpréter les valeurs du tableau.

Prenons pour exemple la méthode de l'appariement au descripteur. L'évaluation de base fournie par les panélistes consiste en un niveau du CECRL, pouvant s'étaler de A1 à C2. On peut compléter ces niveaux dans le tableau (comme étiquettes), le cas échéant on ne peut plus alors appliquer les méthodes décrites précédemment puisqu'elles requièrent un tableau avec des valeurs numériques. Ce qu'il est possible de faire dans une telle situation est de remplacer les étiquettes A1 à C2 respectivement par les chiffres de 1 à 6, puis ensuite procéder comme il est décrit ci-dessus. Dans la littérature, des alternatives sont suggérées pour une telle procédure et certains auteurs pensent que ce n'est pas possible puisque les chiffres utilisés pour compléter le tableau (1 à 6) ne relèvent pas d'une échelle d'intervalle. Il s'agit d'un argument fort, mais on ne doit pas alors recourir aux techniques de décomposition de la variance ou à celle de la Théorie Classique des Tests. Si toutefois, on les applique, cela pourrait fournir des informations utiles, même si l'interprétation reste acrobatique. On peut alors avoir recours à des indices de consistance qui reposent totalement sur les caractéristiques ordinales des données. Le coefficient de concordance W de Kendall constitue alors un bon indicateur[34], [35].

[33] Bien entendu, sous la condition de cohérence sur l'ensemble du tableau : utiliser des pourcentages pour une moitié des colonnes et des proportions pour l'autre moitié conduirait à des résultats étranges et serait totalement inutile.

[34] Pour une bonne introduction, voir Siegel and Castellan (1988).

[35] Il existe également des techniques valables pour pratiquer des analyses quantitatives sur des tableaux contenant des données nominales, où les modalités A1 à C2 sont considérées simplement comme des étiquettes. Ces techniques sont connues sous différents noms, comme l'analyse d'homogénéité ou l'analyse des correspondances multiples. Une référence pratique peut être consultée dans OECD (2005), Chapitre 10.

7.4.3. Exactitude et consistance de la méthode de détermination des scores de césure

Quelle que soit la façon dont on procède au cours de la phase de familiarisation et pendant les tours de discussion, si l'on insiste sur le fait que les panélistes peuvent librement donner leurs jugements, en toute indépendance et sans crainte d'une quelconque sanction, il est inévitable d'avoir des variations dans les jugements. Il ne s'agit pas nécessairement d'un mauvais point, parce que les panélistes sont conviés avec leurs compétences individuelles mais sont priés de parvenir à une décision de groupe raisonnable. En outre, si le processus de sélection des panélistes a été conduit avec une attention soutenue, de telle sorte que les panélistes sont représentatifs de leurs pairs, cela signifie qu'avec un autre échantillon de même taille on devrait observer des résultats similaires à ceux observés avec l'échantillon sélectionné.

7.4.3.1. Erreur standard du score de césure

Que seraient les scores de césure si l'on impliquait la population totale des juges considérés comme des experts en la matière, en fait la population parente ? Si l'on prenait le jugement moyen (du score de césure) des panélistes de l'échantillon, on obtiendrait une estimation de cette population totale, et l'erreur standard (SE_S) de cette estimation est donnée par l'écart-type (SD_S) des scores de césure individuels divisé par la racine carrée du nombre de panélistes n :

$$SE_S = \frac{SD_S}{\sqrt{n}}$$

Dans la littérature, cette erreur standard est généralement comparée à l'erreur standard de mesure du test et il est généralement admis que cette erreur standard ne doit pas être supérieure à l'erreur standard de mesure. Certains auteurs sont cependant plus stricts. Cohen et al (1999) exigent que l'erreur standard soit au moins inférieure à la moitié de l'erreur de mesure, alors que Jaeger (1991) considère qu'elle doit être d'au moins un quart de la valeur de l'erreur de mesure. Norcini et al (1981) suggèrent que l'erreur standard des points de césure ne devrait pas être de plus de deux items sur cent. Ceci signifie que pour un test de 50 items, l'erreur standard du score de césure devrait être au plus d'un.

Le standard 2.14 de AERA/APA/NCME (1999) stipule :

> «Que l'on devrait reporter les erreurs de mesure autour du voisinage de chaque score de césure, et ce qu'ils soient spécifiés pour la sélection ou la classification »

Les applications simples de la Théorie Classique des Tests reportent une valeur unique pour l'erreur standard de mesure, ce qui implique que les scores (en tant qu'indicateurs du score vrai) sont identiquement précis indépendamment de la valeur du score vrai. Néanmoins, par la mise en œuvre de la TRI, nous savons parfaitement que l'erreur standard de l'estimation de l'habileté dépend de la valeur de la variable elle-même (voir le concept d'information du test dans l'annexe G du Supplément au Manuel).

Dans le cadre de la Théorie Classique des Tests, il y a eu des tentatives pour parvenir à différentes valeurs de l'erreur standard de mesure en fonction du niveau de score (Feldt et al 1985). Une formule adéquate pour exprimer l'erreur standard à différents niveaux de score pour des tests constitués d'items binaires est proposée par Keats (1957):

$$SEM(X) = \sqrt{\frac{X(k-X)}{k-1} \times \frac{1-\rho_{xx'}}{1-KR_{21}}}$$

Dans cette formule :

- X représente le score;
- k représente le nombre d'items ;
- $\rho_{xx'}$ est la fidélité du test;
- KR_{21} est l'une des formules de Kuder-Richardson, qui exprime la fidélité d'un test homogène, pour des items de difficulté (pratiquement) identique. La formule du KR_{21} est :

$$KR_{21} = \frac{k}{k-1}\left[1 - \frac{k\bar{p}\bar{q}}{SD_X^2}\right]$$

où $\bar{p}$ est la moyenne des valeurs p et $\bar{q} = 1 - \bar{p}$.

Notez que la *SEM(X)* donne un résultat différent, dépendant ou conditionnel du score *X*. Ainsi, elle est souvent appelée l'erreur standard de mesure conditionnelle. Ses valeurs sont grandes pour des scores à proximité du milieu de l'étendue des scores et diminuent au fur et à mesure que le score décroit ou augmente. Ainsi, si l'on choisit un critère pour juger l'erreur standard du score de césure (par exemple exiger qu'elle soit inférieure à la moitié de l'erreur standard de mesure) cela conduira à une exigence impliquant que plus l'erreur de mesure est petite plus le score de césure sera éloigné du milieu de l'étendue des scores.

7.4.3.2. Une situation paradoxale

Il est admis que dans les applications de la TRI, on obtient les estimations les plus précises de l'habileté latente des candidats pour ceux ayant environ la moitié des items corrects, c'est à dire pour un score aux alentours de la moitié entre le score le plus faible possible et le score le plus élevé possible, alors que les résultats présentés sur l'erreur standard de mesure conditionnelle indiquent le contraire. Pour comprendre cette apparente contradiction, on doit prendre en compte que l'étendue des scores d'un test est délimitée de bas en haut, par le score minimal qui est généralement de zéro. Avec 50 items, où chacun vaut un point, le score maximum est de 50. Dans la TRI, le concept de base n'est pas le score au test mais une variable latente abstraite non bornée, qui peut varier de moins l'infini à plus l'infini. Une façon adéquate d'exprimer la relation entre la variable latente et le score est de représenter la fonction caractéristique du test[36]. Dans la figure 7.2., une courbe caractéristique d'un test de 50 items est proposée. Bien que la courbe présente une allure générale en demi-cloche, elle n'est pas très régulière ; les irrégularités sont dues aux combinaisons particulières des paramètres de discrimination et de difficulté des items[37].

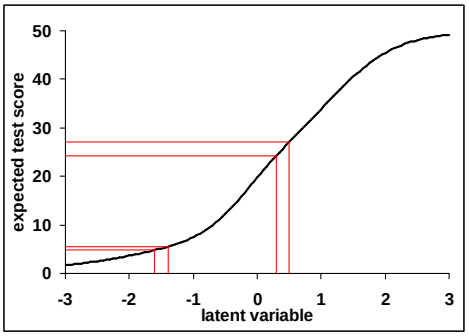


Figure 7.2: une courbe caractéristique de test

Sur l'axe horizontal deux intervalles sont représentés, chacun ayant une largeur de 0.2. Celui de gauche va de -1.6 à -1.4 et en correspondance les scores attendus au test vont de 4.82 à 5.54 (une étendue de 0.72 points). Le deuxième intervalle, qui a la même étendue sur l'axe horizontal (de 0.3 à 0.5) conduit à une correspondance du point de vue du score au test qui s'étage de 24.26 à 27 points (une étendue de 2.74 points, soit environ quatre fois l'étendue du premier intervalle).

Si une méthode pour déterminer les scores de césure a été utilisée en localisant le point de césure sur l'échelle latente, comme dans la méthode du marque-page ou dans la variation de cette méthode proposée par le Cito, l'erreur standard est exprimée dans l'unité de cette échelle. Mais pour la plupart des utilisateurs un point de césure exprimé selon les scores (au

[36] Des détails supplémentaires au sujet de cette fonction peuvent être consultés dans la section 6.8.3.
[37] Il est conseillé, lorsqu'on a recours à la TRI, de construire la courbe caractéristique du test : elle permet de rendre explicite la relation entre un concept abstrait (la variable latente) et des faits observables (les scores au test). Les paramètres de la courbe de la figure 7.2., ont été choisis pour mettre l'accent sur cette irrégularité.

test) est nécessaire, par conséquent une estimation de l'erreur standard sur l'échelle du score doit également être fournie. C'est pourquoi le recours à la courbe caractéristique du test peut s'avérer utile[38].

7.4.3.3. Exactitude et consistance des décisions

Déterminer les scores de césure implique une décision reposant sur les performances individuelles. Si le score de césure pour A2/B1 est fixé à 23/24 cela implique la décision que tout candidat obtenant un score inférieur à 24 à l'examen ne sera pas placé au niveau B1. De cette manière, on exprime l'intention d'affecter un certain niveau à un candidat s'il le mérite vraiment. Mais certaines décisions pourraient être erronées et il serait alors utile de distinguer les sources d'erreur. Nous proposons dans ce qui suit un exemple concret pour comprendre.

Supposons que le candidat Jean obtient un score de 22 au test.

- Avec un score de césure à 23/24, Jean ne se verra pas attribuer le niveau B1. Mais si l'on répliquait la procédure de détermination des scores de césure avec un échantillon différent de panélistes, nous pourrions parvenir à un score de césure légèrement différent pour A2/B1 de telle sorte que Jean se verrait assigner le niveau B1 avec un score de 22. Il reste donc une incertitude sur nos décisions de par la variabilité de la moyenne des scores de césure à travers les répliques de la procédure de détermination des scores de césure. Cette incertitude est quantifiée par l'erreur standard des scores de césure, comme cela a été discuté précédemment.

- Même si l'on prenait une unique procédure de détermination des scores de césure, l'on pourrait se tromper à l'égard de Jean, notamment si Jean avait été dans un mauvais jour au moment du test (ce qui impliquerait une erreur de mesure négative) alors qu'en moyenne il aurait un score supérieur au point de césure A2/B1. La variation entre les scores observés et les scores vrais est exprimée par la fidélité du test (ou par le concept d'erreur standard de mesure). Par conséquent, au cours de la validation d'une procédure de détermination des scores de césure, il est indispensable de relier les caractéristiques de la détermination des scores de césure à celles du test lui-même pour obtenir une idée précise des sources d'erreur et d'inconsistance.

- Le troisième type d'erreur, qui peut être fait au cours de la détermination des scores de césure, repose sur des erreurs systématiques. Si des membres du panel sont trop indulgents, cela pourrait conduire à des scores de césure excessivement bas et donc à une catégorisation des candidats en B1 alors qu'ils ne le mériteraient pas. Les erreurs systématiques influencent directement la validité externe de la procédure. Ce point sera discuté de manière plus détaillée dans la prochaine section.

Dans cette section, nous nous concentrerons essentiellement sur la deuxième source de variabilité : la variation au niveau des décisions dues à la relative fidélité du test. Nous pouvons avoir une bonne idée des effets du manque de variabilité en constituant un échantillon de candidats qui passe deux fois le même test et en construisant un tableau de fréquences pour voir les candidats qui sont classés de la même façon. Les indices d'accord (absolu ou le kappa de Cohen) donneraient une indication de la consistance des décisions.
Malheureusement, administrer à deux reprises le même test aux mêmes candidats est rarement réalisable dans un contexte d'examen, c'est pourquoi on a recours à des modèles psychométriques pour dériver les mesures de consistance à partir de l'administration du test. Une approche intéressante est proposée par Livingston and Lewis (1995), nous la discutons brièvement ici. En partant du travail de Lord (1965), ils considèrent une distribution des scores vrais pouvant être estimée à partir de la distribution des scores observés d'un

[38] Cependant, notez (voir Section 6.8.3.) que la conversion des scores au test sur la variable latente via la courbe caractéristique du test implique l'utilisation de l'estimation de la probabilité maximale qui peut être sévèrement biaisée quand les points de césure sont extrêmes.

échantillon représentatif de candidats, ou en utilisant un modèle à deux ou quatre paramètres[39].

Si la distribution est connue (ou précisément estimée), et si les scores de césure sont donnés, alors :

- La proportion de la population qui sera assignée à chaque catégorie dans le cadre de multiples points de césure peut être déterminée.
- A partir des hypothèses du modèle et de la fidélité du test, on peut déterminer quelle proportion de la population sera catégorisée dans chaque niveau sur la base du score au test.

Dans la partie gauche du tableau 7.9., un exemple de tableau est proposé pour trois catégories (niveaux). Les lignes indiquent la catégorie vraie. Avec la colonne « Marg » (pour marginale), on peut voir que 16.04% de la population appartient au niveau A2, 27.34% à B1 et 56.62% à B2. La fidélité du test a été estimée à 0.9. Si un test de même fidélité (pas nécessairement celui qui fait l'objet de l'étude, mais un test présentant des caractéristiques psychométriques identiques) est administré à une même population, on s'attend à ce que 21.17% des candidats soient catégorisés en A2 sur la base de leur score (cf. ligne du bas), et que 14.95% soient vraiment catégorisés en A2. A partir de la diagonale du tableau, nous pouvons déterminer un indice d'accord absolu qui vaut 0.1495 + 0.2002 + 0.4426 = 0.7922.

Tableau 7.9: exactitude de la décision

	Un Test				**Le test faisant l'objet de l'étude**			
	A2	**B1**	**B2**	**Marg**	**A2**	**B1**	**B2**	**Marg**
T(A2)	0.1495	0.0109	0.0000	***0.1604***	0.1511	0.0102	0.0000	***0.1614***
T(B1)	0.0617	0.2002	0.0115	***0.2734***	0.0624	0.1874	0.0119	***0.2618***
T(B2)	0.0005	0.1232	0.4426	***0.5662***	0.0005	0.1154	0.4611	***0.5769***
Marg	***0.2117***	***0.3343***	***0.4540***	***1***	***0.2140***	***0.3130***	***0.4730***	***1***

La partie gauche du tableau 7.9., a été estimée à partir de la distribution des scores observés de 1000 candidats, où 214, 313 et 473 ont été respectivement assignés aux niveaux A2, B1 et B2. On peut toutefois voir que la fréquence attendue en A2 n'est pas 214 mais de 211.7 (= 1000 x 0.2117). Pour adapter ce tableau de telle sorte que les proportions de chaque groupe correspondent exactement à celles observées, on doit multiplier chaque proportion du tableau (à l'exception des marges) par la proportion observée et diviser par la proportion attendue en colonne. Par exemple, pour la première ligne et la première colonne, nous trouvons 0.1495 x 0.2140/0.2117 = 0.1511. Les valeurs des neuf cellules sont représentées dans la partie droite du tableau 7.9. Les lignes marginales sont simplement la somme des valeurs de chaque colonne. L'indice d'accord absolu pour ce tableau ajusté est 0.7996.

Outre le fait de disposer d'une information valable quant à la précision des décisions par un indice d'accord, les deux tableaux indiquent également une différence marquée au niveau des faux positifs et des faux négatifs : la proportion de faux positifs (ceux qui sont classés au-dessus de ce qu'ils méritent) est environ de 2% tandis que le taux des faux négatifs est d'environ 18%.

Pour évaluer la consistance des décisions, c'est-à-dire dans quelle mesure les décisions différentes ou identiques seraient prises si deux administrations de test étaient utilisées, deux tableaux similaires à ceux du tableau 7.9.peuvent être conçus. Ces tableaux sont représentés dans le tableau 7.10. La seule différence entre ces deux tableaux (7.9. et 7.10.) tient en la signification des lignes. Alors que dans le tableau 7.9., les lignes indiquent la

[39] Dans le modèle à deux paramètres, il est considéré que le score vrai (la proportion d'items corrects) suit une distribution de type beta; dans le cas à quatre paramètres, il est également considéré que le score vrai minimum et le score vrai maximum peuvent être différents de zéro et un respectivement et qu'ils doivent également être estimés depuis les données observées. Les détails techniques du modèle sont particulièrement compliqués.

classification sur la base du score vrai, dans le tableau 7.10., les lignes indiquent la classification sur la base d'une administration indépendante du test. Ainsi, la partie gauche du tableau indique de façon jointe les probabilités de classifications reposant sur deux administrations indépendantes (un test et un autre de même fidélité) tandis que la partie droite donne accès aux probabilités pour l'administration et celle d'un autre test de même fidélité.

Dans ce dernier cas, les erreurs de mesure se produisent au cours des deux administrations, ainsi l'indice d'accord sera plus faible dans le cas du test de précision. Pour les deux cas du tableau 7.10., l'indice d'accord est d'environ 0.77.

Tableau 7.10: consistance de la décision[40]

	Un test				Ce test			
	A2	**B1**	**B2**	**Marg**	**A2**	**B1**	**B2**	**Marg**
A2	0.1663	0.0448	0.0007	***0.2117***	0.1681	0.0419	0.0007	***0.2107***
B1	0.0448	0.2212	0.0683	***0.3343***	0.0453	0.2071	0.0712	***0.3236***
B2	0.0007	0.0683	0.3851	***0.4540***	0.0007	0.0640	0.4012	***0.4658***
Marg	***0.2117***	***0.3343***	***0.4540***	***1***	***0.2140***	***0.3130***	***0.4730***	***1***

La différence la plus remarquable entre les tableaux 7.9. et 7.10. est que, dans le dernier cas, les deux tableaux sont essentiellement symétriques, la proportion dans la cellule (A2, B1) étant (approximativement) la même que la proportion dans la cellule (B1, A2). Pour la partie gauche, la symétrie est complète, cela est nécessairement le cas puisque c'est le résultat de deux administrations totalement indépendantes de deux tests parallèles. Cela signifie que dans ce cas la différence entre les faux négatifs et les faux positifs n'a pas de signification ; ils peuvent seulement être considérés d'une manière significative à partir des tableaux d'exactitude.

Pour voir l'influence de la variation des scores de césure, les tableaux d'exactitude peuvent être de nouveau calculés avec des scores de césure différents, et le résultat peut être alors comparé, en particulier concernant leurs taux de faux positifs et faux négatifs.

Une méthode moins sophistiquée pour calculer la consistance de la décision nous vient de Subkoviak (1988). Une consultation bien documentée, avec les tableaux nécessaires pour mettre en œuvre la méthode, peut être effectuée dans le chapitre 16 de Cizek and Bunch (2007). La méthode de Livingston et Lewis est plus polyvalente parce qu'elle est applicable à la fois pour des situations avec de multiples scores de césure et des situations où les items à crédit partiel et binaires sont utilisés, pondérés identiquement ou non.

7.5. Validation externe

Le principal résultat d'une procédure de détermination des scores de césure est une règle de décision pour assigner les candidats à un petit nombre de niveau du CECRL sur la base de leurs performances à un examen. Généralement, la performance au test a déjà été résumée par un nombre unique, le score au test.

Dans ce manuel, l'accent a été mis sur le fait que les procédures permettant de parvenir à une telle règle de décision sont complexes et chronophages, qu'il y a de nombreux pièges possibles, et que le résultat n'est jamais parfait ; notamment en raison de l'erreur de mesure dans le test et de la variance résiduelle dans le jugement des panélistes. Si toutes les procédures ont été suivies très attentivement, si l'examen dispose d'une validité de contenu

[40] Les tableaux 7.9 et 7.10 ont été calculés à partir du programme BB-CLASS développé par R.L. Brennan, librement accessible par "Center of Advanced Studies in Measurement and Assessment (CASMA) of the University of Iowa". Le programme peut être téléchargé à partir du lien suivant www.education.uiowa.edu/casma/ Lorsque le téléchargement est effectué, un manuel est inclus ainsi que les données et un fichier permettant d'aboutir aux tableaux 7.9. et 7.10. Bien qu'il y ait de nombreuses variations techniques dans l'utilisation du programme, les valeurs par défaut donneront généralement de bons résultats.

adéquate et d'un haut degré de fidélité, et si l'erreur standard des scores de césure est faible, on pourrait penser que le travail est accompli et résumer les résultats par un tableau indiquant l'exactitude des décisions, comme dans la partie gauche du tableau 7.9., tout en tenant compte des limites.

Selon ce raisonnement, le point faible est qu'un tel résultat dépend totalement des procédures mises en œuvre par la même personne ou le même groupe de personnes et des données collectées en une seule occasion sur un seul groupe de candidats, et sur une seule situation d'examen. Cela pourrait être considéré comme étant trop restreint pour garantir la véracité, c'est-à-dire la validité, d'une affirmation telle que : « si un étudiant obtient un score de 39 ou plus à mon test, il peut à juste titre être considéré du niveau B2 ». En général, la faiblesse réside dans le contraste entre la particularité des procédures et la généralité des affirmations.

La validation externe vise à fournir des preuves en provenance de sources indépendantes et qui corroborent les résultats et les conclusions de ses propres procédures. Parmi l'ensemble des preuves fournies, toutes ne sont pas indépendantes de la même façon vis-à-vis de l'information que l'on doit utiliser dans la détermination des scores de césure ; de même parmi l'ensemble des preuves fournies toutes ne sont pas convaincantes avec le même poids.

- Les preuves pourraient provenir de résultats des mêmes candidats sur un autre test ou une autre procédure d'évaluation.

- Les preuves pourraient être fournies par une autre procédure de détermination des scores de césure en utilisant le même panel ou un panel indépendant, conduit par les mêmes organisateurs ou par une équipe indépendante.

Voici un résumé du type de preuves qui pourrait être fournies pour justifier l'affirmation relative aux règles de décision qui émanent de ses propres procédures pour relier son examen au cadre. On pourrait tenter de tout faire mais ce serait irréaliste parce que la collection de preuves serait particulièrement couteuse ; en outre, toutes les études ne corroboreraient les résultats pas de façon comparable.

Dans cette section, quelques exemples de validation externe seront discutées et des arguments, ainsi que leurs limites et leur caractère persuasif (ou l'absence de ce caractère), seront présentés. Cependant, en premier lieu, une remarque générale doit être faite. Dans la théorie des tests, le problème de la validité externe est généralement considéré en montrant la correspondance entre les résultats au test et des critères externes. Parfois, les mesures du critère externe sont, d'une certaine manière, considérées comme absolues. Mais en réalité, aucun critère n'est parfaitement valide. Prenons le succès académique comme exemple. Obtenir un master à l'université peut être considéré sans erreur de mesure. Un master est alors certainement utile, mais non absolu, en termes de critères des habiletés mentales. En effet, quelques étudiants pourraient échouer à l'université pour des raisons largement indépendantes de leur habileté mentale et quelques étudiants pourraient réussir sans que ce soit mérité ; aucun système d'examen n'est infaillible. Ainsi, il est préférable de considérer toutes les mesures de critères comme faillibles de la même façon que les tests le sont, c'est-à-dire qu'une part de leur variance est indésirable et non pertinente pour montrer la validité de la procédure d'un test, comme avec les résultats de la détermination des scores de césure.

7.5.1. Validation croisée

Comme cela a déjà été abordé au cours du chapitre 6, la principale faiblesse des deux méthodes (populaires) centrées sur le candidat, la méthode des groupes contrastés et la méthode du groupe limite, est liée au fait que l'information sur les candidats impliqués provient, d'une certaine manière, d'une source non divulguée, le jugement de leur propre enseignant. Ce jugement peut (et devrait) être considéré comme un résultat de test, mais, en général, il est particulièrement difficile d'obtenir de l'information sur les qualités psychométriques de ces jugements. Il n'y a pas d'opportunité pour discuter ces résultats, puisqu'ils relèvent d'opinions privées, celles des enseignants.
De plus, en déterminant les scores de césure avec ces méthodes, la construction des tables de décision est conseillée, elle maximise la correspondance entre le score au test et le jugement des enseignants. Ceci implique que les points de césure sont dépendants de l'avis d'un petit nombre d'enseignants, généralement un échantillon de petite taille ou au mieux de taille modérée. Statistiquement parlant, cet effet « tire parti de la chance» et il est important de montrer, par une technique de validation croisée, comment cet effet est significatif. Il suffit simplement pour cela d'utiliser les résultats (les scores de césure) provenant de la procédure ayant permis de les déterminer et de les appliquer à un échantillon indépendant. La comparaison des indices de qualité sur l'échantillon originel et sur l'échantillon de validation croisée donne une indication du degré de généralisation des résultats. L'indice d'accord absolu ou le kappa de Cohen peuvent être utilisés ici en qualité d'indice de la qualité, puisque tous les candidats sont assignés à un niveau par le jugement de l'enseignant et par la règle de décision issue de la procédure de détermination des scores de césure.

Il y a plusieurs façons de mettre en œuvre une validation croisée :

- L'échantillon originel peut être scindé en deux (de façon aléatoire). Une moitié est utilisée pour mettre en œuvre la procédure de détermination des scores de césure, l'autre moitié est utilisée pour la validation croisée. On peut également procéder d'une façon plus équilibrée en utilisant chaque moitié de l'échantillon pour la procédure de détermination des scores de césure et l'autre moitié pour la validation croisée, étant donné que la détermination des scores de césure consiste à établir des tableaux et à partir de ces tableaux, de prendre des décisions. Bien qu'une telle procédure soit certainement valable et conseillée, elle devient significative lorsque l'échantillon total est suffisamment large pour produire deux sous-échantillons de taille conséquente. En outre, son pouvoir persuasif est limité. Le critère d'information provenant des mêmes sources (les enseignants), il ne sera pas possible de détecter, dans la validation croisée, si, par exemple, les enseignants ont tendance à être indulgents.

- Pour opérer un contrôle, on peut subdiviser l'échantillon des candidats de manière à avoir tous les étudiants de la moitié des enseignants en tant qu'échantillon permettant la détermination des scores de césure et l'autre moitié pour la validation croisée. Si les tailles d'échantillon sont suffisantes, on peut même avoir recours à quatre échantillons. Premièrement, on subdivise les enseignants en deux moitiés, puis on divise l'échantillon des candidats de chaque enseignant en deux moitiés équivalentes.

- La procédure précédente peut être aisément appréhendée comme un cas de véritable validation. Si la taille de l'échantillon utilisé pour déterminer les scores de césure n'est pas suffisamment importante pour opérer une subdivision, on peut utiliser l'échantillon total pour déterminer les scores de césure puis collecter les données sur un échantillon totalement indépendant, en provenance d'autres écoles. La validation exige une administration du test (ou de l'examen) sur cet échantillon de validation autant que la demande aux enseignants d'évaluer les candidats sur les niveaux du CECRL. Mais en principe, cette procédure ne diffère pas de la précédente, puisque l'échantillon permettant la détermination des scores de césure et l'échantillon de validation peut être facilement interchangé.

Parmi les méthodes, discutées dans le chapitre 6, permettant la détermination des scores de césure, la méthode des groupes contrastés et celle du groupe limite ont un statut spécial puisqu'un critère de mesure (le jugement des enseignants) est un composant de la méthode de détermination des scores de césure elle-même. On pourrait penser que cela est nécessairement vrai pour toutes les méthodes centrées sur le candidat, mais ce n'est pas le cas. Prenons par exemple le cas de la méthode des corpus de productions. Dans cette méthode l'information relative aux candidats, et à disposition des panélistes, est leur performance, ainsi que quelques informations de rang (cf. le classement ordonné des dossiers, même si cela n'est pas strictement nécessaire). Aucune information, pas même le niveau du CECRL auquel les étudiants sont, n'est fournie aux panélistes. La méthode repose totalement sur la performance des candidats à l'examen. Il en est quasiment de même pour toutes les méthodes centrées sur le test et discutées dans le chapitre 6 : les points de césure sont totalement déterminés par les jugements des panélistes sur le matériel du test. Même en leur fournissant une information relative à l'impact (comme la distribution des candidats sur l'ensemble des niveaux), ils sont confrontés aux conséquences de leurs décisions et ça ne reflète pas une catégorisation selon les niveaux du CECRL provenant d'une autre source. C'est pourquoi, le concept de validation croisée ne prend pas sens pour ces méthodes.

La validation externe de ces procédures de détermination des scores de césure impliquera par conséquent la comparaison des résultats de la procédure de détermination des scores de césure (la règle de décision) avec les résultats d'une autre règle de décision. Cette comparaison pourrait prendre essentiellement deux formes : l'utilisation des distributions marginales ou des classifications croisées. Elles vont être maintenant discutées.

7.5.2. Comparaison des distributions marginales

Supposons que des données d'un échantillon représentatif aient été calibrées par un modèle de la TRI, et qu'une règle de décision pour assigner aux candidats un des quatre niveaux A1, A2, B1, B2 du CECRL, ait été, disons, dérivée de la méthode du marque-page. Ainsi, les candidats appartenant à l'échantillon de calibrage pourraient être classés dans l'un de ces quatre niveaux. Si l'on dispose d'information sur un autre échantillon, également représentatif de la même population parente, et classé à partir d'une autre méthode, par exemple par le jugement de leurs enseignants, on peut alors construire un tableau 2x4 comme celui le tableau 7.11. Dans ce tableau, l'échantillon 1 fait référence à l'échantillon de calibrage alors que l'échantillon 2 fait référence à un échantillon indépendant de validation.

Tableau 7.11: distributions marginales sur les niveaux (occurrences)

	A1	**A2**	**B1**	**B2**	**Total**
Echantillon 1	98	124	165	84	471
Echantillon 2	39	74	78	63	254
Total	137	198	243	147	725

Parce que les échantillons sont de taille différente, la comparaison par une simple inspection du tableau est délicate. Convertir les observations en pourcentage (en ligne) rendra la comparaison plus facile. Les résultats sont représentés dans le tableau 7.12., et montrent que dans l'échantillon indépendant il y a relativement plus de candidats qui sont placés aux niveaux A2 et B2 et moins aux niveaux A1 et B1 que dans l'échantillon de calibrage. On peut tester statistiquement cette différence par un chi² qui vaut ici 7.94, p = 0.047 (avec trois degrés de liberté), ce qui signifie qu'il y a une différence significative dans l'assignation des niveaux entre les deux méthodes[41].

[41] Le test du chi² doit être mis en œuvre sur les occurrences (Tableau 7.9), et non sur les pourcentages du Tableau 7.10.

Tableau 7.12: distributions marginales sur les niveaux (pourcentages)

	A1	A2	B1	B2	Total
Echantillon 1	20.8	26.3	35.0	17.8	100.0
Echantillon 2	15.4	29.1	30.7	24.8	100.0

Cet exemple, si simple qu'il puisse paraitre, illustre la délicatesse du processus de validation. Sur un fondement statistique (le test du chi²), il pourrait être déduit que les différences systématiques dans l'attribution des niveaux du CECRL reposent sur les deux méthodes. Mais cela n'explique pas pour autant l'origine de ces différences. Prenons comme exemple le niveau B2, là où la différence est la plus large. Il se pourrait que la méthode du marque-page ait conduit à un point de césure B1/B2 trop sévère. Ce n'est cependant pas du tableau que l'on peut le déduire, puisqu'il se pourrait aussi que les enseignants aient été trop indulgents en attribuant le B2. Chercher ce qui s'est réellement produit ici exigerait un nombre supplémentaire d'études et de données. Interviewer les enseignants sur leur raisonnement et leur raison d'attribuer le niveau B2 pourrait révéler qu'ils n'ont pas bien intégré la description du CECRL pour le B2. Par ailleurs, ils peuvent avoir été inégaux dans leurs jugements, en attribuant une attention restreinte à quelques « être capable de » du B2 ; peut-être ceux ayant fait l'objet de discussion au cours de la méthode du marque-page. Inversement, l'examen utilisé pour déterminer les points de césure pourrait être insuffisant et avoir négligé un nombre d'aspects que les enseignants prennent en compte pour effectuer un jugement holistique sur le niveau de leurs étudiants. Un tableau comme le tableau 7.12., peut être utilisé pour cerner le problème et pour, au mieux, suggérer une explication possible ; il faut toutefois faire preuve d'une bonne dose de créativité pour détecter les causes réelles de ces différences.

7.5.3. Table de décision

Si les deux jeux des règles de décision peuvent être appliqués au même échantillon de candidat, plus d'information peut alors être obtenue. Les résultats d'une méthode de détermination des scores de césure (les règles de décision) peuvent être, en général, directement appliqués à un échantillon de candidats, par exemple un échantillon de calibrage. Si l'on disposait d'un autre jeu de règles de décision, soit en provenance des jugements holistiques des enseignants soit d'une autre méthode de détermination des scores de césure, et que ces règles soient applicables au même échantillon de candidats, on pourrait alors concevoir une table de décision présentant conjointement les probabilités (ou les occurrences) pour chaque paire de niveau. Ces tableaux sont comparables à la partie droite du tableau 7.10., à une différence *essentielle* près : les colonnes font référence à l'attribution des niveaux depuis la méthode qui fait l'objet de l'étude (comme dans le cas du tableau 7.10.) mais les lignes reposent sur l'assignation des niveaux depuis un ensemble *indépendant* de règles de décision, et non pas depuis des hypothèses comme c'était le cas lors du jugement de la consistance de la décision. Si l'ensemble indépendant de règles de décision conduit vraiment à la même chose que les règles de décision issues de la méthode de détermination des scores de césure, c'est-à-dire si les deux ont la même validité de construit et la même fidélité, alors le tableau de décision devrait être essentiellement le même que celui de la partie droite du tableau 7.10. Par conséquent, construire et comparer les deux tables pourraient mettre à jour des informations utiles :

- Les distributions marginales pourraient être comparées de la même manière que celle discutée précédemment avec les échantillons indépendants.

- Les indices d'accord (absolu, adjacent et le kappa de Cohen) pourraient être calculés sur les deux tables puis être comparés.

- La comparaison des cellules en dehors des diagonales des deux tableaux est la plus pertinente pour la validation. Il a été fait mention précédemment à l'égard de l'évaluation de la consistance de la décision de l'aspect symétrique de la table de décision. Dans le cas de la validation à partir d'un autre jeu de règles de décision la symétrie ou le manque

de symétrie est un résultat purement empirique et pourrait être utile pour comprendre la validité de la méthode de détermination des scores de césure. Le concept des faux positifs et des faux négatifs prend ici toute sa place. On doit toutefois clairement définir ce qui est entendu par ces termes dans un contexte de validation. Il pourrait être éclairant de définir les *faux négatifs* comme les cas où les décisions selon les points de césure qui font l'objet de l'étude conduisent à un niveau *plus faibles* que les règles selon les critères ; et les *faux positifs* comme les situations où l'assignation d'un niveau selon les conclusions de la procédure de détermination des scores de césure est à un niveau *plus élevé.* Si dans l'étude de validation, le taux de faux positifs est plus élevé que celui de faux négatifs, cela signifie que la méthode de détermination des scores de césure qui fait l'objet de l'étude est plus indulgente que les règles selon les critères ; pour la situation inverse, il s'agira de plus de sévérité[42].

Exercice travaillé. Un exercice travaillé peut aider à illustrer comment les tableaux de décision pourraient être utilisés pour relier les résultats au test aux autres données d'évaluation, par exemple le jugement holistique, selon les niveaux du CECRL, par les enseignants. Le principe en utilisant des tableaux à deux entrées n'est pas complexe en lui-même. Le principal problème avec le recours au jugement holistique des enseignants comme critère externe n'est pas l'analyse. Les enseignants doivent absolument connaître vraiment (a) les niveaux du CECRL et (b) la compétence des individus concernés ; cela pourrait donc ne pas être pratique avec des enseignants qui voient des classes de 30 élèves seulement à quelques reprises au cours de la semaine.

North (2000b) rapporte l'utilisation des jugements des enseignants comme critère externe pour référencer les banques d'items pour l'anglais, l'allemand, le français et l'espagnol sur l'échelle Eurocentres, qui distingue neuf niveaux. Des points de césure provisoires ont été proposés auparavant avec une variante simplifiée de la méthode du marque-page. Au cours de l'étude de validation externe, les évaluations des enseignants étaient utilisées pour vérifier, à travers la validation externe indépendante, les points de césure proposés pendant le développement de la banque d'items pour l'allemand. Les enseignants furent interrogés pour assigner à chaque élève de leur classe un niveau pour le domaine testé par la banque d'items : connaissance du système langagier. La figure 7.3., montre la relation entre la performance standard (axe X) et les jugements des enseignants (critère) sur l'axe Y.

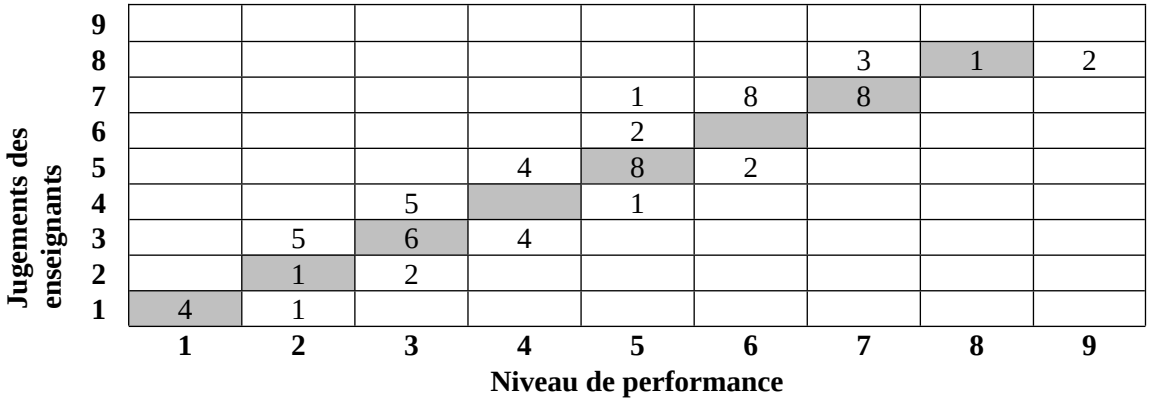


Figure 7.3: table de décision à 9 niveaux

La relation entre la classification par la performance et par les enseignants apparait régulière et équilibrée, avec une corrélation de .93. Néanmoins, seuls 28 des 68 sujets (soit 41%) ont effectivement été affectés au même niveau, et ce en dépit de la forte corrélation. Il y a huit

[42] Des analyses plus sophistiquées pourraient être conduites ; par exemple, en choisissant parmi des méthodes polyvalentes comme les analyses de type log-linéaire pour mieux situer les différences significatives. Pour plus d'information, on peut consulter Fienberg (1977) pour une introduction facilement accessible ou Fienberg et al (1975) si l'on souhaite une information plus élaborée.

apprenants placés au niveau 7 par les enseignants et au niveau 6 par le programme. Ceci est dû à un seul enseignant indulgent. Toutefois, même si ces huit candidats ont été affectés à la bonne place dans le tableau, seulement 50% des candidats auraient reçu exactement la même affectation de l'enseignant et du test. L'indice d'accord adjacent est néanmoins de 67/68=0.985 : seul un candidat a été assigné à deux catégories au-dessus par les enseignants.

L'échelle Eurocentres coupe les niveaux du CECRL en deux (à l'exception du niveau A1). Si une table de décision est créée en utilisant seulement les niveaux du CECRL, comme c'est le cas dans le tableau 7.13., la proportion de classification correcte augmente considérablement, de 41 à 73.5%, puisque 50 des 68 apprenants reçoivent maintenant le même niveau du CECRL, que ce soit par les enseignants ou par l'application des scores de césure.[43]. L'indice d'accord adjacent est égal à un.

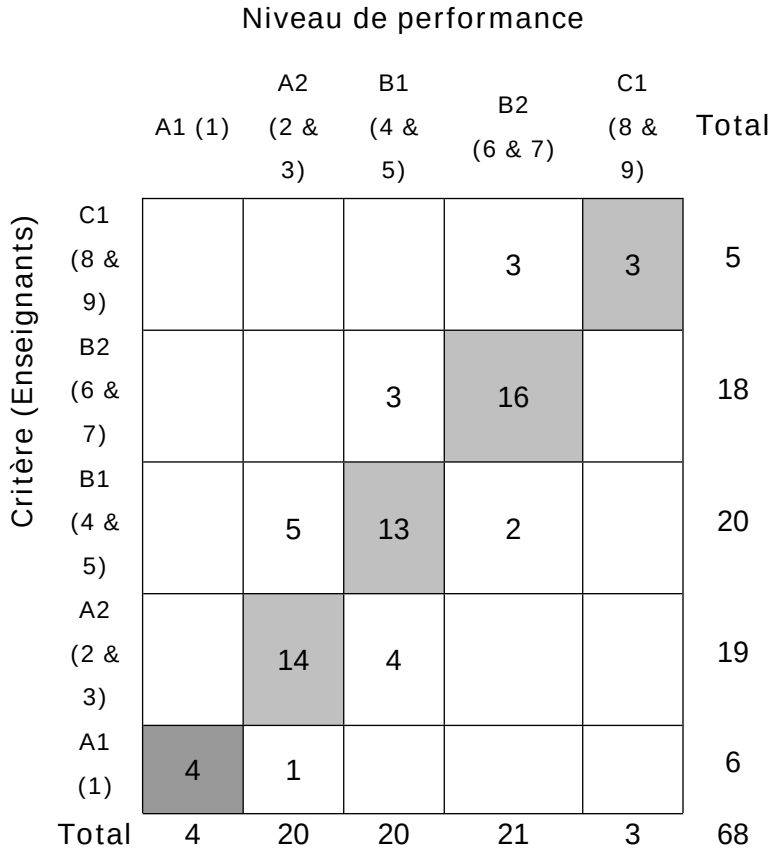


Figure 7.4: table de décision pour cinq niveaux

Si les évaluations des enseignants sont utilisées, il serait sage de considérer une telle procédure évaluative comme une forme de test et de porter attention à la validité interne comme pour un test. A cet égard, des remarques sont listées ci-après.

- Si le jugement est uniquement holistique, comment peut-on alors estimer sa fidélité ? D'un point de vue psychométrique, cela revient à utiliser un test à un item ; il n'y a donc pas d'espace pour le calcul des indices de la consistance interne. Dans un pareil cas, on

[43] Exprimer les niveaux Eurocentres dans les termes du CECRL peut être justifié par ce qu'un nombre considérable de descripteur du CECRL ont leurs origines dans l'échelle Eurocentres. En effet, les descripteurs d'Eurocentres survivent mieux au processus de validation qualitative que ceux de la plupart des autres échelles, puisque les formulations d'Eurocentres tendent à être concrètes et positives. La corrélation sur les rangs pour 73 descripteurs de l'interaction et de la production est de .88. La classification partagée montrée par une table de décision est de 70% (See North 2000a: 337.)

devrait concevoir une procédure de re-test, et on doit considérer le problème sous l'angle de la faisabilité d'un jugement répété.

- Même avec des jugements comprenant des listes de vérification des descripteurs, le correcteur a besoin de bien connaître la compétence du candidat. Les procédures de détermination des scores de césure centrées sur le candidat, discutées dans le chapitre 6, impliquent que le correcteur puisse seulement juger un nombre limité de candidats (ses étudiants). En outre, en sollicitant les enseignants pour corriger leurs propres apprenants, l'on pourrait avoir à faire face au fait qu'ils pourraient exagérer les différences entre leurs apprenants les plus faibles et les plus forts.

On peut probablement éviter de rencontrer les problèmes qui viennent d'être mentionnés avec plusieurs correcteurs qui donnent des jugements sur des échantillons de comportements plus faciles à observer comme les productions écrites. L'utilisation de juges, indépendants du processus de détermination des scores de césure et correctement préparés, ainsi que d'outils d'évaluation appropriés (voir section B du Supplément au Manuel) est une option qui a été utilisé de façon réussie en Finlande. La variance des correcteurs pourrait alors être étudiée avec une étude de type G (voir section E sur Supplément au Manuel) ou par une analyse Rash à multiples facettes (Linacre 1989). On peut par exemple y avoir recours avec le programme FACETS (Linacre 2008). Ce modèle, qui prend en compte une troisième facette (le correcteur), estime la sévérité/indulgence des correcteurs et en tient compte pour l'estimation des habiletés des candidats.

7.5.4. Quelques scénarii

Nous avons indiqué dans les paragraphes précédents que toutes les procédures de validation visaient à comparer différents jeux de règles de décision, soit en utilisant des échantillons indépendants de candidats soit sur le même échantillon. Au cours de cette sous-section, quelques scénarii seront décrits. Cette description pourrait aider à élaborer une décision sur la base d'une sage et solide comparaison.

Une distinction importante entre les méthodes de détermination des scores relève de la différence entre les méthodes centrées sur le candidat et celles centrées sur le test. Il semble naturel par conséquent de diriger la validation de la méthode appartenant à l'une des classes vers une comparaison avec la méthode appartenant à l'autre classe. Il faut rester prudent en accomplissant une pareille comparaison. Prenons comme exemple de méthodes contrastées celle du marque-page (ou la variante du Cito) et celle du corpus de productions. Il y a plusieurs arguments plaidant contre un tel scénario.

- La méthode du marque-page est adaptée pour les tests ou examens qui peuvent être calibrés en ayant recours à la TRI, par exemple des tests fortement itémisés. En revanche, la méthode du corpus de productions repose sur des jugements holistiques et est particulièrement adaptée pour des examens qui ne sont généralement pas adaptés pour les analyses de la TRI, comme les épreuves de production orale ou de production écrite. Les conséquences seront qu'au moins une des méthodes souffrira d'une approche inappropriée, ce qui aura tendance à rendre les comparaisons caduques.

- Considérons un examen avec un degré de complexité tel qu'il permet d'avoir recours à des méthodes aussi différentes pour la détermination des scores de césure que celle du marque-page ou du corpus de productions. D'un point de vue pratique, les mettre en œuvre pourrait être irréaliste puisque les deux méthodes exigent une préparation spécifique, et donc un temps lui étant consacré[44]. De l'ennui et/ou de la fatigue chez les panélistes pourraient être prohibitif pour une telle approche complexe ; ce serait également le cas avec un défaut de ressources.

[44] Généralement, les panélistes invités pour participer à la détermination des scores de césure pour un examen de langue ne connaissent pas très bien la TRI. Leur donner une introduction à ces notions est difficile car très chronophage. Même si cela est faisable, cela ne doit pas être sous-estimé.

D'un autre côté, il est toujours possible (si les ressources sont suffisantes) d'appliquer deux méthodes différentes en utilisant deux panels d'experts indépendants et de mettre en œuvre les deux méthodes à des périodes différentes. La mise en œuvre de deux procédures coûteuses pourrait ne pas être adaptée à certains contextes de détermination des scores de césure, mais en même temps cela pourrait se révéler pertinent dans le cadre de projets aux enjeux internationaux.

Un compromis intéressant pourrait être trouvé en combinant une méthode centrée sur le test avec la méthode des groupes contrastés ou la méthode du groupe limite, si les panélistes peuvent donner des jugements holistiques sur un nombre suffisamment important de candidats ayant participé au test qui fait l'objet de l'étude. Nous attirons toutefois votre attention sur l'exemple traité précédemment.

7.5.4.1. Tirer parti du calibrage de la TRI

En utilisant un modèle de la TRI pour relier les items ou les tâches entre eux, on dispose de nombreuses opportunités pour mettre en regard différentes méthodes de détermination des scores de césure. Le cas échéant, on peut tirer un avantage du fait que la relation des items à l'habileté sous-jacente (latente) est connue (à un degré suffisant de précision) par une étude de calibrage. Ici, nous décrirons un scenario qui utilise cette relation, de façon explicite, dans une étude de validation d'une procédure particulière de détermination des scores de césure. Pour illustrer ce point, nous prendrons la variante du Cito de la méthode du marque-page.

La méthode implique que les deux procédures relatives aux scores de césure présentent des ensembles différents d'items pour les panélistes. Les panélistes, dans les deux procédures, pourraient ou non être les mêmes personnes. Dans le cas de personnes différentes, on doit prendre garde, pour les deux sessions correspondantes aux deux méthodes, à l'équivalence des panels d'experts du point de vue de leur composition. L'ensemble pour la première procédure de détermination des scores de césure pourrait être constitué de tous les items (ou un sous-ensemble) utilisés dans l'examen A, alors que le second jeu d'items contiendrait des items (ou un sous-ensemble) de l'examen B. Comme c'est le cas avec toutes les procédures relatives aux scores de césure qui reposent sur la TRI, les scores de césure sont définis dans le domaine de la variable latente. En utilisant les techniques discutées dans la section 6.8.3., ces repères sur la variable latente doivent être traduits en scores de césure, pour lesquels les caractéristiques d'items sont connues. En particulier, on pourrait transcrire ces scores de césure pour un test principalement constitué d'items utilisés au cours de la procédure de détermination des scores de césure ou à un test principalement constitué d'items non utilisés pour la procédure. La situation est résumée dans le tableau 7.13.

Les cellules grisées sont les conditions pour lesquelles les items utilisés sont en étroite relation (soit identiques, soit un large sous-ensemble des items) avec les items de l'examen. Les cellules en clair sont les plus sensibles : les items utilisés pour fixer les scores de césure sont des items autres que ceux véritablement utilisés pour l'examen.

Tableau 7.13: dispositif de procédure de détermination des scores de césure pour une paire

	Détermination des scores de césure à partir des items appartenant à	
	Examen A	**Examen B**
Score de césure pour ex.A		
Score de césure pour ex. B		

Parce que (virtuellement) personne ne prend part aux deux examens, les comparaisons empiriques sont seulement judicieuses à l'intérieur des lignes du tableau 7.13. ; ce qui signifie essentiellement que pour le même examen deux jeux de scores de césure ont été fixés et que l'on doit vérifier dans quelle mesure ils conduisent aux mêmes conclusions ou à des conclusions différentes en élaborant un tableau de décisions comme celles décrites

précédemment dans cette section et illustrées par le tableau 7.12. Cette procédure évaluative pourrait être mise en œuvre aux deux lignes du tableau 7.13. et offrirait alors l'opportunité de vérifier si une explication des différences de résultats (de par les deux procédures) est consistante avec les différences trouvées avec les mêmes méthodes à une autre occasion (par exemple, la ligne complémentaire du tableau 7.13.).

7.5.4.2. Utilisation des "Etre capable de"

Pour exploiter le CECRL dans la validation externe, une méthode consiste à évaluer les candidats qui alimenteront les données pour le test étudié avec des listes de vérification du type « Portfolio Européen des Langues » constitué de 30 à 50 descripteurs pertinents. De cette manière, chaque descripteur peut être inclus comme un item séparé dans l'analyse selon la TRI et ainsi calibré sur la même échelle d'habileté latente. Les jugements peuvent provenir des enseignants, ou des candidats eux-mêmes à travers l'auto-évaluation.

En combinaison avec la variante du Cito de la méthode du marque-page, cette information peut être utilisée pour valider la détermination des scores de césure comme dans l'exemple de la figure 7.5. Cette figure est la même que la figure 6.5. à une exception près, trois « être capable de » (calibrés comme des items) ont été ajoutés à la représentation. Supposons que le point de césure pour A2/B1 ait été fixé comme l'indique la ligne verticale sur la représentation selon la méthode décrite dans la section 6.9. Supposons également que les trois lignes en pointillé sur cette représentation formalisent les trois « être capable de » pour le niveau B1. Pour les deux situés plus bas, on constate que la performance correspondante au point de césure « maîtrise totale » est quasiment atteinte, alors que pour celui du haut on est guère éloigné de la maîtrise limite. Cette information (collectée de préférence avec plus de trois descripteurs de compétence), concaténée au contenu de ces descripteurs de compétence, donne une image détaillée de ce que signifie le point de césure selon les termes des descripteurs du CECRL.

Cette façon de valider la détermination des scores de césure peut en réalité être utilisée d'au moins deux manières. Une figure, comme la figure 7.5., peut être construite à la fin de la procédure de détermination des scores de césure pour évaluer la validité des résultats. Le cas échéant, la détermination des scores de césure et la validation sont considérées comme un processus linéaire. Les jugements en relation avec les descripteurs de compétences sont utilisés comme un critère externe d'une étude de validité externe. Mais si les résultats de la validation sont décevants (en indiquant par exemple que les panélistes ont été trop indulgents), la procédure de détermination des scores de césure, dans son ensemble, peut être perçue comme étant un échec et une perte de temps. Une approche plus efficace consiste à incorporer ce type d'information à la procédure de détermination des scores de césure, par exemple entre deux tours de jugements, comme information pertinente relative aux conséquences de la détermination des scores de césure et comme arguments pour adapter/moduler les jugements précédemment établis.

Il est vrai qu'avec cette dernière approche, la validation n'est pas véritablement indépendante de la procédure de détermination des scores de césure elle-même, puisque l'information sur les « être capable de » est utilisée au cours de la procédure elle-même. Toutefois, cela pourrait permettre un gain de temps notable. Une bonne documentation des résultats de l'ensemble des tours de jugements peut être un argument convaincant pour une validation totalement indépendante (validation externe au sens classique).

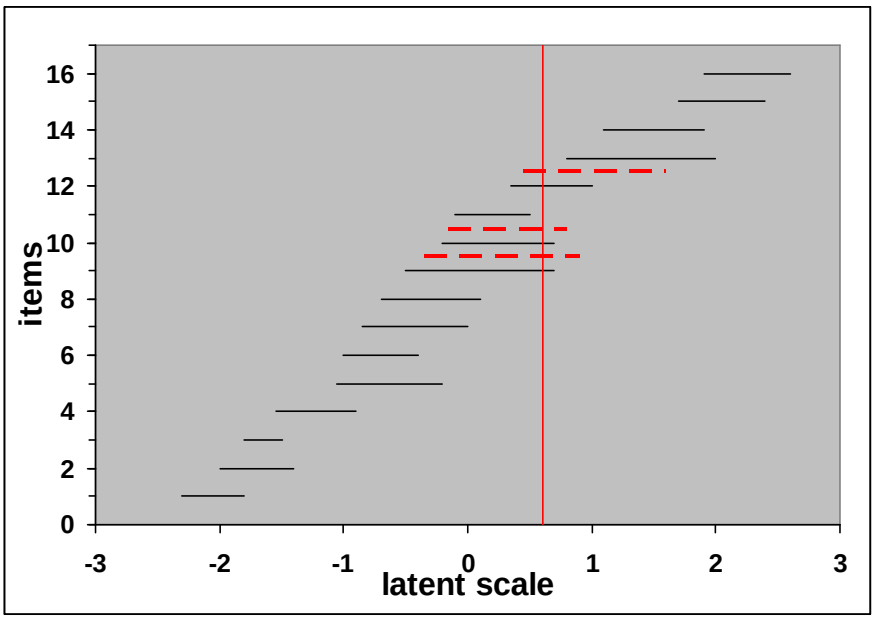


Figure 7.5: cartographie d'items avec des descripteurs de compétence

Avec des données de type « être capable de », on peut s'appuyer soit sur les évaluations des enseignements soit sur l'auto-évaluation. Le choix entre ces deux derniers reste problématique. La confiance dans les données issues de l'auto-évaluation devrait conduire, peut être de façon erronée, à conclure sur un caractère trop strict des scores de césure. Par conséquent, il est préférable de collecter à la fois les évaluations des enseignants et les données de l'auto-évaluation, ce qui permet d'ajouter du poids à l'argument de validité.

7.5.4.3. Détermination des scores de césure sur plusieurs langues

Dans le chapitre 6 (section 6.8.3), une procédure générale a été décrite pour relier différents examens ou tests au CECRL (exemple un test de français et un test d'anglais). La procédure repose essentiellement sur le plurilinguisme des panélistes. En effet, l'assignation d'un niveau pour un candidat, de façon identiquement juste dans les deux langues, ne peut être prise en compte. La fragilité de cette procédure tient au fait qu'on ne peut considérer pour acquis que tous les panélistes impliqués dans la procédure de détermination des scores de césure sont suffisamment compétents dans les langues concernées.

Pour jauger la signification des résultats, on doit mettre en œuvre une procédure de contrôle. Prenons l'exemple de l'anglais et du français :

- Au cours de la détermination des scores de césure impliquant deux langues, un équilibre doit être trouvé au niveau de l'expérience des panélistes. Cela devrait signifier que la moitié d'entre eux sont des locuteurs natifs de l'anglais et l'autre moitié du français, alors que l'ensemble doit avoir l'autre langue comme principale spécialité.

- Une procédure de détermination des scores de césure pour chaque langue est conseillée, puisqu'un contexte plurilinguistique pourrait créer des références spécifiques (de part le contexte inhabituel par exemple), ce qui rendrait les résultats impropres à la généralisation.

Ces deux considérations impliquent déjà un dispositif particulièrement compliqué pour tester la validité de la détermination des scores de césure adaptée à la validation croisée sur les langues. Idéalement, il faudrait :

- Une procédure de détermination des scores de césure dans laquelle la moitié des panélistes ont l'anglais comme langue native et le français comme langue de première

spécialisation, et l'autre moitié ont le français comme langue native et l'anglais comme langue de première spécialisation;

- une détermination des scores de césure pour le français dans laquelle la moitié des panélistes sont des locuteurs natifs de français et l'autre moitié sont des spécialistes du français;

- une détermination des scores de césure pour l'anglais dans laquelle la moitié des panélistes sont des locuteurs natifs de l'anglais et l'autre moitié sont des spécialistes de l'anglais.

Idéalement, les trois conditions mentionnées ci-dessus font référence à des panélistes indépendants. Mettre en œuvre un tel dispositif offre une possibilité pour comparer les points de césure à travers les langues, et, via le partage de l'expertise dans la détermination des scores de césure sur plusieurs langues, on pourrait faire des suggestions sur la manière dont on peut améliorer la procédure, ou bien suggérer son abandon. L'expérience du séminaire de calibrage sur les langues qui s'est tenu à Sèvres en Juin 2008 (Breton et al) a elle été très bénéfique.

7.6. Conclusion

La discussion relative à la validation externe dans ce chapitre pourrait apparaitre décevante à l'égard de nombreux points. En fait, elle ne propose pas de distinction claire entre ce qui relève du bon et du mauvais. Elle ne prescrit pas non plus de façon claire et univoque ce qu'il convient de faire pour une situation donnée.

Voici ci-après, en deux points, quelques raisons à cela :

Premièrement, il n'y a aucune autorité qui détient la vérité et la divulgation reste problématique. Les organismes de test aspirent à découvrir cette vérité encore non connue en effectuant un choix méthodologique (et/ou des méthodes psychométriques) approprié. C'est en faisant part de ces travaux à la communauté que dans le futur nous pourrons nous approcher de la vérité de si près que nous pourrons considérer que nous avons résolu le problème. A l'opposé, nous croyons que ce qui constitue un « B1 » est essentiellement une convention pratique, mais la formulation est si claire et si consistante que deux professionnels du monde des langues s'y référant signifieront essentiellement la même chose, même si leur culture, leur formation et leur expérience sont différentes et se réfèrent à des langues cibles différentes. Le CECRL constitue un système de référence dont l'objectif est de rendre de telles affirmations possibles. Du point de vue des études de validation, cela signifie que toute étude de validation, peut, en principe, offrir une critique constructive pouvant conduire à une référence plus affinée, équilibrée et élaborée ; ce qui est vrai de toute expérimentation d'hypothèses, de construits et de théories.

Deuxièmement, même dans le cas d'un système de référence largement accepté, les éléments déterminants des performances à un test de langue ou à un examen sont si variés (et pas toujours totalement compris) que toute tentative pour classer les études pour relier les performances au CECRL (soit en bonnes ou mauvaises) doivent être considérées comme simplistes et catégoriques. En fait, nous tentons de développer un système qui offre un éclairage sur les points forts et les points faibles de toute tentative, il ne serait donc pas réaliste de dresser un verdict catégorique et fini.

Est-ce une bonne ou mauvaise nouvelle ? Nous pensons qu'il s'agit simplement d'un état de l'art. Des conclusions plus fines devraient être dressées à partir d'une méta-analyse rigoureuse, qui pourrait résumer les résultats d'un large nombre d'études de validation, rigoureusement conduites sur les prochaines années. Il appartient à la génération présente de fournir les données nécessaires et la documentation pour cette méta-analyse. (Voir Plake 2008 pour une revue des enjeux et des recommandations).

Il faut ainsi espérer que les acteurs des procédures de détermination des scores de césure feront leur maximum pour prendre connaissance de l'information fournie dans ce Manuel, dans le Supplément au Manuel et les autres sources d'informations pertinentes. Il faut également espérer que ces procédures seront conduites et les rapports établis de façon transparente. En les analysant et en les comparant, les savoir-faire pour déterminer les scores de césure progresseront. La crédibilité accordée aux décisions sur les points de césure progressera également, ainsi que la portée des conséquences qui en découleront.

Les utilisateurs du manuel devraient considérer :

- *la meilleure manière d'obtenir les preuves de validité exigées ;*
- *quelles techniques ils seront capables de mettre en œuvre et dans quelle mesure ils auront besoin d'un support technique ;*
- *s'ils peuvent élaborer un argumentaire sur la validité à propos de la qualité du test et des procédures qui y sont associées (validité interne) de la qualité des procédures suivies pour relier l'examen au cadre, et en particulier pour la détermination des scores de césure (validité procédurale), de l'existence de résultats corroborés par des analyses indépendantes (validité externe) ;*
- *comment ils s'assurent, le cas échéant, que les points de césure sont comparables à travers les langues ;*
- *s'il y a, en particulier, suffisamment de preuves pour soutenir la validité des scores de césure ;*
- *comment ils mettront à la disposition de leurs collègues les détails de leurs conclusions.*

Références

AERA/APA/NCME (1999): American Educational Research Association, American Psychological Association, National Council on Measurement in Education: *Standards for Educational and Psychological Testing*. Washington, D.C.: American Educational Research Association. (ISBN 0-935302-25-5)

Alderson, J. C. (2005): *Diagnosing Foreign Language Proficiency*. London: Continuum.

Alderson, J. C., Clapham, C. and Wall, D. (1995): *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

Alderson, J. C., Figueras, N., Kuijpers, H., Nold. G., Takala, S. and Tardieu, C. (2006): Analysing Tests of Reading and Listening in relation to the CEFR: the experience of the Dutch CEFR Construct Project. *Language Assessment Quarterly* 3 (1): 3–30.

American Educational Research Association (1999): *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Angoff, W. H. (1971): Scales, Norms and Equivalent Scores. In: Thorndike, R. L. (ed.) *Educational Measurement* (2nd Edition), pp. 508–600. Washington, D.C.: American Council on Education.

Beacco, J-C. and Porquier, R. (2008): *Niveau A2 pour le français : Un référentiel*. Paris: Didier.

Beacco, J-C., Porquier, R. and Bouquet, S. (2004): *Niveau B2 pour le français : Un référentiel*. Paris: Didier. (2 vols)

Beacco, J-C., De Ferrari, M., Lhote, G. and Tagliante, C. (2006): *Niveau A1.1 pour le français / référentiel DILF livre*. Paris: Didier.

Beacco, J-C., Porquier, R. and Bouquet, S. (2007): *Niveau A1 pour le français : Un referential*. Paris: Didier.

Berk, R.A. (1986): A Consumer's Guide to Setting Performance Standards on Criterion Referenced Tests. *Review of Educational Research*, 56, 137–172.

Bolton, S., Glaboniat, M., Lorenz, H., Müller, M., Perlmann-Balme, M. and Steiner, S. (2008): *Mündlich: Mündliche Produktion und Interaktion Deutsch: Illustration der Niveaustufen des Gemeinsamen europäischen Referenzrahmens*. Berlin: Langenscheidt.

Breton, Jones, Laplannes, Lepage and North, (forthcoming): Séminaire interlangues / Cross language benchmarking seminar, CIEP Sèvres, 23–25 June 2008: Report. *Strasbourg: Council of Europe.*

Cizek, G. J. (ed.) (2001): *Setting Performance Standards: concepts, methods and perspectives*. Mahwah, NJ: Lawrence Erlbaum.

Cizek, G.J. and Bunch, M.B. (2007): *Standard Setting: a guide to establishing and evaluating performance standards on tests*. Thousand Oaks: Sage.

Cohen, A., Kane, M. and Crooks, T. (1999): A Generalized Examinee-Centered Method for Setting Standards on Achievement Tests. *Applied Measurement in Education*, 12, 343–366.

Council of Europe (2001a): *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Europe (2001b): *Cadre européen commun de référence pour les langues: Apprendre, enseigner, évaluer*. Paris: Didier.

Council of Europe (2002): *Seminar on Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEF), Helsinki, 30 June 30–2 July 2002: Report*. DG IV / EDU / LANG (2002) 15. Strasbourg: Council of Europe.

Council of Europe (2003): *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (*"*CEFR*" DGIV/EDU/LANG (2003) 5. Strasbourg: Council of Europe.

Davidson, F. and Lynch, B. (1993): Criterion-referenced language test development: a prolegomenon. In: Huhta, A., Sajavaara, K. & Takala, S. (eds.), *Language Testing: New Openings*. Jyvaskyla, Finland: University of Jyvaskyla, pp.73–89.

Davidson, F. and Lynch, B. (2002): *Testcraft: A Teacher's Guide to Writing and Using Language Test Specifications*. Yale University Press.

Downing, S. M. and Haladyna, T. M. (eds.) (2006): *Handbook of Test Development.* Earlbaum.

Ebel, R. L. and Frisbee, O. A. (1986): *Essentials of Educational Measurement (4th edition)*. Englewood Cliffs, N.J.: Prentice Hall.

Feldt, L. S., Steffen, M. and Gupta, N. C. (1985): A Comparison of Five Methods for Estimating the Standard Error of Measurement at Specific Score Levels. *Applied Psychological Measurement,* 9, 351–361.

Ferrara, S., Perie, M. and Johnson, E. (2002): *Matching the Judgmental Task with Standard Setting Panelist Expertise: the item-descriptor (ID) matching procedure.* Washington DC: American Institutes for Research.
Fienberg, S. E. (1977): *The Analysis of Cross-classified Categorical Data.* Cambridge, Massachusetts: The MIT Press.
Fienberg, S.E., Bishop, Y. M. M. and Holland, P. W. (1975): *Discrete Multivariate Analysis*. Cambridge (Massachusetts): The MIT Press.
Glaboniat, M., Müller, M., Schmitz, H., Rusch, P., Wertenschlag, L., (2002/5): *Profile Deutsch*. Berlin: Langenscheidt, ISBN 3-468-49463-7.
Instituto Cervantes (2007): *Niveles de Referencia para el español, Plan Curricular del Instituto Cervantes*. Madrid: Biblioteca Nueva.
Jaeger, R. M. (1991): Selection of Judges for Standard-setting. *Educational Measurement: Issues and Practice*, 10, 3–6.
Kaftandjieva, F. (2007): Quantifying the Quality of Linkage between Language Examinations and the CEF. In Carlsen, C. and Moe, E. (eds.) *A Human Touch to Language Testing*. Oslo: Novus Press, 34–42.
Keats, J. A. (1957): Estimation of Error Variances of Test Scores. *Psychometrika* 22, 29–41.
Kingston, N. M., Kahl, S. R., Sweeny, K. P. and Bay, L. (2001): Setting Performance Standards using the Body of Work Method. In Cizek G. J. (ed.), *Setting Performance Standards: Concepts, methods and perspectives.* Mahwah, NJ: Erlbaum, pp. 219–248.
Kolen, M. L. and Brennan, R-L. (2004): *Test Equating, Scaling and Linking*. New York: Springer.
Lepage, S. and North, B. (2005): *Guide for the organisation of a seminar to calibrate examples of spoken performance in line with the scales of the Common European Framework of Reference for Languages*. Strasbourg: Council of Europe DGIV/EDU/LANG (2005) 4.
Linacre, J. M. (1989): *Multi-faceted Measurement*. Chicago: MESA Press.
Linacre, J. M. (2008): *A User's Guide to FACETS. Rasch Model Computer Program*. ISBN 0-941938-03-4. www.winsteps.com.
Livingston, S. A. and Lewis, C. (1995): Estimating the Consistency and Accuracy of Classification based on Test Scores. *Journal of Educational Measurement*, 32, 179–197.
Lord, F. (1965): A Strong True-score Theory, with Applications. *Psychometrika*, 30, 239–270.
Lynch, B. and Davidson, F. (1994): Criterion-referenced language test development: linking curricula, teachers and tests. *TESOL Quarterly* 28:4, pp. 727–743.
Lynch, B. and Davidson, F. (1998): Criterion Referencing. In: Clapham, C. & Dorson, D. (eds.) *Language Testing and Assessment*, Volume 7, Encyclopedia of Language and Education. Dordrecht: Kluwer Academic Publishers, pp. 263–273.
Milanovic, M. (2002): *Language Examining and Test Development*. Strasbourg: Language Policy Division, Council of Europe.
Mitzel, H. C., Lewis, D. M., Patz, R. J. & Green, D. R. (2001): The Bookmark Procedure: psychological perspectives. In Cizek G. J. (ed.) *Setting Performance Standards: concepts, methods and perspectives* (pp. 249-281). Mahwah, NJ: Erlbaum.
Norcini, J., Lipner, R., Langdon, L., and Strecker, C. (1987): A Comparison of Three Variations on a Standard-Setting Method. *Journal of Educational Measurement*, 24, 56–64.
North, B. (2000a): *The Development of a Common Framework Scale of Language Proficiency*. New York: Peter Lang.
North, B. (2000b): Linking Language Assessments: an example in a low-stakes context. *System* 28, 555–577.
North, B. and Schneider, G. (1998): Scaling descriptors for language proficiency scales. *Language Testing* 15/2: 217–262.
OECD (2005): *Pisa 2003 Technical Report*. Paris: OECD.
Parizzi, F. and Spinelli, B. (forthcoming): *Profilo della Lingua Italiana*, Firenze: La Nuova Italia.
Plake, B. S. (2008): Standard Setters: Stand Up and Take a Stand! *Educational Measurement: Issues and Practice* 27/1: 3–9.
Reckase, M. D. (2006a): A Conceptual Framework for a Psychometric Theory for Standard Setting with Examples of Its Use for Evaluating the Functioning of Two Standard Setting Methods. *Educational Measurement: Issues and Practice*, 2006, 25(2), 4–18.
Reckase, M. D. (2006b): Rejoinder: Evaluating Standard Setting Methods Using Error Models Proposed by Schulz. *Educational Measurement: Issues and Practice*, 2006, 25 (3), 14–17.

Schneider, G. and North, B. (2000): *Fremdsprachen können – was heisst das? Skalen zur Beschreibung, Beurteilung und Selbsteinschätzung der fremdsprachlichen Kommunikationsfähigkeit*. Chur/Zürich: Ruegger Verlag.
Siegel, S. and Castellan, N. J. (1988): *Non-parametric statistics for the behavioral sciences*. New York: McGraw-Hill.
Subkoviak, M. J. (1988): A Practitioner's Guide to Computation and Interpretation of Reliability for Mastery Tests. *Journal of Educational Measurement,* 13, 265–276.
Thorndike, R.L. (ed.) (1971): *Educational Measurement* (2nd Edition), pp. 508–600. Washington, D.C.: American Council on Education.
Van der Schoot, F. (2001): *Standaarden voor Kerndoelen Basisonderwijs* [Standards for Primary Objectives in Primary Education]. PhD thesis. Arnhem: Cito.
van Ek, Jan A. (1976): *The Threshold level in a European Unit/credit System for Modern Language Learning by Adults*. Strasbourg: Council of Europe.
van Ek, J. A. and Trim, J. L. M., (2001a): *Waystage*. Cambridge: CUP, ISBN 0-521-56707-6
van Ek, J. A. and Trim, J. L. M., (2001b): *Threshold 1990*. Cambridge: CUP, ISBN 0-521-56707-8
van Ek, J. A. and Trim, J. L. M., (2001c): *Vantage*. Cambridge: CUP, ISBN 0-521-56705-X
Verhelst, N. D. and Verstralen, H. H. F. M. (2008): Some Considerations on the Partial Credit Model. *Psicológica,* 29, 229–254.
Weir, C. (1993): *Understanding and Developing Language Tests*. Hemel Hempstead UK: Prentice Hall.

Annexes

Annexe A. Fiches et échelles pour la description et la spécification. Chapitre 1 et 4

Partie A1 : Caractéristiques principales des niveaux du CECRL (chapitre 1)

Partie A2 : Fiches pour la description des examens (chapitre 4)

Partie A3 : Spécification : activités langagières communicatives (chapitre 4)

Partie A4 : Spécification : compétence langagière communicative (chapitre 4)

Partie A5 : Résultat des analyses (chapitre 4)

Annexe B. Grilles d'analyse de contenu Chapitre 4

Partie B1 : Grille d'analyse de contenu pour la réception orale et la réception écrite
Partie B2 : Grille d'analyse de contenu pour la production orale et la production écrite

Annexe C. Fiches et échelles pour la standardisation et le calibrage (chapitre 5)

Partie A1 : Caractéristiques principales des niveaux du CECRL Chapitre 1

Niveau		Tableau A1. Caractéristiques principales : Interaction et production (CECR partie 3.6, simplifiée)
Utilisateur expérimenté		On ne saurait trop insister sur le fait qu'au Niveau C2 on n'a pas l'ambition d'égaler la compétence du locuteur natif ou presque. La recherche initiale autant qu'un projet utilisant les descripteurs du CECR pour évaluer la compétence en langue maternelle (North 2002 : CECRL Etudes de cas) ont montré l'existence de locuteurs bilingues bien au-dessus du niveau le plus élevé défini (C2). Wilkins a identifié un septième niveau de « Compétence ambilingue » dans sa proposition de 1978 pour une échelle européenne d'unités de crédits.
	C2	**Le Niveau C2** a pour but de caractériser le degré de **précision, d'adéquation et d'aisance de la langue** que l'on trouve dans le discours de ceux qui ont été des apprenants de haut niveau. Les descripteurs inventoriés ici comprennent *: transmettre les subtilités de sens avec précision en utilisant, avec une raisonnable exactitude, une gamme étendue de modalisateurs ; avoir une bonne maîtrise des expressions idiomatiques et familières accompagnée de la conscience des connotations ; revenir en arrière et reformuler une difficulté sans heurts de sorte que l'interlocuteur s'en aperçoive à peine*.
	C1	Le **Niveau C1** semble être caractérisé par le bon accès à **une large gamme de discours qui permet une communication aisée et spontanée** comme on le verra dans les exemples suivants : *peut s'exprimer avec aisance et spontanéité presque sans effort. A une bonne maîtrise d'un répertoire lexical large dont les lacunes sont facilement comblées par des périphrases. Il y a peu de recherche notable de certaines expression ou de stratégies d'évitement ; seul un sujet conceptuellement difficile peut empêcher que le discours ne se déroule naturellement*. Les capacités discursives qui caractérisent le niveau précédent se retrouvent au Niveau C1 avec encore plus d'aisance, par exemple *: peut choisir une expression adéquate dans un répertoire disponible de fonctions du discours pour introduire ses commentaires afin de mobiliser l'attention de l'auditoire ou de gagner du temps en gardant cette attention pendant qu'il/elle réfléchit ; produit un discours clair, bien construit et sans hésitation qui montre l'utilisation bien maîtrisée des structures, des connecteurs et des articulateurs.*
Utilisateur indépendant	**B2+**	B2+ correspond à une performance B2 confirmée. L'accent y est mis sur l'argumentation, et la conscience de la langue qui apparaît en B2 se poursuit ici. Néanmoins, on peut aussi interpréter l'accent mis sur l'argumentation et le discours social comme une importance nouvelle accordée aux capacités discursives. Ce nouveau degré de compétence discursive apparaît dans la gestion de la conversation (stratégies de coopération) : *est capable de donner un retour d'informations et une suite aux déclarations et aux déductions des autres locuteurs et, ce faisant, de faciliter l'évolution de la discussion ; de mettre en relation adroitement sa propre contribution et celle des autres locuteurs*. Il apparaît également dans la relation logique/cohésion : *utilise une variété de mots de liaison efficacement pour indiquer le lien entre les idées ; soutient systématiquement une argumentation qui met en valeur les points significatifs et les points secondaires pertinents.* Enfin, c'est à ce niveau que se concentrent les descripteurs portant sur la négociation.
	B2	Le **Niveau B2** marque une coupure importante avec ceux qui les précèdent. Par exemple, ce degré se concentre sur **l'efficacité de l'argumentation** *: rend compte de ses opinions et les défend au cours d'une discussion en apportant des explications appropriées ; des arguments et des commentaires ; développe un point de vue en soutenant tour à tour les avantages et les inconvénients des différentes options ; développe une argumentation en défendant ou en critiquant un point de vue donné ; prend une part active dans une discussion informelle dans un contexte familier ; fait des commentaires, exprime clairement son point de vue, évalue les choix possibles, fait des hypothèses et y répond.* En second lieu, à ce niveau, on est capable de bien se débrouiller dans le **discours social**, par exemple : *comprendre dans le détail ce que l'on vous dit dans une langue standard courante même dans un environnement bruyant ; prendre l'initiative de la parole, prendre son tour de parole au moment voulu et clore la conversation lorsqu'il faut, même si cela n'est pas toujours fait avec élégance ; intervenir avec un niveau d'aisance et de spontanéité qui rend possibles les échanges avec les locuteurs natifs sans imposer de contrainte à l'une ou l'autre des parties*. Enfin, ce niveau se caractérise par une **conscience de la langue** : *corriger les fautes qui ont débouché sur des malentendus ; prendre note des « fautes préférées » et contrôler consciemment le discours pour les traquer. En règle générale, corriger les fautes et les erreurs aussitôt qu'on en prend conscience.*
	B1+	B1+ correspond à une performance B1 confirmée**.** On y retrouve les deux mêmes traits caractéristiques auxquels s'ajoute un certain nombre de descripteurs qui se concentrent sur la quantité d'information échangée, par exemple : *apporte l'information concrète exigée dans un entretien ou une consultation (par exemple, décrit des symptômes à un médecin) mais avec une précision limitée ; explique pourquoi quelque chose pose problème ; donne son opinion sur une nouvelle, un article, un exposé, une discussion, un entretien, un documentaire et répond à des questions de détail complémentaires – les résume ; mène à bien un entretien préparé en vérifiant et confirmant l'information même s'il doit parfois faire répéter l'interlocuteur dans le cas où sa réponse est longue ou rapidement énoncée ; décrit comment faire quelque chose et donne des instructions détaillées ; échange avec une certaine assurance une grande quantité d'informations factuelles sur des questions habituelles ou non dans son domaine.*
	B1	Le **Niveau B1** correspond aux spécifications du **Niveau seuil**. Deux traits le caractérisent particulièrement. Le premier est la **capacité à poursuivre une interaction et à obtenir ce que l'on veut**, par exemple : *en règle générale, suit les points principaux d'une discussion assez longue à son sujet, à condition que la diction soit claire et la langue standard ; reste compréhensible même si la recherche des mots et des formes grammaticales ainsi que la remédiation sont évidentes, notamment au cours de longs énoncés*. Le deuxième trait est la **capacité de faire face habilement aux problèmes de la vie quotidienne**, par exemple : *se débrouiller dans une situation imprévue dans les transports en commun ; faire face à l'essentiel de ce qui peut arriver chez un voyagiste ou au cours du voyage ; intervenir sans préparation dans des conversations sur des sujets familiers*.

Utilisateur élémentaire	A2 +	Ce niveau A2+ correspond à une performance A2 confirmée avec une participation dans une conversation plus active, encore que limitée et nécessitant une aide, par exemple *: comprend assez bien pour se débrouiller dans des échanges simples et courants sans effort excessif ; se fait comprendre pour échanger des idées et des informations sur des sujets familiers dans des situations quotidiennes prévisibles à condition que l'interlocuteur aide, le cas échéant ; se débrouille dans les situations quotidiennes dont le contenu est prévisible bien qu'en devant adapter le message et chercher ses mots ;* de manière plus significative, une meilleure capacité à poursuivre un monologue, par exemple, *exprime ses impressions en termes simples ; fait une longue description des données quotidiennes de son environnement comme les gens, les lieux, une expérience professionnelle ou académique ; décrit des activités passées et des expériences personnelles ; décrit des occupations quotidiennes et des habitudes ; décrit des projets et leur organisation ; explique ce qu'il/elle aime ou n'aime pas.*
	A2	C'est au niveau A2 que l'on trouvera la plupart des descripteurs qui indiquent les rapports sociaux tels que *: utilise les formes quotidiennes de politesse et d'adresse ; accueille quelqu'un, lui demande de ses nouvelles et réagit à la réponse ; invite et répond à une invitation ; discute de ce qu'il veut faire, où, et fait les arrangements nécessaires ; fait une proposition et en accepte une*. C'est ici que l'on trouvera également les descripteurs relatifs aux sorties et aux déplacements: *mener à bien un échange simple dans un magasin, un bureau de poste ou une banque ; se renseigner sur un voyage ; utiliser les transports en commun : bus, trains et taxis, demander des informations de base, demander son chemin et l'indiquer, acheter des billets ; fournir les produits et les services nécessaires au quotidien et les demander*.
	A1	Le Niveau A1 est le niveau le plus élémentaire d'utilisation de la langue à titre personnel – celui où l'apprenant *est capable d'interactions simples ; peut répondre à des questions simples sur lui-même, l'endroit où il vit, les gens qu'il connaît et les choses qu'il a et en poser ; peut intervenir avec des énoncés simples dans les domaines qui le concernent ou qui lui sont familiers et y répondre également* en ne se contentant pas de répéter des expressions toutes faites et pré-organisées.

Tableau A2. Caractéristiques principales : Réception					
	Les thèmes	L'action	Ce qui est compris	Le support	Les limitations
C1	Les thèmes abstraits et complexes de la vie sociale, professionnelle ou du monde de l'éducation, en rapport ou non avec son domaine ou sa spécialité	Suit, peut être avec un peu de difficulté		Les films faisant un usage important de l'argot et d'expressions idiomatiques Les annonces publiques de mauvaise qualité dont le son est déformé	Peut avoir besoin par moments de : confirmer des détails (à l'aide d'un dictionnaire ou du locuteur) s'il ne s'agit pas de son domaine Relire des parties difficiles
		Comprend	Les points de détail fins Les opinions implicites et explicites Une gamme étendue d'expressions idiomatiques et de tournures courantes Les changements de registres Les comportements et les relations implicites	Différents types de textes longs et complexes De longs discours – conférences, discussions, débats – même quand ils sont mal structurés Les interactions et les débats complexes avec des intervenants extérieurs Une gamme étendue de textes enregistrés ou radiodiffusés, même ce ne sont pas des textes standards Tout type de correspondance	
B2+	Une gamme étendue de thèmes familiers ou non de la vie sociale, professionnelle ou du monde de l'éducation	Suit, peut être avec un peu de difficulté		Une conversation animée entre locuteurs natifs	Une langue standard, non idiomatique Des structures de discours appropriées Un faible bruit de fond A parfois besoin de confirmer des détails (à l'aide d'un dictionnaire ou du locuteur) - s'ils ne sont pas de son domaine si les conditions énoncées ci-dessus ne sont pas réunies
		Comprend		La langue parlée, les émissions en direct Les textes spécialisés (hautement spécialisés dans le domaine)	

B2	• Les thèmes assez familiers, concrets et abstraits en rapport avec son centre d'intérêt ou sa spécialité	• Suit, peut être avec un peu de difficulté	• La plupart de ce qui est dit	Les discussions sur lui/elle par des locuteurs natifs	• Une langue standard • Des repères et des indications avec des marqueurs explicites • Si des locuteurs natifs parlant ensemble modifient leur façon de parler
		• Parcourt rapidement	• Ce qui est pertinent • Si une étude plus approfondie vaut la peine • Les détails spécifiques	• Les textes longs et complexes • Les actualités, des articles, des reportages	
		• Comprend avec une autonomie assez grande	• Les idées principales • La ou les significations essentielles • Les raisonnements complexes • L'état d'esprit, le ton du locuteur/de l'auteur de l'écrit	• Les longs discours : conférences, conversations, présentations, comptes rendus, discussions • Les textes complexes à la fois d'un point de vue du genre et d'un point de vue linguistique • Les discussions techniques ; des instructions longues et complexes ; des détails sur les conditions et des avertissements • La plupart des programmes télévisés sur des événements actuels • La plupart des documentaires télévisés, des interviews, des émissions-débats, des supports très spécialisés • Les annonces et les messages • La plupart des documentaires radiophoniques, des matériels enregistrés • La correspondance	• S'il ou si elle peut relire des passages difficiles
B1+	• Les thèmes d'usage courant ou en rapport avec le domaine professionnel • Les thèmes en rapport avec son domaine d'intérêt personnel	• Suit, mais pas toujours le détail	• Le raisonnement pour résoudre un problème	• Un texte argumentatif	• Une langue standard (accent familier) et simple • Des repères avec des marqueurs explicites et des indications
		• Parcourt	• L'information recherchée	• Les textes plus longs • Les textes différents, différentes parties d'un texte	
		▪ Comprend	• Les informations factuelles et claires • Les messages d'ordre général • Les conclusions principales • Les détails spécifiques	• Les textes argumentatifs • Les conférences et les conversations dans son domaine • Une grande partie des programmes de télévision, des interviews, des conférences courtes, des reportages d'actualité • La plupart des documentaires radiophoniques et des textes enregistrés	

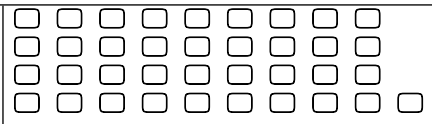
B1	• Les thèmes familiers traités habituellement dans le domaine de l'éducation, du travail ou des loisirs • Les thèmes en rapport avec son domaine d'intérêt personnel	• Suit, mais pas toujours le détail	Les points essentiels	• Les discours longs sur lui/elle • Beaucoup de films dans lesquels l'image et l'action aident à comprendre • Les programmes TV : interviews, courtes conférences, actualités, reportages • Les articles de journaux, simples éclairs	• Une langue - claire - standard - simple • Un débit assez lent
		• Comprend de manière satisfaisante	• Les points importants • Les informations pertinentes	• Les textes factuels clairs • Les récits courts • Les descriptions d'événements, de sentiments et de souhaits • Les indications détaillées • Les conversations courtes • Les bulletins d'information à la radio ou des documents enregistrés plus simples • Les écrits quotidiens : lettres, brochures, de courts documents officiels • Les renseignements techniques simples : par exemple des modes d'emploi	
A2+	• Les thèmes familiers et concrets	• Identifie	• Les points importants	• Les émissions télévisées d'actualités sur des événements, des accidents, etc., …où l'image accompagne le commentaire	
		• Comprend assez pour satisfaire ses besoins		• Les types simples de lettres et de télécopie standard (demandes d'information, commandes, confirmations) • Les textes courts utilisant une langue plus simple, et d'usage très courant et quotidien et liée au domaine professionnel • Les règlements, par exemple sur la sécurité	• L'expression doit se faire dans une langue simple
A2	• Les faits quotidiens prévisibles • Les lieux de première priorité : très personnel, famille, achats, voisinage, travail	• Identifie	• Les informations spécifiques, prévisibles • Les thèmes de discussion • Les changements de thèmes • Une idée du contenu	• Les documents plus simples quotidiens : prospectus, menus, inventaires, horaires, brochures, lettres • Les discussions sur lui/elle • Les articles de journaux courts décrivant des événements • Les émissions d'actualités télévisées factuelles	• Une articulation claire et lente

		• Comprend	• Le point important • L'information essentielle	• Les textes courts et simples comprenant un lexique utilisé très fréquemment dont une partie faisant partie du lexique international • Les indications simples indiquant comment aller de A à B • Les messages, annonces et passages enregistrés simples et clairs • Les instructions simples concernant un appareil d'usage courant (ex : le téléphone) • Les lettres personnelles simples et courtes • Les panneaux et des affiches de la vie quotidienne : indications, instructions, risques	
A1	• Les situations les plus communes de la vie quotidienne	• Identifie	• Les mots, phrases, noms familiers • Une idée du contenu	• Les informations simples • Les textes d'information plus simples	• Une articulation très lente, réalisée avec beaucoup d'attention, avec de longues pauses permettant d'assimiler le sens du message • Les noms familiers, des mots et des phrases simples • La possibilité de relire /de faire répéter
		• Comprend	(Idée générale)	• Les textes très simples et courts avec un support visuel, une seule phrase à la fois : - messages sur une carte postale - itinéraires - descriptions	

Partie A2 : Fiches pour la description des examens (chapitre 4)

Fiche A1 : Description générale de l'examen

DESCRIPTION GENERALE DE L'EXAMEN			
1. Informations générales			
Intitulé de l'examen			
Langue évaluée			
Organisme examinateur			
Versions de l'examen analysées (date)			
Type d'examen	□ international □ national □ régional □ propre à l'institution		
Objectifs			
Population cible	□ 1er cycle du sec. □ 2ème cycle du sec. □ enseignement supérieur □ adultes		
Nombre de candidats par an			
2. Quel est l'objectif général ?			
3. Quels sont les objectifs plus particuliers ? Si possible, décrivez les besoins des utilisateurs potentiels à qui cet examen est destiné.			
4. Quels sont les principaux domaines traités ?	□ public □ personnel □ professionnel □ éducationnel		
5. Quelles sont les activités communicatives testées ?	□ 1 réception orale □ 2 réception écrite □ 3 interaction orale □ 4 interaction écrite □ 5 production orale □ 6 production écrite □ 7 capacités intégrées □ 9 médiation orale □ 10 médiation écrite □ 11 Utilisation de la langue □ 12 autres (préciser) :……. ……	Intitulé de l'épreuve …… …… …….. …….. …….. …… …… …… …… …… …… …… …… …… …… ……	Durée
6. Quel est le poids des différentes épreuves par rapport au résultat global ?			
7. Faites une courte description de la structure de chaque épreuve.			
8. Quels sont les types de réponses attendues ?	□ choix multiple □ vrai/faux □ appariement □ classement □ phrases lacunaires □ phrase à compléter □ texte lacunaire fermé / test de closure fermé □ texte lacunaire ouvert /test de closure ouvert	Epreuve utilisée in (Indiquez le numéro) □	

	□ question ouverte à réponse courte □ réponse développée (texte, monologue) □ interaction avec l'examinateur □ interaction avec des pairs □ autre	
9. Quels sont les renseignements fournis aux candidats et aux enseignants ?	□ objectif général □ domaine(s) principaux □ épreuves □ tâches □ exemples d'épreuves □ vidéo illustrant l'oral	□ exemples de feuilles de réponses □ critères de correction □ barème de notation □ échantillons de performances standards du niveau de réussite □ fac-similé de diplôme
10. Où peut-on les trouver ?	□ sur le site internet □ dans des librairies □ dans les centres d'examens □ à la demande de l'institution □ autre	
11. Sous quelle forme les résultats sont-ils délivrés ?	□ note globale □ note par épreuve	□ note globale plus profil graphique □ profil par épreuve

Fiche A2 : Conception de l'examen

Elaboration de l'examen	**Brève description et/ou références**
1. Quel organisme a décidé de la nécessité de cet examen ?	❑ L'institution ❑ Un institut culturel ❑ Le ministère de l'Education ❑ Le ministère de la Justice ❑ Autre (préciser) :
2. Si un organisme extérieur est impliqué, quelle est son influence sur la conception et l'élaboration ?	□ définit les objectifs généraux □ fixe le niveau de compétence en langue □ fixe le domaine et le contenu de l'examen □ fixe le format de l'examen et le type de tâches □ autre (préciser) :
3. S'il n'y avait pas d'implication d'un organisme extérieur, quels sont les autres paramètres qui ont influencé la conception et l'élaboration de l'examen ?	□ une analyse de besoins □ une description interne des objectifs de l'examen □ une description interne du niveau de langue □ un référentiel ou un programme □ le profil des candidats
4. Lors de l'élaboration des épreuves a-t-on tenu compte des différents profils des candidats ?	□ origine linguistique (L1) □ acquis linguistiques antérieurs □ âge □ niveau d'instruction □ milieu socio-économique □ facteurs socioculturels □ origine ethnique □ sexe
5. Qui rédige les items ou élabore les tâches du test?	
6. Les rédacteurs d'épreuves bénéficient-ils de conseils pour en garantir la qualité ?	□ formation □ lignes directrices □ listes de contrôle □ exemples de tâches valides, fiables et appropriées □ descriptions calibrées sur les niveaux du cadre □ descriptions calibrées sur d'autres niveaux

7. Donne-t-on une formation aux rédacteurs ?	□ oui □ non
8. Les épreuves font-elles l'objet d'une discussion avant leur utilisation ?	□ oui □ non

9. Si oui, qui y participe ?	□ les collègues, individuellement □ un groupe interne de discussion □ une commission externe d'examen □ des personnes impliquées, en interne □ des personnes impliquées, en externe
10. Les épreuves sont-elles prétestées ?	□ oui □ non
11. Si oui, comment ?	
12. Sinon, pourquoi ?	
13. La fiabilité du test est-elle évaluée ?	□ oui □ non
14. Si oui, comment ?	□ recueil de données et mesures psychométriques □ autre (préciser) :
15. Les différents aspects de la validité sont-ils évalués ?	□ validité apparente □ validité de contenu □ validité convergente □ validité prédictive □ validité de construct
16. Si oui, décrivez de quelle façon	

Fiche A3 : Correction

Correction : épreuve de _____	Remplir un exemplaire de cette fiche pour chaque épreuve. **Brève description et/ou référence**
1. Comment les tâches sont-elles corrigées ?	Tâches concernant la réception : □ lecteur optique □ examinateur Tâches concernant la production ou tâches intégrées : □ examinateur formé □ enseignants
2. Où sont corrigées les tâches ?	□ par un organisme central □ localement : □ par des équipes locales □ par des correcteurs individuels
3. Quels sont les critères de sélection des correcteurs ?	
4. Comment l'exactitude de la notation est-elle recherchée ?	□ contrôles réguliers du coordinateur □ formation des correcteurs/examinateurs □ sessions de formation à la standardisation des évaluations □ utilisation d'exemples standards de tâches : □ calibrées par rapport au CECR □ calibrées par rapport à une autre description de niveaux □ non calibrées par rapport au CECR ou à toute autre description
5. Décrire les spécifications des critères de notation des épreuves de production ou des épreuves intégrées.	□ note globale pour chaque tâche □ notes pour différents aspects de chaque tâche □ échelle de notation pour la performance globale □ grille de notation pour différents aspects de la performance □ échelle de notation pour chaque tâche □ grille de notation pour différents aspects de chaque épreuve □ échelle de notation par niveau, sans lien avec le CECRL □ échelle de notation par niveau en liaison avec le CECR L
6. Les épreuves intégrées ou de production font-elles ou non l'objet d'une double correction ?	□ correction simple □ deux correcteurs simultanément □ double correction des copies d'écrits / des enregistrements

	des productions orales □ autre (préciser) :
7. S'il y a double correction, que fait-on en cas de désaccord entre les correcteurs ?	□ appel à un troisième correcteur dont la note sera celle qui sera gardée □ appel à un troisième correcteur et choix des deux notes les plus proches □ moyenne des deux notes □ consensus entre deux correcteurs après discussion □ autre (préciser) :
8. L'accord inter- correcteurs est-il mesuré ?	□ oui □ non
9. L'accord intra-correcteur est-il mesuré ?	□ oui □ non

Fiche A4 : Notation

Notation : Epreuve de _____	**Remplir un exemplaire de cette fiche pour chaque épreuve.** **Brève description et/ou référence**
1. Y a t-il une note d'admissibilité ou un niveau ?	□ admissibilité □ mentions
2. Décrivez la démarche suivie pour définir les notes d'admissibilité, les niveaux et les points de césure	
3. Si l'on ne rend compte que de l'échec ou de la réussite, comment en définit-on les points de césure ?	
4. S'il y a des niveaux, comment définit-on leur seuil ?	
5. Comment assure-t-on la cohérence de ces normes ?	

Fiche A5 : Communication des résultats

Résultats	**Brève description et/ou référence**
1. Sous quelle forme se présentent les résultats délivrés aux candidats ?	□ note globale ou échec/réussite □ note ou échec/réussite pour chaque épreuve □ note globale plus profil graphique par épreuve □ profil des performances pour chaque épreuve
2. Sous quelle forme rend-on compte des résultats ?	□ note brute □ niveaux indéterminés (par exemple « C ») □ niveau sur une échelle donnée □ profils diagnostiques
3. Sur quel type de document sont indiqués les résultats ?	□ lettre ou courriel □ rapport □ certificat/diplôme
4. Donne-t-on des renseignements aux candidats pour les aider à interpréter les résultats ? Donnez des détails.	
5. Les candidats ont-ils le droit de voir leurs copies corrigées et notées ?	
6. Les candidats ont-ils le droit de demander une nouvelle correction ?	

Fiche 6 : Analyse et révision de l'examen

Analyse de l'examen et révision après passation	**Brève description et/ou référence**
1. Recueille-t-on des retours d'information sur l'examen ?	□ oui □ non
2. Si oui, par qui ?	□ experts internes (collègues) □ experts externes □ organismes locaux d'évaluation □ personnel administrant l'examen

	□ enseignants □ candidats
3. Tient-on compte du retour d'information pour les versions révisées de l'examen ?	□ oui □ non
4. Recueille-t-on des données pour procéder à des analyses sur les examens ?	□ sur tous les examens □ sur un échantillon de candidats : combien :_____ ; combien de fois :_____ □ non
5. Si oui, dites comment on recueille les données.	□ pendant les pré-tests □ pendant la passation de l'examen □ après la passation
6. Pour quelles caractéristiques fait-on une analyse des données recueillies ?	□ la difficulté □ la discrimination □ la fiabilité □ la validité
7. Dites quelles méthodes analytiques ont été mises en œuvre (par exemple, en termes de procédures psychométriques).	
8. Analyse-t-on les performances de candidats appartenant à des groupes différents ? Si oui, dites comment.	
9. Décrivez les moyens mis en œuvre pour garantir la confidentialité des données.	
10. Les concepts de mesure appropriés sont-ils expliqués aux utilisateurs du test ? Si oui, dites comment.	

Fiche A7 : Justification des décisions

Justification des décisions prises	**Brève description et/ou référence**
Justifiez les décisions prises relatives à l'examen ou aux tâches en question. Un cycle de révision de l'examen est-il mis en place ? Par qui ? Quelles procédures pour revoir les décisions ?	

Fiche A8 : Impression initiale du niveau global

Impression initiale du niveau global de l'examen par rapport au CECRL		
□ **A1**	□ **B1**	□ **C1**
□	□	□
□ **A2**	□ **B2**	□ **C2**
□	□	□
Brève justification, référence de la documentation		

Partie A3 : Spécification : activités langagières communicatives (chapitre 4)

A3.1 Réception

Réception orale

Fiche A9 : Réception orale

Réception orale	Brève description et/ou référence
1. Dans quels contextes (domaines, situations…) attend-on des candidats qu'ils prouvent leur compétence ? ➢ Le Tableau 5 dans le CECRL 4.1 peut servir de référence.	
2. Quels sont les thèmes de communication que les candidats doivent être capables de traiter ? ➢ Les listes du CECRL 4.2 peuvent servir de référence.	
3. Quels types de tâches, d'activités communicatives et quelles stratégies les candidats doivent-ils être capables de mettre en œuvre ? ➢ Les listes du CECRL 4.3,4.4.2.1,7.1,7.2 et 7.3 peuvent servir de référence.	
4. Quels types et quelle longueur de textes attend-on que les candidats soient capables de traiter ? ➢ Les listes du CECRL 4.6.2 et 4.6.3 peuvent servir de référence.	
5. Après avoir pris connaissance de l'échelle de Compréhension générale de l'oral reproduite ci-dessous, dites et justifiez à quel(s) niveau(x) de l'échelle l'épreuve devrait se situer. ➢ Les sous échelles de réception orale du CECRL 4.4.2.1 énumérées à la suite de l'échelle peuvent servir de référence.	**Niveau :** **Justification (y compris références documentaires)**

	COMPRÉHENSION GÉNÉRALE DE L'ORAL
C2	Peut comprendre toute forme de langue orale qu'elle soit en direct ou à la radio et quel qu'en soit le débit.
C1	Peut suivre une intervention d'une certaine longueur sur des sujets abstraits ou complexes même hors de son domaine mais peut avoir besoin de faire confirmer quelques détails, notamment si l'accent n'est pas familier. Peut reconnaître une gamme étendue d'expressions idiomatiques et de tournures courantes en relevant les changements de registre. Peut suivre une intervention d'une certaine longueur même si elle n'est pas clairement structurée et même si les relations entre les idées sont seulement implicites et non explicitement indiquées.
B2	Peut comprendre une langue orale standard en direct ou à la radio sur des sujets familiers et non familiers se rencontrant normalement dans la vie personnelle, sociale, universitaire ou professionnelle. Seul un très fort bruit de fond, une structure inadaptée du discours ou l'utilisation d'expressions idiomatiques peuvent influencer la capacité à comprendre.
	Peut comprendre les idées principales d'interventions complexes du point de vue du fond et de la forme, sur un sujet concret ou abstrait et dans une langue standard, y compris des discussions techniques dans son domaine de spécialisation. Peut suivre une intervention d'une certaine longueur et une argumentation complexe à condition que le sujet soit assez familier et que le plan général de l'exposé soit indiqué par des marqueurs explicites.
B1	Peut comprendre une information factuelle directe sur des sujets de la vie quotidienne ou relatifs au travail en reconnaissant les messages généraux et les points de détail, à condition que l'articulation soit claire et l'accent courant.
	Peut comprendre les points principaux d'une intervention sur des sujets familiers rencontrés régulièrement au travail, à l'école, pendant les loisirs, y compris des récits courts.

A2	Peut comprendre assez pour pouvoir répondre à des besoins concrets à condition que la diction soit claire et le débit lent.
	Peut comprendre des expressions et des mots porteurs de sens relatifs à des domaines de priorité immédiate (par exemple, information personnelle et familiale de base, achats, géographie local, emploi).
A1	Peut comprendre une intervention si elle est lente et soigneusement articulée et comprend de longues pauses qui permettent d'en assimiler le sens.

Sous échelles correspondant à la réception orale	**Français**
➢ Comprendre une interaction entre locuteurs natifs	Page 55
➢ Comprendre en tant qu'auditeur	Page 56
➢ Comprendre des annonces et des instructions orales	Page 56
➢ Comprendre des émissions de radio et des enregistrements	Page 56
➢ Comprendre des émissions de télévision et des films	Page 59
➢ Reconnaître des indices et faire des déductions	Page 60
➢ Prendre des notes	Page 77

Fiche A10 : Réception écrite

Réception écrite	**Brève description et/ou référence**
1. Dans quels contextes (domaines, situations…) attend-on des candidats qu'ils prouvent leur compétence ? ➢ Le Tableau 5 dans le CECRL 4.1 peut servir de référence.	
2. Quels sont les thèmes de communication que les candidats doivent être capables de traiter ? ➢ Les listes du CECRL 4.2 peuvent servir de référence.	
3. Quels types de tâches, d'activités communicatives et quelles stratégies les candidats doivent-ils être capables de mettre en œuvre ? ➢ Les listes du CECRL 4.3, 4.4.2.2, 7.1,7.2 et 7.3 peuvent servir de référence.	
4. Quels types de textes et quelle longueur de texte attend-on que les candidats soient capables de traiter ? ➢ Les listes du CECRL 4.6.2 et 4.6.3 peuvent servir de référence.	
5. Après avoir pris connaissance de l'échelle de Compréhension générale de l'écrit reproduite ci-dessous, dites et justifiez à quel(s) niveau(x) de l'échelle l'épreuve devrait se situer. ➢ Les sous échelles de compréhension de l'écrit du CECRL 4.4.2.2 énumérées à la suite de l'échelle peuvent servir de référence.	**Niveau :** **Justification (y compris références documentaires)**

	COMPRÉHENSION GÉNÉRALE DE L'ÉCRIT
C2	Peut comprendre et interpréter de façon critique presque toute forme d'écrit, y compris des textes (littéraires ou non) abstraits et structurellement complexes ou très riches en expressions familières. Peut comprendre une gamme étendue de textes longs et complexes en appréciant de subtiles distinctions de style et le sens implicite autant qu'explicite.
C1	Peut comprendre dans le détail des textes longs et complexes, qu'ils se rapportent ou non à son domaine, à condition de pouvoir relire les parties difficiles.
B2	Peut lire avec un grand degré d'autonomie en adaptant le mode et la rapidité de lecture à différents textes et objectifs et en utilisant les références convenables de manière sélective. Possède un vocabulaire de lecture large et actif mais pourra avoir des difficultés avec des expressions peu fréquentes.
B1	Peut lire des textes factuels directs sur des sujets relatifs à son domaine et à ses intérêts avec un niveau satisfaisant de compréhension.
A2	Peut comprendre de courts textes simples sur des sujets concrets courants avec une fréquence élevée de langue quotidienne ou relative au travail.
	Peut comprendre des textes courts et simples contenant un vocabulaire extrêmement fréquent, y compris un vocabulaire internationalement partagé.
A1	Peut comprendre des textes très courts et très simples, phrase par phrase, en relevant des noms, des mots familiers et des expressions très élémentaires et en relisant si nécessaire.

Sous échelles correspondant à la réception écrite	**Français**
➢ Comprendre la correspondance	Page 58
➢ Lire pour s'orienter	Page 58
➢ Lire pour s'informer et discuter	Page 58
➢ Lire des instructions	Page 59
➢ Reconnaître des indices et faire des déductions	Page 60
➢ Prendre des notes	Page 77

A3.2 Interaction

Fiche A11 : Interaction orale

Interaction orale	**Brève description et/ou référence**
1. Dans quels contextes (domaines, situations…) attend-on des candidats qu'ils prouvent leur compétence ? ➢ Le Tableau 5 dans le CECRL 4.1 peut servir de référence.	
2. Quels sont les thèmes de communication que les candidats doivent être capables de traiter ? ➢ Les listes du CECRL 4.2 peuvent servir de référence.	
3. Quels types de tâches, d'activités communicatives et quelles stratégies d'interaction les candidats doivent-ils être capables de mettre en œuvre ? ➢ Les listes du CECRL 4.3,4.4.2.,7.1,7.2 et 7.3 peuvent servir de référence.	
4. Quels textes et types de textes attend-on que les candidats soient capables de traiter ? ➢ Les listes du CECRL 4.6.2 et 4.6.3 peuvent servir de référence.	
5. Après avoir pris connaissance de l'échelle de l'Interaction orale générale reproduite ci-dessous, dites et justifiez à quel(s) niveau(x) de l'échelle l'épreuve devrait se situer. ➢ Les sous échelles d'interaction orale du CECRL 4.4.3.1 énumérées à la suite de l'échelle peuvent servir de référence.	**Niveau :** **Justification (y compris références documentaires)**

	INTERACTION ORALE GÉNÉRALE
C2	Possède une bonne maîtrise d'expressions idiomatiques et de tournures courantes, avec une conscience du sens connotatif. Peut exprimer avec précision des nuances fines de signification, en utilisant assez correctement une gamme étendue de modalités. Peut revenir sur une difficulté et la restructurer de manière si habile que l'interlocuteur s'en rende à peine compte.
C1	Peut s'exprimer avec aisance et spontanéité, presque sans effort. Possède une bonne maîtrise d'un vaste répertoire lexical lui permettant de surmonter facilement des lacunes par des périphrases avec apparemment peu de recherche d'expressions ou de stratégies d'évitement. Seul un sujet conceptuellement difficile est susceptible de gêner le flot naturel et fluide du discours.
B2	Peut utiliser la langue avec aisance, correction et efficacité dans une gamme étendue de sujets d'ordre général, éducationnel, professionnel et concernant les loisirs, en indiquant clairement les relations entre les idées. Peut communiquer spontanément avec un bon contrôle grammatical sans donner l'impression d'avoir à restreindre ce qu'il/elle souhaite dire et avec le degré de formalisme adapté à la circonstance.
	Peut communiquer avec un niveau d'aisance et de spontanéité tel qu'une interaction soutenue avec des locuteurs natifs sera tout à fait possible sans entraîner de tension d'une part ni de l'autre. Peut mettre en valeur la signification personnelle de faits et d'expériences, exposer ses opinions et les défendre avec pertinence en fournissant explications et arguments.
B1	Peut communiquer avec une certaine assurance sur des sujets familiers habituels ou non en relation avec ses intérêts et son domaine professionnel. Peut échanger, vérifier et confirmer des informations, faire face à des situations moins courantes et expliquer pourquoi il y a une difficulté. Peut exprimer sa pensée sur un sujet abstrait ou culturel comme un film, des livres, de la musique, etc.
	Peut exploiter avec souplesse une gamme étendue de langue simple pour faire face à la plupart des situations susceptibles de se produire au cours d'un voyage. Peut aborder sans préparation une conversation sur un sujet familier, exprimer des opinions personnelles et échanger de l'information sur des sujets familiers, d'intérêt personnel ou pertinents pour la vie quotidienne (par exemple, la famille, les loisirs, le travail, les voyages et les faits divers).
A2	Peut interagir avec une aisance raisonnable dans des situations bien structurées et de courtes conversations à condition que l'interlocuteur apporte de l'aide le cas échéant. Peut faire face à des échanges courants simples sans effort excessif ; peut poser des questions, répondre à des questions et échanger des idées et des renseignements sur des sujets familiers dans des situations familières prévisibles de la vie quotidienne.
	Peut communiquer dans le cadre d'une tâche simple et courante ne demandant qu'un échange d'information simple et direct sur des sujets familiers relatifs au travail et aux loisirs. Peut gérer des échanges de type social très courts mais est rarement capable de comprendre suffisamment pour alimenter volontairement la conversation.
A1	Peut interagir de façon simple, mais la communication dépend totalement de la répétition avec un débit plus lent, de la reformulation et des corrections. Peut répondre à des questions simples et en poser, réagir à des affirmations simples et en émettre dans le domaine des besoins immédiats ou sur des sujets très familiers.

Sous échelles correspondant à l'interaction orale	**français**
➢ Comprendre un locuteur natif	Page 62
➢ Conversation	Page 62
➢ Discussion informelle	Page 63
➢ Discussions et réunions formelles	Page 64
➢ Coopération à visée fonctionnelle	Page 65
➢ Obtenir des biens et des services	Page 66
➢ Echange d'informations	Page 67
➢ Interviewer et être interviewé	Page 68

Fiche A12 : Interaction écrite

Interaction écrite	Brève description et/ou référence
1. Dans quels contextes (domaines, situations…) attend-on des candidats qu'ils prouvent leur compétence ? ➢ Le Tableau 5 dans le CECRL 4.1 peut servir de référence.	
2. Quels sont les thèmes de communication que les candidats doivent être capables de traiter ? ➢ Les listes du CECRL 4.2 peuvent servir de référence.	
3. Quels types de tâches, d'activités communicatives et quelles stratégies les candidats doivent-ils être capables de mettre en œuvre ? ➢ Les listes du CECRL 4.3, 4.4.2.1,7.1,7.2 et 7.3 peuvent servir de référence.	
4. Quels textes et types de textes attend-on que les candidats soient capables de traiter ? ➢ Les listes du CECRL 4.6.2 et 4.6.3 peuvent servir de référence.	
5. Après avoir pris connaissance de l'échelle de l'Interaction écrite générale reproduite ci-dessous, dites et justifiez à quel(s) niveau(x) de l'échelle l'épreuve devrait se situer. ➢ Les sous échelles d'interaction écrite du CECRL 4.4.3.4 énumérées à la suite de l'échelle peuvent servir de référence.	**Niveau :** **Justification (y compris références documentaires)**

	INTERACTION ÉCRITE GÉNÉRALE
C2	*Comme C1*
C1	Peut s'exprimer avec clarté et précision, en s'adaptant au destinataire avec souplesse et efficacité.
B2	Peut relater des informations et exprimer des points de vue par écrit et s'adapter à ceux des autres.
B1	Peut apporter de l'information sur des sujets abstraits et concrets, contrôler l'information, poser des questions sur un problème ou l'exposer assez précisément.
	Peut écrire des notes et lettres personnelles pour demander ou transmettre des informations d'intérêt immédiat et faire comprendre les points qu'il/elle considère importants.
A2	Peut écrire de brèves notes simples en rapport avec des besoins immédiats.
A1	Peut demander ou transmettre par écrit des renseignements personnels détaillés.

Sous échelles correspondant à interaction écrite	**Français**
➢ Correspondance	Page 69
➢ Notes, messages et formulaires	Page 69

A3.3.Production

Fiche A13 : Production orale

Production orale	Brève description et/ou référence
1. Dans quels contextes (domaines, situations…) attend-on des candidats qu'ils prouvent leur compétence ? ➢ Le Tableau 5 dans le CECRL 4.1 peut servir de référence.	
2. Quels sont les thèmes de communication que les candidats doivent être capables de traiter ? ➢ Les listes du CECRL 4.2 peuvent servir de référence.	
3. Quels types de tâches, d'activités communicatives et quelles stratégies les candidats doivent-ils être capables de mettre en œuvre ? ➢ Les listes du CECRL 4.3, 4.4.2.1,7.1,7.2 et 7.3 peuvent servir de référence.	
4. Quels textes et types de textes attend-on que les candidats soient capables de traiter ? ➢ Les listes du CECRL 4.6.2 et 4.6.3 peuvent servir de référence.	
5. Après avoir pris connaissance de l'échelle de la production orale générale reproduite ci-dessous, dites et justifiez à quel(s) niveau(x) de l'échelle l'épreuve devrait se situer. ➢ Les sous échelles d'interaction écrite du CECRL 4.4.1.1 énumérées à la suite de l'échelle peuvent servir de référence.	**Niveau :** **Justification (y compris références documentaires)**

	PRODUCTION ORALE GÉNÉRALE
C2	Peut produire un discours élaboré, limpide et fluide, avec une structure logique efficace qui aide le destinataire à remarquer les points importants et à s'en souvenir.
C1	Peut faire une présentation ou une description d'un sujet complexe en intégrant des arguments secondaires et en développant des points particuliers pour parvenir à une conclusion appropriée.
B2	Peut méthodiquement développer une présentation ou une description soulignant les points importants et les détails pertinents.
	Peut faire une description et une présentation détaillées sur une gamme étendue de sujets relatifs à son domaine d'intérêt en développant et justifiant les idées par des points secondaires et des exemples pertinents.
B1	Peut assez aisément mener à bien une description directe et non compliquée de sujets variés dans son domaine en la présentant comme une succession linéaire de points.
A2	Peut décrire ou présenter simplement des gens, des conditions de vie, des activités quotidiennes, ce qu'on aime ou pas, par de courtes séries d'expressions ou de phrases non articulées.
A1	Peut produire des expressions simples isolées sur les gens et les choses.

Sous échelles correspondant à la production orale	**Français**
➢ Monologue suivi : décrire l'expérience	Page 49
➢ Monologue suivi : argumenter (par exemple dans un débat)	Page 50
➢ Annonces publiques	Page 50
➢ S'adresser à un auditoire	Page 50

Fiche A14 : Production écrite

Production écrite	Brève description et/ou référence
1. Dans quels contextes (domaines, situations…) attend-on des candidats qu'ils prouvent leur compétence ? ➢ Le Tableau 5 dans le CECRL 4.1 peut servir de référence.	
2. Quels sont les thèmes de communication que les candidats doivent être capables de traiter ? ➢ Les listes du CECRL 4.2 peuvent servir de référence.	
3. Quels types de tâches, d'activités communicatives et quelles stratégies les candidats doivent-ils être capables de mettre en œuvre ? ➢ Les listes du CECRL 4.3, 4.4.2.1,7.1,7.2 et 7.3 peuvent servir de référence.	
4. Quels textes et types de textes attend-on que les candidats soient capables de traiter ? ➢ Les listes du CECRL 4.6.2 et 4.6.3 peuvent servir de référence.	
5. Après avoir pris connaissance de l'échelle de la production écrite générale reproduite ci-dessous, dites et justifiez à quel(s) niveau(x) de l'échelle l'épreuve devrait se situer. ➢ Les sous échelles d'interaction écrite du CECRL 4.4.1.2 énumérées à la suite de l'échelle peuvent servir de référence.	**Niveau :** **Justification (y compris références documentaires)**

	PRODUCTION ÉCRITE GÉNÉRALE
C2	Peut écrire des textes élaborés, limpides et fluides, dans un style approprié et efficace, avec une structure logique qui aide le destinataire à remarquer les points importants.
C1	Peut écrire des textes bien structurés sur des sujets complexes, en soulignant les points pertinents les plus saillants et en confirmant un point de vue de manière élaborée par l'intégration d'arguments secondaires, de justifications et d'exemples pertinents pour parvenir à une conclusion appropriée.
B2	Peut écrire des textes clairs et détaillés sur une gamme étendue de sujets relatifs à son domaine d'intérêt en faisant la synthèse et l'évaluation d'informations et d'arguments empruntés à des sources diverses.
B1	Peut écrire des textes articulés simplement sur une gamme de sujets variés dans son domaine en liant une série d'éléments discrets en une séquence linéaire.
A2	Peut écrire une série d'expressions et de phrases simples reliées par des connecteurs simples tels que "et", "mais" et "parce que".
A1	Peut écrire des expressions et phrases simples isolées.

Sous échelles correspondant à la production écrite	**Français**
➢ Ecriture créative	Page 52
➢ Essais et rapports	Page 52

A3.4 Capacités intégrées

Quelles combinaisons de capacités sont proposées dans les épreuves de l'examen ?
Préciser ces combinaisons dans la fiche 15, puis, pour chaque combinaison, remplissez la fiche 16.

Fiche A15 : Combinaison de capacités intégrées

Combinaisons de capacités intégrées		**Epreuve dans laquelle elles apparaissent**
Réception orale et prise de notes	□	
Réception orale et production orale	□	
Réception orale et production écrite	□	
Réception écrite et prise de notes	□	
Réception écrite et production orale	□	
Réception écrite et production écrite	□	
Réception orale et écrite et prise de notes	□	
Réception orale et écrite et production orale	□	
Réception orale et écrite et production écrite	□	

Fiche A16 : Capacités intégrées

	Répondez pour chacune des combinaisons citées ci-dessus
Capacités intégrées	**Brève description et/ou référence**
1. Quelles sont les combinaisons qui apparaissent ? ➢ Reportez-vous aux réponses données dans la Fiche A15.	
2. Quelles sont les activités de texte à texte ? ➢ Le Tableau 6 dans le CECRL 4.6.4 peut servir de référence.	
3. Dans quels contextes (domaines, situations…) attend-on des candidats qu'ils prouvent leur compétence ? ➢ Le Tableau 5 dans le CECRL 4.1 peut servir de référence.	
4. Quels sont les thèmes de communication que les candidats doivent être capables de traiter ? ➢ Les listes du CECRL 4.2 peuvent servir de référence.	
5. Quels types de tâches, d'activités communicatives et quelles stratégies les candidats doivent-ils être capables de mettre en œuvre ? ➢ Les listes du CECRL 4.3, 4.4.2.1,7.1,7.2 et 7.3 peuvent servir de référence.	
6. Quels types de textes attend-on que les candidats soient capables de traiter ? ➢ Les listes du CECRL 4.6.2 et 4.6.3 peuvent servir de référence.	
7. Après avoir pris connaissance de l'échelle Traiter un texte reproduite ci-dessous ainsi que des échelles réception de l'oral/écrit et de Production écrite déjà données, dites et justifiez à quel(s) niveau(x) de l'échelle l'épreuve devrait se situer. ➢ La sous échelle Prendre des notes du CECRL 4.6.3 peut servir de référence.	**Niveau :** **Justification (y compris références documentaires)**

	TRAITER UN TEXTE
C2	Peut faire le résumé d'informations de sources diverses en recomposant les arguments et les comptes rendus dans une présentation cohérente du résultat général
C1	Peut résumer de longs textes difficiles
B2	Peut résumer un large éventail de textes factuels et de fiction en commentant et en critiquant les points de vue opposés et les thèmes principaux Peut résumer des extraits de nouvelles (information), d'entretiens ou de documentaires traduisant des opinions, les discuter et les critiquer Peut résumer l'intrigue et la suite des événements d'un film ou d'une pièce
B1	Peut collationner des éléments d'information issus de sources diverses et les résumer pour quelqu'un d'autre
	Peut paraphraser simplement de courts passages écrits en utilisant les mots et le plan du texte
	Peut prélever et reproduire des mots et des phrases ou de courts énoncés dans un texte court qui reste dans le cadre de sa compétence et de son expérience limitées
A2	Peut copier des textes courts en script ou en écriture lisible
A1	Peut copier des mots isolés et des textes courts imprimés normalement

A3.5 Médiation

Fiche A17 : Médiation orale

Médiation orale	Brève description et/ou référence
1. Quelles sont les activités de texte à texte ? ➢ Le Tableau 6 dans le CECRL 4.6.4 peut servir de référence.	
2. Quelles sont les activités de médiation évaluées ? ➢ La liste du CECRL 4.4.4.1 peut servir de référence	
3. Dans quels contextes (domaines, situations…) attend-on des candidats qu'ils prouvent leur compétence ? ➢ Le Tableau 5 dans le CECRL 4.1 peut servir de référence.	
4. Quels sont les thèmes de communication que les candidats doivent être capables de traiter ? ➢ Les listes du CECRL 4.2 peuvent servir de textes de référence.	
5. Quels types de tâches, d'activités communicatives et quelles stratégies les candidats doivent-ils être capables de mettre en œuvre ? ➢ Les listes du CECRL 4.3, 4.4.2.1,7.1,7.2 et 7.3 peuvent servir de référence	
6. Quels types de textes attend-on que les candidats soient capables de traiter ? ➢ Les listes du CECRL 4.6.2 et 4.6.3 peuvent servir de référence.	
7. Le CECRL ne présente pas d'échelle pour la Traduction. En généralisant à partir des échelles de Réception orale, Traiter un texte et Production orale, dites et justifiez à quel(s) niveau(x) l'épreuve devrait se situer.	**Niveau :** **Justification (y compris références documentaires)**

Fiche A18 : Médiation écrite

Médiation écrite	Brève description et/ou référence
1. Quelles sont les activités de texte à texte ? ➢ Le Tableau 6 dans le CECRL 4.6.4 peut servir de référence.	
2. Quelles sont les activités de médiation évaluées ? ➢ La liste du CECR 4.4.4.2 peut servir de référence.	
3. Dans quels contextes (domaines, situations….) attend-on des candidats qu'ils prouvent leur compétence ? ➢ Le Tableau 5 dans le CECRL 4.1 peut servir de référence.	
4. Quels sont les thèmes de communication que les candidats doivent être capables de traiter ? ➢ Les listes du CECR 4.2 peuvent servir de textes de référence.	
5. Quels types de tâches, d'activités communicatives et quelles stratégies les candidats doivent-ils être capables de mettre en œuvre ? ➢ Les listes du CECRL 4.3, 4.4.2.1,7.1,7.2 et 7.3 peuvent servir de référence.	
6. Quels types de textes attend-on que les candidats soient capables de traiter ? ➢ Les listes du CECRL 4.6.2 et 4.6.3 peuvent servir de référence.	
7. Le CECRL ne présente pas d'échelle pour la Traduction. En généralisant à partir des échelles de Réception écrite, Traiter un texte et Production écrite, dites et justifiez à quel(s) niveau(x) l'épreuve devrait se situer.	**Niveau :** **Justification (y compris références documentaires)**

Partie A4 : Spécification : compétence langagière communicative (chapitre 4)

Fiches portant sur la compétence sont de nouveau proposées dans l'ordre suivant :
1. Réception
2. Interaction
3. Production
4. Médiation

A4 .1 Réception

Ces échelles du CECRL correspondant le mieux aux capacités de réception ont été utilisées pour élaborer le tableau A3, auquel on peut se référer dans cette partie. Les descripteurs des « niveaux plus » ne sont pas mentionnés dans le tableau A3. Les tableaux d'origine qui ont été pris en compte et dont certains définissent des niveaux plus, comprennent :

Compétence linguistique
- Etendue linguistique générale page 87
- Etendue du vocabulaire page 88

Compétence sociolinguistique
- Correction sociolinguistique page 95

Compétence pragmatique
- Développement thématique page 97
- Cohérence et cohésion page 98
- Précision page 101

Compétence stratégique
- Reconnaître des indices et faire des déductions page 60

Fiche A19 : Aspects de la compétence langagière pour la réception

Compétence linguistique	**Brève description et/ou référence**
1 Quelle étendue de la compétence lexicale et grammaticale attend-on que les candidats soient capables de maîtriser ? ➢ Les listes du CECR 5.2.1.1 et 5.2.1.2 peuvent servir de référence.	
2 Après avoir pris connaissance de l'échelle de Compétence linguistique du Tableau A.3, dites et justifiez à quel(s) niveau(x) l'examen devrait se situer.	**Niveau**
	Justification (y compris références documentaires)
Compétence sociolinguistique	**Brève description et/ou référence**
3 Quelles compétences sociolinguistiques attend-on que les candidats soient capables de mettre en œuvre : marqueurs linguistiques, règles de politesse, adéquation des registres, dialectes/accent, etc. ? ➢ Les listes du CECR 5.2.2 peuvent servir de référence.	
4 Après avoir pris connaissance de l'échelle de Compétence sociolinguistique du Tableau A. 3, dites et justifiez à quel(s) niveau(x) l'examen devrait se situer.	**Niveau**
	Justification (y compris références documentaires)

Compétence pragmatique	Brève description et/ou référence
5 Quelles compétences pragmatiques attend-on que les candidats soient capables de mettre en œuvre : compétences discursives, fonctionnelles ? ➢ Les listes du CECR 5.2.3 peuvent servir de référence.	
6 Après avoir pris connaissance de l'échelle de Compétence pragmatique du Tableau A.3, dites et justifiez à quel(s) niveau(x) l'examen devrait se situer.	**Niveau**
	Justification (y compris références documentaires)
Compétence stratégique	**Brève description et/ou référence**
7 Quelles compétences stratégiques attend-on que les candidats soient capables d'utiliser ? ➢ Les listes du CECR 4.4.2.4 peuvent servir de référence.	
8 Après avoir pris connaissance de l'échelle de Compétence stratégique du Tableau A.3, dites et justifiez à quel(s) niveau(x) l'examen devrait se situer.	**Niveau**
	Justification (y compris références documentaires)

TABLEAU A3 : ELEMENTS QUALITATIFS PERTINENTS POUR LA RECEPTION				
	LINGUISTIQUES **D'après « Etendue linguistique générale »** et **« Etendue du vocabulaire »**	**SOCIOLINGUISTIQUES** **D'après « Correction sociolinguistique »**	**PRAGMATIQUES** **D'après « Développement thématique » et « Précision »**	**STRATEGIQUES** **Reconnaître des indices et faire des déductions**
C2	*Peut comprendre avec précision une gamme très étendue de discours, apprécier l'insistance et la discrimination. Ne montre aucun signe d'incompréhension.* *Possède une bonne maîtrise d'un vaste répertoire lexical d'expressions idiomatiques et courantes avec la conscience du niveau de connotation sémantique*	*Manifeste une bonne maîtrise des expressions idiomatiques et dialectales avec la conscience des niveaux connotatifs de sens. Apprécie complètement les implications sociolinguistiques et socioculturelles de la langue utilisée par les locuteurs natifs et peut réagir en conséquence.*	*Peut comprendre avec précision des nuances de sens assez fines en utilisant une gamme étendue de procédés de modalisation (par exemple, adverbes exprimant le degré d'intensité, propositions restrictives)* *Peut comprendre l'insistance et la différenciation sans ambiguïté.*	*Comme C1*
C1	*Possède une bonne maîtrise d'un vaste répertoire lexical.* *Bonne maîtrise d'expressions idiomatiques et familières.*	*Peut reconnaître un large éventail d'expressions idiomatiques et dialectales et apprécier les changements de registre ; peut devoir toutefois confirmer tel ou tel détail, en particulier si l'accent n'est pas familier.* *Peut suivre des films utilisant largement l'argot et des expressions idiomatiques.* *Peut comprendre la langue avec efficacité et souplesse dans des relations sociales, y compris pour un usage affectif, allusif ou pour plaisanter.*	*Peut comprendre des descriptions et des récits compliqués avec des thèmes secondaires et certains plus développés.* *Peut comprendre avec précision les qualificatifs des opinions et des affirmations relatifs aux degrés, par exemple, de certitude/doute, croyance/doute, similitude, etc.*	*Est habile à utiliser les indices contextuels, grammaticaux et lexicaux pour en déduire une attitude, une humeur, des intentions et anticiper la suite.*
B2	*Possède une gamme assez étendue de langue pour comprendre des descriptions, des points de vue et des arguments sur la plupart des sujets pertinents pour sa vie quotidienne tels que la famille, les loisirs et centres d'intérêt, le travail, les voyages et l'actualité.*	*Peut, avec quelque effort, suivre des discussions rapides et familières.*	*Peut comprendre une description ou un récit, reconnaître les points saillants, des détails et des exemples.* *Peut comprendre une information détaillée de façon fiable.*	*Peut utiliser différentes stratégies de compréhension dont l'écoute des points forts et le contrôle de la compréhension par les indices textuels.*
B1	*Possède suffisamment de moyens linguistiques pour se débrouiller et un vocabulaire suffisant pour comprendre la plupart des textes sur des sujets tels que la famille, les loisirs et centres d'intérêt, le travail, les voyages et l'actualité.*	*Peut répondre à un large éventail de fonctions langagières en utilisant leurs expressions les plus courantes de manière neutre.* *Peut reconnaître les règles de politesse importantes.* *Est conscient des différences les plus significatives entre les coutumes, les usages, les attitudes, les valeurs et les croyances qui prévalent dans la communauté concernée et celles de sa propre communauté et en recherche les indices.*	*Peut, avec une exactitude relative, comprendre un récit ou une description linéaire.* *Peut comprendre les points principaux d'une idée ou d'un problème avec une certaine précision.*	*Peut identifier des mots inconnus à l'aide du contexte sur des sujets relatifs à son domaine et à ses intérêts.* *Peut extrapoler du contexte le sens de mots inconnus et en déduire le sens de la phrase à condition que le sujet en question soit familier.*
A2	*Possède un vocabulaire suffisant pour se débrouiller dans des situations courantes au contenu prévisible et pour répondre à des besoins simples de type concret.*	*Peut se débrouiller dans des échanges sociaux très courts en utilisant les formes quotidiennes polies d'accueil et de contact.* *Peut faire des invitations, des excuses et y répondre.*	*Peut comprendre une histoire ou une description consistant en une succession de points.* *Peut comprendre un échange d'information limité, simple et direct sur des sujets familiers et habituels.*	*Peut utiliser le sens général d'un texte ou d'un énoncé court sur des sujets quotidiens concrets pour déduire du contexte le sens probable de mots inconnus.*
A1	*Possède un choix élémentaire d'expressions simples pour les informations sur soi et les besoins de type courant.*	*Peut comprendre les formes de politesse les plus élémentaires d'accueil et prise de congé, de présentation ; dire merci, s'il vous plaît, excusez-moi, etc.*	*Pas de descripteur disponible*	*Pas de descripteur disponible*

A 4 .2 Interaction

Ces échelles du CECRL correspondant le mieux à l'interaction ont été utilisées pour élaborer le tableau A4, auquel on peut se référer dans cette partie. Les descripteurs des« niveaux plus » ne sont pas mentionnés dans le tableau A4. Les tableaux d'origine qui ont été pris en compte et dont certains définissent des niveaux plus, comprennent :

Compétence linguistique

- Etendue linguistique générale page 87
- Etendue du vocabulaire page 88
- Maîtrise du vocabulaire page 89
- Correction grammaticale page 90

Compétence sociolinguistique

- Correction sociolinguistique page 95

Compétence pragmatique

- Souplesse page 97
- Tours de parole page 97
- Aisance à l'oral page 100
- Précision page 101

Compétence stratégique

- Tours de parole page 70
- Coopérer page 71
- Faire clarifier page 71
- Compensation page 54
- Contrôle et correction page 54

Fiche A20 : Aspects de la compétence langagière en interaction

Compétence linguistique	**Brève description et/ou référence**
1. Quelle étendue de la compétence lexicale et grammaticale attend-on que les candidats soient capables de maîtriser ? ➢ Les listes du CECRL 5.2.1.1 et 5.2.1.2 peuvent servir de référence.	
2. Quelle étendue de la compétence phonologique et orthographique attend-on que les candidats soient capables d'utiliser ? ➢ Les listes du CECRL 5.2.1.4 et 5.2.1.5 peuvent servir de référence.	
3. Après avoir pris connaissance des échelles « Etendue » et « Correction » du Tableau A4, dites et justifiez à quel(s) niveau(x) l'examen devrait se situer. ➢ Les échelles pour la Maîtrise du système phonologique du CECRL 5.2.1.4 et pour la Maîtrise de l'orthographe 5.2.1.5 peuvent aussi servir de référence.	**Niveau** **Justification (y compris références documentaires)**
Compétence sociolinguistique	**Brève description et/ou référence**
4. Quelles compétences sociolinguistiques attend-on que les candidats soient capables de mettre en œuvre : marqueurs linguistiques, règles de politesse, adéquation des registres, dialectes/accent, etc. ? ➢ Les listes du CECRL 5.2.2 peuvent servir de référence.	
5. Après avoir pris connaissance de l'échelle de Compétence sociolinguistique du Tableau A.4, dites et justifiez à quel(s) niveau(x) l'examen devrait se situer.	**Niveau** **Justification (y compris références documentaires)**
Compétence pragmatique	**Brève description et/ou référence**
6. Quelles compétences pragmatiques attend-on que les candidats soient capables de mettre en œuvre : compétences discursives, fonctionnelles ? ➢ Les listes du CECRL 5.2.3 peuvent servir de référence.	
7. Après avoir pris connaissance de l'échelle pour l'Aisance du Tableau A4, dites et justifiez à quel(s) niveau(x) l'examen devrait se situer.	**Niveau** **Justification (y compris références documentaires)**
Compétence stratégique	**Brève description et/ou référence**
8. Quelles compétences stratégiques attend-on que les candidats soient capables d'utiliser : ➢ Le débat du CECRL 4.4.3.5 peut servir de référence.	
9. Après avoir pris connaissance de l'échelle pour l'Interaction du Tableau A.4, dites et justifiez à quel(s) niveau(x) l'examen devrait se situer.	**Niveau** **Justification (y compris références documentaires)**

A4 .3 Production

Ces échelles du CECRL correspondant le mieux aux capacités de production ont été utilisées pour élaborer le tableau A5, auquel on peut se référer dans cette partie. Les descripteurs des« niveaux plus » ne sont pas mentionnés dans le tableau A5. Les tableaux d'origine qui ont été pris en compte et dont certains définissent des niveaux plus, comprennent :

Compétence linguistique

- Etendue linguistique générale — Français page 87
- Etendue du vocabulaire — Français page 88
- Maîtrise du vocabulaire — Français page 89
- Correction grammaticale — Français page 90

Compétence sociolinguistique

- Correction sociolinguistique — Français page 95

Compétence pragmatique

- Souplesse — Français page 97
- Développement thématique — Français page 97
- Cohésion et cohérence — Français page 98
- Aisance à l'oral — Français page 100
- Précision — Français page 101

Compétence stratégique

- Planification — Français page 53
- Compensation — Français page 54
- Contrôle et correction — Français page 54

Fiche A21 : Aspects de la compétence langagière en production

Compétence linguistique	**Brève description et/ou référence**
1. Quelle étendue de la compétence lexicale et grammaticale attend-on que les candidats soient capables de maîtriser ? ➢ Les listes du CECRL 5.2.1.1 et 5.2.1.2 peuvent servir de référence.	
2. Quelle étendue de la compétence phonologique et orthographique attend-on que les candidats soient capables d'utiliser ? Les listes du CECRL 5.2.1.4 et 5.2.1.5 peuvent servir de référence.	
3. Après avoir pris connaissance des échelles de Etendue et Correction du Tableau 5, dites et justifiez à quel(s) niveau(x) l'examen devrait se situer. ➢ Les échelles pour la Maîtrise du système phonologique en CECRL 5.2.1.4 et pour la Maîtrise de l'orthographe en 5.2.1.5 peuvent aussi servir de référence.	**Niveau** **Justification (y compris références documentaires)**
Compétence sociolinguistique	**Brève description et/ou référence**
4. Quelles compétences sociolinguistiques attend-on que les candidats soient capables de mettre en œuvre : marqueurs linguistiques, règles de politesse, adéquation des registres, dialectes/accent, etc. ? ➢ Les listes du CECRL 5.2.2 peuvent servir de référence	
5. Après avoir pris connaissance de l'échelle de Compétence sociolinguistique du Tableau A5 dites et justifiez à quel(s) niveau(x) l'examen devrait se situer.	**Niveau** **Justification (y compris références documentaires)**

Compétence pragmatique	Brève description et/ou référence
6. Quelles compétences pragmatiques attend-on que les candidats soient capables de mettre en œuvre : compétences discursives, fonctionnelles ? ➢ Les listes du CECRL 5.2.3 peuvent servir de référence	
7. Après avoir pris connaissance de l'échelle pour la Compétence pragmatique du Tableau A5, dites et justifiez à quel(s) niveau(x) l'examen devrait se situer.	**Niveau**
	Justification (y compris références documentaires)
Compétence stratégique	**Brève description et/ou référence**
8. Quelles compétences stratégiques attend-on que les candidats soient capables d'utiliser ? ➢ Le débat du CECR 4.4.1.3 peut servir de référence	
9. Après avoir pris connaissance de l'échelle de Compétence stratégique du Tableau A5, dites et justifiez à quel(s) niveau(x) l'examen devrait se situer.	**Niveau**
	Justification (y compris références documentaires)

TABLEAU A4 : ELEMENTS QUALITATIFS PERTINENTS POUR L'INTERACTION ORALE

	ETENDUE LINGUISTIQUE **D'après « Etendue linguistique générale », « Etendue du vocabulaire », « Souplesse »**	**CORRECTION LINGUISTIQUE** **D'après « Correction grammaticale » et « Maîtrise du vocabulaire »**	**SOCIOLINGUISTIQUE** **D'après « Correction sociolinguistique »**	**AISANCE** **Aisance, Souplesse**	**INTERACTION** **D'après « Tours de parole », « Coopérer », « Faire clarifier »**
C2	Montre une grande souplesse dans la reformulation d'idées en les présentant sous des formes linguistiques variées pour accentuer l'importance, marquer une différence et lever l'ambiguïté. Possède aussi une bonne maîtrise d'un répertoire d'expressions idiomatiques et courantes	Peut maintenir constamment un niveau élevé de correction grammaticale même lorsque l'attention se porte ailleurs (par exemple, la planification ou l'observation des réactions de l'autre)	Apprécie complètement les implications sociolinguistiques et socioculturelles de la langue utilisée par les locuteurs natifs et peut réagir en conséquence. Peut jouer efficacement le rôle de médiateur entre les locuteurs de la langue cible et de celle de sa communauté d'origine en tenant compte des différences socioculturelles et sociolinguistiques	Peut s'exprimer longuement dans un discours naturel et sans effort en évitant ou en contournant les difficultés de sorte que l'interlocuteur ne s'en rend pas compte	Peut intervenir habilement et avec facilité en utilisant des expressions non-verbales ou l'intonation apparemment sans effort. Peut relier naturellement sa propre contribution à celle d'autres interlocuteurs en prenant la parole à son tour, faisant des références et des allusions, etc.
C1	Possède une bonne maîtrise d'un répertoire lui permettant de choisir la façon de s'exprimer clairement de manière appropriée sur un large éventail de sujets académiques, professionnels ou de loisirs sans restrictions sur ce qu'il/elle veut dire	Peut maintenir un niveau élevé de correction grammaticale ; les erreurs sont rares, difficiles à repérer et généralement corrigées rétrospectivement	Peut utiliser la langue avec efficacité et souplesse dans des relations sociales, y compris pour un usage affectif, allusif ou pour plaisanter	Peut s'exprimer avec aisance et spontanéité presque sans effort. Seul un sujet conceptuellement difficile est susceptible de gêner le flot naturel et fluide du discours	Peut choisir une expression adéquate dans un répertoire courant de fonctions discursives en préambule à ses propos pour obtenir la parole ou la garder et relier habilement sa propre contribution à celles de ses interlocuteurs
B2	Possède une gamme assez étendue de langue pour pouvoir faire des descriptions claires, exprimer son point de vue et développer une argumentation sans chercher ses mots de manière évidente et en utilisant des phrases complexes	A un niveau relativement élevé de correction grammaticale. Ne fait pas de fautes conduisant à des malentendus et peut corriger la plupart de ses fautes	Peut, avec quelque effort, comprendre et participer à des échanges dans un groupe même si le discours est rapide et familier. Peut poursuivre une relation suivie avec des locuteurs natifs sans les amuser ou les irriter sans le vouloir ou les mettre en situation de se comporter autrement qu'avec un locuteur natif	Peut s'adapter aux changements de sujet, de style et de ton rencontrés normalement dans une conversation. Peut parler relativement longtemps avec un débit assez régulier ; bien qu'il/elle puisse hésiter pour chercher tournures et expressions, on remarque peu de longues pauses	Peut commencer un discours, prendre la parole au bon moment et terminer la conversation quand il/elle le souhaite bien que parfois sans élégance. Peut faciliter le développement de la discussion sur un terrain connu en confirmant sa compréhension, en invitant les autres à participer, etc.
B1	Possède suffisamment de moyens linguistiques pour s'en sortir avec quelques hésitations et quelques périphrases sur des sujets tels que la famille, les loisirs et centres d'intérêt, le travail, les voyages et l'actualité	Peut se servir avec une correction suffisante d'un répertoire de tournures et expressions fréquemment utilisées et associées à des situations plutôt prévisibles	Peut s'exprimer et répondre aux fonctions langagières de base telles que l'échange d'information et la demande et exprimer simplement une idée et une opinion. Est conscient des règles de politesse importantes et se conduit de manière appropriée	Peut exploiter avec souplesse une gamme étendue de langue simple afin d'exprimer l'essentiel de ce qu'il/elle veut dire. Peut s'exprimer avec une certaine aisance. Malgré quelques problèmes de formulation ayant pour conséquence pauses et impasse, est effectivement capable de continuer à parler sans aide	Peut commencer, poursuivre et terminer une simple conversation en tête-à-tête sur des sujets familiers ou d'intérêt personnel. Peut reformuler en partie les dires de l'interlocuteur pour confirmer une compréhension mutuelle
A2	Peut utiliser des modèles de phrases élémentaires et communiquer des informations limitées dans des situations courantes de la vie quotidienne à l'aide de phrases mémorisées, de groupes de mots et d'expressions toutes faites	Peut utiliser des structures simples correctement mais commet encore systématiquement des erreurs élémentaires	Peut se débrouiller dans des échanges sociaux très courts en utilisant les formes quotidiennes polies d'accueil et de contact. Peut faire des invitations, des excuses et y répondre	Peut se faire comprendre dans une brève intervention même si la reformulation, les pauses et les faux démarrages sont très évidents. Peut développer des expressions apprises par la simple recombinaison de leurs éléments	Peut indiquer qu'il/elle suit ce qui se dit mais est rarement en mesure de comprendre suffisamment pour poursuivre la conversation. Peut attirer l'attention
A1	Possède un répertoire élémentaire de mots isolés et d'expressions simples relatives à soi et à des situations concrètes particulières	A un contrôle limité de structures syntaxiques et de formes grammaticales appartenant à un répertoire mémorisé	Peut établir un contact social de base en utilisant les formes de politesse les plus élémentaires : accueil et prise de congé ; présentation et dire merci, s'il vous plaît, excusez-moi, etc.	Peut se débrouiller avec des énoncés très courts, isolés, généralement stéréotypés, avec de nombreuses pauses pour chercher ses mots pour prononcer les moins familiers et pour remédier à la communication	Peut intervenir simplement mais la communication repose entièrement sur la répétition, la reformulation et la remédiation

TABLEAU A5 : ELEMENTS QUALITATIFS PERTINENTS POUR LA PRODUCTION

	ETENDUE LINGUISTIQUE **D'après « Etendue linguistique générale », « Etendue du vocabulaire »**	**CORRECTION LINGUISTIQUE** **D'après « Correction grammaticale », « Maîtrise du vocabulaire », « Maîtrise du système phonologique »**	**SOCIOLINGUISTIQUE** **D'après « Correction sociolinguistique »**	**PRAGMATIQUE** **Aisance à l'oral, Souplesse**	**PRAGMATIQUE** **Développement thématique, Précision, Cohérence et cohésion**	**STRATEGIQUE** **Compensation, Contrôle et correction**
C2	Possède une grande souplesse pour reformuler les idées de différentes manières afin d'exprimer avec précision des nuances fines de sens pour insister, discriminer et lever l'ambiguïté. Possède également une bonne maîtrise d'expressions idiomatiques et courantes	Peut maintenir constamment un niveau élevé de correction grammaticale même lorsque l'attention se porte ailleurs (par exemple, la planification ou l'observation des réactions de l'autre)	Apprécie complètement les implications socioculturelles de la langue utilisée par les autres locuteurs et peut réagir en conséquence	Peut s'exprimer longuement avec spontanéité dans une langue courante, en évitant ou contournant les difficultés de telle sorte que l'interlocuteur ne s'en rend pas compte	Peut créer un texte cohérent et cohésif en utilisant de manière complète et appropriée les structures organisationnelles adéquates et une grande variété d'articulateurs	Peut substituer à un mot qui lui échappe un terme équivalent de manière si habile que l'on s'en rende à peine compte
C1	Possède une bonne maîtrise d'une vaste étendue de langue lui permettant de choisir la formulation appropriée pour s'exprimer clairement de manière appropriée sur une gamme importante de sujets généraux, académiques, professionnels ou sur les loisirs sans avoir à restreindre ce qu'il/elle veut dire	Peut maintenir constamment un niveau élevé de correction grammaticale ; les erreurs sont rares, difficiles à repérer et généralement corrigées aussitôt	Peut utiliser la langue avec efficacité et souplesse dans des relations sociales, y compris pour un usage affectif, allusif ou pour plaisanter	Peut s'exprimer avec aisance et spontanéité presque sans effort ; seul un sujet conceptuellement difficile est susceptible de gêner le flot naturel et fluide du discours	Peut produire un texte clair, fluide et bien structuré, démontrant un usage contrôlé de moyens linguistiques de structuration et d'articulation. Peut faire des descriptions et des récits compliqués avec des thèmes secondaires et certains plus développés et arriver à une conclusion adéquate	Peut contourner une difficulté rencontrée et reformuler ce qu'il/elle veut dire sans interrompre complètement le fil du discours
B2	Possède une gamme assez étendue de langue pour pouvoir faire des descriptions claires, exprimer son point de vue et développer une argumentation sans chercher ses mots de manière évidente et en utilisant des phrases complètes	Possède un niveau relativement élevé de correction grammaticale. Ne fait pas de fautes conduisant à des malentendus et peut corriger la plupart d'entre elles rétrospectivement	Peut s'exprimer de façon appropriée à la situation et éviter des erreurs grossières de formulation	Peut parler relativement longtemps avec un débit assez régulier ; bien qu'il/elle puisse hésiter en cherchant tournures et expression, on remarque peu de longues pauses	Peut faire une description ou un récit clair en développant et argumentant les points importants à l'aide de détails et d'exemples significatifs. Peut utiliser un nombre limité d'articulateurs pour relier ses énoncés en un discours clair et cohérent bien qu'il puisse y avoir quelques « sauts » dans une longue intervention	Peut utiliser des périphrases et des paraphrases pour dissimuler des lacunes lexicales et structurales. Peut relever ses erreurs habituelles et surveiller consciemment son discours afin de les corriger
B1	Possède suffisamment de moyens linguistiques et d'un vocabulaire suffisant pour s'en sortir avec quelques hésitations et quelques périphrases sur des sujets tels que la famille, les loisirs et centres d'intérêt, le travail, les voyages et l'actualité	Peut se servir avec une correction suffisante d'un répertoire de tournures et expressions fréquemment utilisées et associées à des situations plutôt prévisibles	Pas de descripteur disponible	Peut exploiter avec souplesse une gamme étendue de langue simple afin d'exprimer l'essentiel de ce qu'il/elle veut dire. Peut discourir de manière compréhensible même si les mots pour chercher ses mots et ses phrases et pour faire ses corrections sont évidents, notamment dans les séquences plus longues de production libre	Peut relier une série d'éléments courts, simples et distincts afin de raconter ou de décrire, avec une relative aisance, quelque chose de simple et de linéaire	Peut utiliser un mot simple signifiant quelque chose de semblable au concept recherché et solliciter une « correction ». Peut recommencer, avec une tactique différente, s'il y a rupture de communication
A2	Peut utiliser des modèles de phrases élémentaires et communiquer à l'aide de phrases mémorisées de groupes de mots et d'expressions toutes faites pour transmettre des informations limitées sur de simples situations quotidiennes	Peut utiliser correctement des structures simples mais commet encore systématiquement des erreurs élémentaires	Pas de descripteur disponible	Peut se faire comprendre dans une brève intervention, même si la reformulation, les pauses et les faux démarrages sont très évidents. Peut développer des expressions apprises par la simple recombinaison de leurs éléments	Peut relier des groupes de mots avec des connecteurs simples tels que « et », « mais » et « parce que »	Pas de descripteur disponible
A1	Possède un répertoire élémentaire de mots et d'expressions élémentaires relatifs à soi et à des situations concrètes particulières	A un contrôle limité de structures syntaxiques et de formes grammaticales simples appartenant à un répertoire mémorisé	Pas de descripteur disponible	Peut se débrouiller avec des énoncés très courts, isolés, généralement stéréotypés, avec de nombreuses pauses pour cher ses mots, pour prononcer les moins familiers et pour remédier à la communication	Peut relier des groupes de mots avec des connecteurs très élémentaires tels que « et » ou « alors »	Pas de descripteur disponible

A4.4 Médiation

C'est de la nature de la médiation que vont dépendre les échelles du CECR qui seront le plus appropriées.
En situation de langue étrangère, on met naturellement l'accent sur les capacités dans la langue étrangère. Pour des activités de médiation effectuées à partir de la langue étrangère vers la langue maternelle, les capacités requises seront essentiellement du domaine de la réception tandis que pour la médiation effectuée à partir de la langue maternelle vers la langue étrangère, ce sont les capacités en production qui seront nécessaires. En ce qui concerne la Médiation entièrement dans la langue étrangère, on fera appel à la fois à la réception et à la production.

Variables	**Type de compétences langagières**	**Descripteurs**
a. dans une langue étrangère	Pour la Réception et la Production	Tableaux A3 et A5
b. d'une langue étrangère à une autre	Pour la Réception et la Production	Tableaux A3 et A5
c. d'une langue étrangère vers la langue maternelle	Pour la Réception	Tableau A3
d. de la langue maternelle vers la langue étrangère	Pour la Production	Tableau A5

Les autres paramètres à prendre en considération sont les variables des capacités (de la réception de l'oral ou de l'écrit vers la production orale ou écrite) et les variables des tâches – selon un registre formel ou informel – comme cela est indiqué dans le CECR 4.4.4.1 (médiation orale) et 4.4.4.2 (médiation écrite).

Ainsi, bien qu'il n'y ait pas de descripteurs pour la Médiation en tant que telle dans le CECRL, toutes les échelles de descripteurs du Chapitre 5 du CECRL auxquelles s'ajoutent les échelles pour les Stratégies de Réception et de Production (incluses respectivement dans les Tableaux A3 et A5) sont pertinentes.

Si l'examen comprend de la Médiation, consulter les Tableaux A3, A4 et/ou A5 pour remplir la Fiche A22

Fiche A22 : Aspects de la compétence langagière en médiation

Compétence linguistique	**Brève description et/ou référence**
1. Quelle étendue de la compétence lexicale et grammaticale attend-on que les candidats soient capables de maîtriser ? ➢ Les listes du CECRL 5.2.1.1 et 5.2.1.2 peuvent servir de référence.	
2. Quel type de relations sémantiques attend-on que les candidats soient capables de traiter ? ➢ Les listes du CECRL 5.2.1.3 peuvent servir de référence.	
3. Quelle étendue de la compétence phonologique et orthographique attend-on que les candidats soient capables de maîtriser ? ➢ Les listes du CECRL 5.2.1.4 et 5.2.1.5 peuvent servir de référence.	
4. L'échelle pour la Maîtrise de l'orthographe du CECRL 5.2.1.5 peut également servir de référence.	**Niveau**
	Justification (y compris références documentaires)

Compétence sociolinguistique	Brève description et/ou référence
5. Quelles compétences sociolinguistiques attend-on que les candidats soient capables de mettre en œuvre : marqueurs linguistiques, règles de politesse, adéquation des registres, dialectes/accent, etc. ? ➢ Les listes du CECRL 5.2.2 peuvent servir de référence	
6. Après avoir pris connaissance de l'échelle de Compétence sociolinguistique du Tableau A3 et A.4, dites et justifiez à quel(s) niveau(x) l'examen devrait se situer.	**Niveau**
	Justification (y compris références documentaires)
Compétence pragmatique	**Brève description et/ou référence**
7. Quelles compétences pragmatiques attend-on que les candidats soient capables de mettre en œuvre : compétences discursives, fonctionnelles ? ➢ Les listes du CECRL 5.2.3 peuvent servir de référence	
8. Après avoir pris connaissance de l'échelle pour la Compétence pragmatique du Tableau A5, dites et justifiez à quel(s) niveau(x) l'examen devrait se situer.	**Niveau**
	Justification (y compris références documentaires)
Compétence stratégique	**Brève description et/ou référence**
9. Quelles stratégiques de réception et de production attend-on que les candidats soient capables d'utiliser ? ➢ Le débat du CECRL 4.4.2.4 et 4.4.1.3 peut servir de référence	
10. Après avoir pris connaissance de l'échelle de Compétence stratégiques des Tableaux A3 et A5, dites et justifiez à quel(s) niveau(x) l'examen devrait se situer.	**Niveau**
	Justification (y compris références documentaires)

Partie A5: Spécification : Résultat des analyses (chapitre 4)

La fiche A23 propose un profil sous forme de graphique de ce que recouvre un examen en relation avec les catégories et aux niveaux du CECRL. Ce tableau est à remplir à la fin du processus de Spécification.

Fiche A23 : Représentation graphique de la relation de l'examen aux niveaux du CECRL (exemple)

C2								
C1								
B2.2								
B2								
B1.2								
B1								
A2.2								
A2								
A1								
Ensemble	Réception orale	Réception écrite	Conversation sociale	Echange d'information	Notes Messages et formulaires	Socio linguistique	Pragmatique	Linguistique

Fiche A24 : Confirmation de l'estimation du niveau global de l'examen

Confirmation de l'estimation du niveau global de l'examen par rapport au CECRL		
☐ **A1**	☐ **B1**	☐ **C1**
☐ **A2**	☐ **B2**	☐ **C2**
Brève justification, référence de la documentation. Si les conclusions de cette fiche sont différentes de celles de la fiche 8, indiquer les raisons principales de ce changement.		

Annexe B

Grilles d'analyse de contenu (chapitre 4)

Partie B1 : Grille d'analyse du contenu du CECRL pour la réception orale et la réception écrite

Les concepteurs de tests ou d'examens peuvent relier les épreuves de réception écrite et orale au CECRL grâce à la grille d'analyse de contenu pour la réception orale et la réception écrite du CECRL [45],. Les informations concernant chaque tâche, texte et item du test ou de l'examen sont indiquées dans la grille avec leurs caractéristiques (par exemple la source/l'origine du texte, le type de discours, le niveau de difficulté estimé, etc…) choisies parmi les options proposées par le CECRL. Une utilisation efficace de la grille suppose de la part de celui/celle qui assure l'analyse une bonne connaissance du CECRL. Un chapitre portant sur la familiarisation avec le CECRL est proposé pour aider à la mise en œuvre de l'ensemble. La grille a été conçue pour être utilisée en ligne mais une version papier est disponible dans ce manuel. On peut ajouter, si cela s'impose, de nouvelles catégories.

Alors que la grille a été avant tout conçue pour analyser des épreuves de réception écrite et orale, elle peut aussi servir d'outil pour les concevoir.

Un lien avec la version en ligne est aussi disponible sur www.coe.portfolio . Le lien direct est www.lancs.ac.uk/fss/projects/grid.

Dans cette partie, la même fiche a été proposée en trois versions :

1. une version vierge ;
2. une version complétée à la suite de l'analyse du panel d'experts débouchant sur des points de césure provisoires ;
3. une troisième version dans laquelle les classements provisoires des items ont été revus à la suite de la comparaison entre les données issues des difficultés estimées et les données empiriques sur ces mêmes difficultés. Des ajustements identiques ont été opérés sur les points de césure.

[45] La grille a été élaborée par un groupe de travail comprenant Charles Alderson (coordinateur du projet), Neus Figueras, Henk Kuijpers, Günther Nold, Sauli takala et Claire Tardieu. Grâce à une subvention du ministère de l'Education néerlandais, le groupe a élaboré une version informatique disponible sur le site www.lancs.ac.uk/fss/projects/grid. Un rapport de ce projet est disponible à la demande auprès du coordinateur du projet : c.alderson@lancaster.ac.uk

Fiche vierge de réception orale

Réception orale/écrite en … (langues)…					
Niveau à atteindre dans le programme :					
Types d'items					
Support					
Durée (total : 45 minutes)					
Authenticité					
Type de discours					
Domaine					
Thème					
Rapport avec le programme (une nouvelle catégorie optionnelle)					
Nombre de locuteurs					
Prononciation					
Contenu					
Grammaire					
Vocabulaire					
Nombre d'écoutes					
Texte proposé compré-hensible au niveau					
Items compréhensibles au niveau (indiquer les codes de l'item)					
A1					
A1/A2					
A2					
A2/B1					
B1					
B1/B2					
B2					
B2/C1					
C1					
C1/C2					
C2					

Echantillon de Spécification d'un test de réception orale

Test	Réception orale en français				
Niveau à atteindre dans le programme: B2.1					
Types d'items	30 items à choix multiple				5 items à réponse ouverte
Support	Interview	Interview	Présentation	Programme de radio	**Actualités**
Durée (total : 45 minutes)	7	12	7	9	10
Authenticité	Modifié	Modifié	Authentique	Authentique	Modifié (abrégé)
Type de discours	Narratif	Argumentatif	Descriptif	Descriptif	Narratif
Domaine	Personnel	Personnel	Public	Public	Public
Thème	Culture pop	Environnement	Affaires/commerce	Loisirs	Société
Rapport avec le programme	Note	**Note**	Note	Note	Note
Nombre de locuteurs	2	2 + 1	2	1	1
Prononciation	Norme française	Norme francophone	Norme française	Norme francophone	Norme française
Contenu	Concret	Concret	Assez abstrait	Assez abstrait	Assez abstrait
Grammaire	Simple	Assez complexe	Plutôt complexe	Plutôt complexe	Assez complexe
Vocabulaire	Uniquement fréquent	Surtout fréquent	Plutôt étendu	Plutôt étendu	Plutôt étendu
Nombre d'écoutes	2	2	2	1	1
Texte proposé compré-hensible au niveau du CECRL					
Items compréhensibles au niveau (indiquer le classement en utilisant les codes d'items)					
A1					
A1/A2					
A2					
A2/B1					
B1	1, 2, 3, 4, 5			25, 27	
B1/B2		6, 7, 8, 10, 12, 14, 15	17	24, 26	Réponse ouverte : 1, 2
B2		9, 11, 13, 16	18, 19, 20	21, 22, 23	Réponse ouverte: 4
B2/C1				28, 29, 30	Réponse ouverte: 3, 5
C1					
C1/C2					
C2					

Points de césure initiaux: < B1: 0; B1: 1–19; B2: 20–30; >B2: 31–35

Echantillon de grille à utiliser après la passation du test

Test	Réception orale en français				
Niveau cible dans le programme: B2.1					
Types d'items	30 items à choix multiple				5 à compléter
Support	Interview	Interview	Présentation	Programme de radio	Actualités
Durée (total : 45 minutes)	7	12	7	9	10
Authenticité	Modifié	Modifié	Authentique	Authentique	Modifié (abrégé)
Type de discours	Narratif	Argumentatif	Descriptif	Descriptif	Narratif
Domaine	Personnel	Personnel	Public	Public	Public
Thème	Culture pop	Environnement	Affaires/commerce	Loisirs	Société
Rapport avec le programme	Note	Note	Note	Note	Note
Nombre de locuteurs	2	2 + 1	2	1	1
Prononciation	Norme française	Norme francophone	Norme française	Norme francophone	Norme française
Contenu	Concret	Concret	Assez abstrait	Assez abstrait	Assez abstrait
Grammaire	Simple	Assez complexe	Plutôt complexe	Assez complexe	Assez complexe
Vocabulaire	Uniquement fréquent	Surtout fréquent	Plutôt étendu	Plutôt étendu	Plutôt étendu
Nombre d'écoutes	2	2	2	1	1
Texte proposé compré-hensible au niveau					
Items compréhensibles au niveau (indiquer les codes d'items après définition des points de césure)					
A1					
A1/A2					
A2					
A2/B1					
B1					
B1/B2					
B2					
B2/C1					
C1					
C1/C2					
C2					

Points de césure finaux:

Echantillon de grille vierge pour un test de réception écrite

Caractéristiques	Texte 1	Texte 2	Texte 3	Texte 4	Texte 5
Source du texte					
Authenticité					
Type de discours					
Domaine					
Thème					
Nature du contenu					
Longueur du texte					
Vocabulaire					
Grammaire					
Texte susceptible d'être compris par des apprenants de niveau du CECRL :					

Items compréhensibles par des apprenants/utilisateurs au niveau CECRL (indiquer le code de l'item)					
A1					
A2					
B1					
B2					
C1					
C2					

Points de césure initiaux : ……Points de césure finaux :……...

Partie B2 : Grilles d'analyse de contenu (du CECRL) pour les tâches de production écrite et de production orale

Ces grilles ont été conçues par un groupe de travail au sein de ALTE dont l'objectif est d'aider les concepteurs d'examen qui utilisent le CECRL et le manuel. Ce groupe de ALTE actualise les grilles en tenant compte des retours d'informations des utilisateurs. On recommande donc aux utilisateurs de télécharger les dernières versions des pages du site de la Division des Politiques linguistiques du conseil de l'Europe : www.coe.int/lang.

Les concepteurs des grilles avaient pour objectif de procurer des outils souples utilisables dans des contextes différents et pour une utilisation multiple.

Il existe deux types de grilles :

Grille d'analyse : utilisée quand des panels d'experts doivent donner leur opinion sur les tâches d'un test ou d'un examen, à l'occasion par exemple de sessions de formation, ou de sessions montrant des échantillons représentatifs ou des exercices de définition de points de césure.

Grille de présentation : utilisée pour présenter une analyse déjà faite, voire des modèles pour la formation et la standardisation, pour un rapport ou une présentation dans des séminaires.

Ces grilles n'ont jamais été conçues pour être utilisées d'une seule manière, et il n'est donc pas possible de donner dans ce Manuel des instructions d'utilisation exhaustives. C'est pour cette raison que seuls deux exemples de la façon dont elles ont été utilisées ont été proposés.

Exemple 1

Grille utilisée :

Grille de production écrite du CECRL : analyse, version 3.0, 2005

Utilisation: calibrage de performances écrites d'un ensemble d'examens locaux.

Procédure :

Dans un atelier de calibrage regroupant 11 experts, la grille a été utilisée comme une activité introductive. On a demandé aux experts de la compléter pour une des tâches, puis de discuter entre eux du degré de pertinence de chaque catégorie de la grille pour relier la tâche à un niveau. Le but de cette activité était d'amener les experts à se concentrer sur la relation entre la tâche et la performance et sur les différents aspects de la difficulté de la tâche.

Une modification de la grille a également été proposée aux experts, en particulier l'insertion de la catégorie « type de texte attendu », en complément de la catégorie « type de texte proposé ».

Dans une colonne supplémentaire, les experts pouvaient indiquer quelles étaient les catégories qu'ils considéraient comme étant déterminantes pour relier une tâche à un niveau. Ils ont eu à prendre cette décision pour chacune des catégories numérotées de 16 à 38. Les catégories le plus souvent mentionnées étaient : « type de texte attendu » (10 fois), « temps permis ou suggéré »(9 fois), « type de texte proposé »(8) « sujet ou thème proposé » (8), « Nombre de mots attendus »(8).Certaines catégories ont donné lieu à des discussions sur (a) l'interprétation des catégories et (b) la possibilité de les appliquer à tous les niveaux (par exemple la catégorie « type de texte proposé » que les experts ont considéré pertinente uniquement dans les niveaux supérieurs.

A noter, les points positifs : Le calibrage se concentre sur les qualités linguistiques du texte, plutôt que sur les aspects d'accomplissement de la tâche. La grille a permis que certains aspects d'accomplissement

de la tâche soient pris en compte dans la discussion sur la qualité du texte, par exemple, la durée de l'épreuve de production écrite.

A noter, les points négatifs : Différentes personnes ont tendance à interpréter différemment certaines catégories (par exemple jusqu'à quel point « rédaction à moitié directive » est directif).

Recommandations :

L'utilisation de la grille dans un atelier avec des concepteurs de tests ou d'examens serait bien utile, car ce serait là l'occasion d'une réflexion sur le niveau de langue d'une tâche, et ainsi sur les caractéristiques qu'une tâche devrait avoir pour susciter la performance attendue.

Une façon de promouvoir une interprétation identique des termes de la grille serait, de la part des organisateurs, de fournir des échantillons représentatifs, accompagnés éventuellement d'une version finale faisant état des conclusions auxquelles ils sont arrivés dans cette activité.

Exemple 2

Grille utilisée :

Grille de production orale du CECRL : analyse et présentation, version 01,09/12/05

Utilisation: calibrage de performances orales d'un examen local.

Procédure :

Pendant la phase de formation, on a montré à 12 experts des vidéos de performances calibrées qui avaient été sélectionnées lors d'un séminaire de calibrage organisé pour la langue concernée en coopération avec le Conseil de l'Europe. Chaque juge devait classer les performances filmées sur les niveaux du CECRL. Les experts devaient d'abord reporter individuellement leur évaluation sur la grille finale puis discuter par groupe de deux puis en grand groupe.

La grille était utilisée pour faire prendre conscience aux experts de la difficulté de la tâche et leur montrer quel genre de catégories pouvait avoir plus d'influence que d'autres sur la difficulté. Comme la performance d'un candidat est en étroite relation avec la réponse que la tâche induit, cela a permis d'avoir une idée de la difficulté de la tâche avant de commencer à évaluer les échantillons de performances.

Dans un deuxième temps, les grilles ont été utilisées de façon identique pour classer les tâches de production orale et évaluer les performances d'échantillons de l'examen local.

A noter, les points positifs : Cette méthode a bien fonctionné car les experts ont eu une idée plus précise des différents aspects de la difficulté d'une tâche et du niveau des performances qui y correspondent. Cela a, en particulier, facilité l'évaluation des tâches proposées dans l'examen local.

A noter, les points négatifs : Une des difficultés de cette méthode est le temps pris pour expliquer les 45 catégories de la grille proposée. La grille a donc été traduite dans la langue utilisée par les experts et une sélection a été faite des catégories utilisées pendant la réunion. La partie 1 d'information générale a été laissée de côté ; dans la partie 2, ont été plus particulièrement pris en compte les conseils 15/16 et le thème 23. En revanche la partie 3 a été gardée dans sa totalité.

Recommandations : Envoyer la grille aux experts avant la réunion sur la standardisation afin qu'ils se familiarisent avec ce document.

La grille du CECRL pour les tâches de production écrite v.3.1 (Présentation)

Cette grille a été conçue par un groupe de travail au sein de ALTE dont l'objectif est d'aider les concepteurs d'examen qui utilisent le *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer* et le *Manuel pour relier les examens au CECRL* disponibles auprès de la Division des Politiques linguistiques du conseil de l'Europe.

Deux versions sont disponibles : la grille d'analyse et la grille de présentation (version simplifiée)

La grille d'analyse est destinée à être utilisée dans des ateliers et des séminaires de calibrage.

- Si le but de l'atelier est d'analyser le contenu et les spécifications d'un test, l'étape adéquate est celle de la Spécification (chapitre 4).
- Si la grille est utilisée pour calibrer des échantillons locaux nouveaux, la partie adéquate du Manuel est la partie 5.6.

La grille de présentation fournit un rapport descriptif de l'analyse des tâches d'un test, telle qu'elle a été faite dans un exercice de calibrage préalable. Si les grilles complétées sont utilisées pour la description d'échantillons représentatifs, elles peuvent être exploitées lors d'une formation à la standardisation (chapitre 5 de ce manuel).

Echantillons des tâches d'un test

Rapport sur l'analyse de	
Langue cible de ce test	
Niveau cible (CECRL) de ce test	
Numéro/nom de la tâche	

Information générale - le test dans son ensemble

1. La durée du test ou de l'examen dans son ensemble
2. L'objectif
3. Le contexte /l'arrière plan de l'examen
4. Les candidats
5. La structure du test

Information générale - l'épreuve de production écrite

6. Le nombre de tâches dans l'épreuve de production écrite
7. La durée de l'ensemble de l'épreuve
8. L'intégration des capacités
9. Le mode de présentation
10. Le niveau CECRL de l'épreuve
11. Le format de l'épreuve de production écrite
12. L'information spécifique – exemple de tâche

13. La distribution des notes
14. L’évaluation de la tâche
15. Le niveau réel
16. Un échantillon de tâche

Echantillon de tâche à mettre ici

i) Tâche proposée/déclencheur		
17	La langue de la tâche proposée/du déclencheur	
18	Le niveau CEFRL de la tâche proposée / du déclencheur	
19	La durée permise ou suggérée pour cette tâche	minutes
20	Les directives /Conseils	
21	Le contenu	
22	Le genre	
23	La fonction rhétorique de la tâche proposée	
24	Le public attendu	
25	Support de la tâche proposée/ du déclencheur	
26	Le thème de la tâche	
27	L’intégration des capacités langagières dans la tâche proposée	

ii) Réponse (description de la réponse écrite suscitée par le déclencheur/la tâche proposée		
28	Le nombre de mots attendus	
29	La fonction rhétorique attendue	
30	L’objectif du texte	
31	Le registre	
32	Le domaine	
33	La compétence grammaticale attendue	
34	La compétence lexicale attendue	
35	La compétence discursive attendue	

36	L’authenticité : situationnelle	
37	L’authenticité : interactionnelle	
38	Le processus cognitif	
39	La connaissance du contenu	

iii) Evaluation de la tâche		
40	Les critères connus	
41	La méthode d’évaluation de la tâche	
42	Les critères d’évaluation	
43	Le nombre et la composition des évaluateurs	

iv) Retours d’informations aux candidats		
44	Les retours d’informations quantitatifs	
45	Les retours d’informations qualitatifs	

46 Exemple de réponse

47 Commentaire

48 Résultats attribués

Notes : les numéros ci-dessous correspondent à ceux des items de la grille.

2. L’objectif du test peut être l’évaluation de la compétence générale, ou d’une compétence sur objectif spécifique. Indiquer l’objectif s’il est spécifique (français pour le droit, allemand pour des objectifs universitaires, etc.).

3. La description de l’arrière plan peut comprendre les raisons pour lesquelles ce test est conçu, une description de l’ensemble des tests dont fait partie ce test, ou d’autres détails de ce type.

4. Décrire le nombre et le profil des candidats (nationalités, âge…).

5. Décrire les autres épreuves du test ou de l’examen (par exemple l’épreuve de production orale, de réception écrite).

6. Au cas où le nombre de tâches dépend des options choisies, le spécifier dans l’introduction (point 5).

8 Les capacités, en plus de la production écrite, qui sont prises en compte dans cette tâche (indépendamment du fait qu’elles soient prises en compte de façon explicite au moment d’évaluer). Choisir entre : aucune, réception écrite, orale, production orale, une combinaison.

9 Sous quel format sont consignées les réponses du candidat. Choisir entre le format manuscrit, la saisie électronique ou l’un ou l’autre ou les deux.

10 CECRL, chapitre 3

11 La description peut comprendre des informations sur le nombre de parties dans l'épreuve, le type de tâche dans chaque partie, la durée allouée à chaque partie.

12 Il est possible d'inclure une description courte de la tâche à ce niveau. La description peut comprendre les buts de la tâche, ce qu'on demande aux candidats de faire et ce qui est attendu pour pouvoir juger que la tâche est totalement accomplie.

13 Décrire comment les points sont répartis dans cette partie de la tâche et ce que les candidats devraient faire pour obtenir la totalité des points.

14 Expliquer comment la tâche est évaluée (par exemple manuellement, automatiquement), quels outils sont utilisés et quels sont les éléments pris en compte dans la décision du niveau.

15 Décrire les mesures prises pour s'assurer que les tâches de production écrite sont au niveau approprié. Cette description peut comprendre le processus de conception de l'épreuve et le pré-test.

16 Placer ici un échantillon de tâche, y compris la consigne et le document déclencheur.

18 Choisir un niveau du CECRL : A1, A2, B1, B2, C1, C2.

19 Si cela n'est pas précisé, la durée attendue.

20 Indiquer jusqu'à quel point la consigne, le document déclencheur ou la tâche proposée déterminent la nature et le contenu de la réponse. Choisir entre : directif, semi-directif ou réponse ouverte.

21 Le contenu de la réponse est-il précisé dans la consigne ? Choisir entre : précisé ou non précisé.

22 Choisir entre : **lettre (domaine professionnel), lettre (domaine personnel), revue, essai, rédaction, rapport, récit, projet, article, fiche**, autre (préciser).

23 Les fonctions attendues dans la réponse. Choisir entre : **décrire (événements), décrire (processus), raconter, commenter, présenter, expliquer, faire une démonstration, donner des instructions, argumenter, persuader, rapporter des événements, donner des opinions, faire des réclamations, suggérer, comparer et opposer, donner des exemples, évaluer, exprimer des possibilités/probabilités, résumer**, autres (préciser), CECRL, pages 98-101.

24 Le public auquel est censé s'adresser la tâche. Choisir entre : **ami/connaissance, enseignant, employeur, employé(e), comité, commission, entreprise, étudiants, grand public (par exemple des articles de journaux),** autres (préciser)

25 Choisir entre **oral, écrit ou visuel** ou une **combinaison**.

26 Le sujet ou le thème. Choisir entre : **identification personnelle, maison et foyer/environnement, vie quotidienne, congés/loisirs, voyages, relations avec les autres, santé et bien-être, éducation, achats, nourriture et boisson, services, lieux, langue étrangère, temps (météo)**, autre (préciser) CECRL page 45.

27 Les capacités langagières que le candidat doit avoir pour comprendre la consigne et le document déclencheur. Choisir entre : **réception écrite, orale ou les deux**.

29 Les fonctions attendues dans la réponse. Choisir entre : **décrire (événements), décrire (processus), raconter, commenter, présenter, expliquer, faire une démonstration, donner des instructions, argumenter, persuader, rapporter des événements, donner des opinions, faire des réclamations, suggérer, comparer et opposer, donner des exemples, évaluer, exprimer des possibilités/probabilités, résumer**, autres (préciser), CECRL, pages 98-101.

30 La ou les fonctions attendues de la réponse. Choisir entre : référentiel (pour donner des faits « objectifs » sur le monde), émotif (pour décrire l'état émotionnel de l'auteur), conatif (pour persuader le ou les lecteur(s), phatique (pour établir et maintenir un contact social avec le lecteur),

métalinguistique (pour clarifier ou vérifier la compréhension), poétique (écrire avec des buts esthétiques).

31 Le registre que les candidats sont supposés adopter dans leur réponse. Choisir entre : **informel, sans marqueurs linguistiques à informel, sans marqueurs, sans marqueurs à formel, formel**. CECRL pages 93 à 96.

32 Le domaine auquel la réponse attendue est censée appartenir. Choisir entre : personnel, public, professionnel, éducationnel. CECRL page 41.

33 Choisir le niveau CECRL : **A1, A2, B1, B2, C1, C2 CECRL** pages 89 à 93

34 Choisir le niveau CECRL : **A1, A2, B1, B2, C1, C2 CECRL** pages 87 à 89

35 Choisir le niveau CECRL : **A1, A2, B1, B2, C1, C2 CECRL** pages 96 à 98

36 Jusqu'à quel point la tâche est un reflet d'une activité de la vie réelle qu'un candidat pourrait réaliser. Choisir entre : **faible, moyen, fort**.

37 Jusqu'à quel point les schémas d'interaction sont susceptibles de refléter ceux d'une tâche de la vie réelle. Choisir entre : **faible, moyen, fort**.

38 La difficulté de résoudre la tâche d'un point de vue non linguistique. Choisir entre : reproduction d'idées connues, transformation des connaissances.

39 Le type de connaissance extralinguistique requise pour résoudre la tâche. Choisir entre : **domaines de connaissance personnelle/de la vie quotidienne, domaines de connaissance générale/non spécialisée, domaines de connaissance spécialisée (**scientifique, en rapport avec les études, etc.) **une large gamme de domaines de la connaissance.**

40 Décrire les critères d'évaluation portés à la connaissance des candidats, soit avant soit pendant l'examen. Si les critères ne sont pas donnés, indiquer où ils peuvent être consultés.

41 Choisir entre : **impressionniste / holistique, échelle descriptive, échelle analytique**.

42 Indiquer les critères utilisés pour la notation. Choisir entre : **étendue grammaticale, correction grammaticale, étendue lexicale, correction lexicale, cohésion et cohérence, accomplissement de la tâche/** contenu, développement des idées, **orthographe**, autres (préciser).

43 Si la correction est manuelle, il y aura un ou plusieurs correcteurs. Cependant, il se peut que dans certains cas, les réponses donnent lieu à une double ou triple correction par d'autres correcteurs de même statut ou de statut supérieur. Quand c'est le cas, ajouter la mention « + dans des cas particuliers » en fonction du nombre de correcteurs.

44 Les retours d'informations quantitatives transmises de façon régulière (pour l'épreuve de production écrite). Choisir entre : **scores bruts, sous forme de pourcentage, classement des candidats, niveau du CECRL, note spécifique à l'examen, échec/réussite,** autres (préciser).

45 Les retours d'informations qualitatives transmises de façon régulière (pour l'épreuve de production écrite). Choisir entre : **commentaires sur chaque critère d'évaluation, commentaires holistiques**, autres (préciser).

46 Proposer un échantillon de réponse.

47 Une explication ou justification du niveau attribué à l'échantillon.

48 Le niveau (ou note) attribué à cet échantillon.

La grille du CECRL pour les tâches de production écrite v.3.1 (Analyse)

Cette grille a été conçue par un groupe de travail au sein de ALTE dont l'objectif est d'aider les concepteurs d'examen qui utilisent le *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer* et le *Manuel pour relier les examens au CECRL* disponibles auprès de la Division des Politiques linguistiques du conseil de l'Europe.

Deux versions sont disponibles : la grille d'analyse et la grille de présentation (version simplifiée)

La grille d'analyse est destinée à être utilisée dans des ateliers et des séminaires de calibrage.

- Si le but de l'atelier est d'analyser le contenu et les spécifications d'un test, l'étape adéquate est celle de la Spécification (chapitre 4).
- Si la grille est utilisée pour calibrer des échantillons locaux nouveaux, la partie adéquate du manuel est la partie 5.6.

La grille de présentation fournit un rapport descriptif de l'analyse des tâches d'un test, telle qu'elle a été faite dans un exercice de calibrage préalable. Si les grilles complétées sont utilisées pour la description d'échantillons représentatifs, elles peuvent être exploitées lors d'une formation à la standardisation (chapitre 5 de ce manuel).

Echantillons des tâches d'un test

Rapport sur l'analyse de	
Langue cible de ce test	
Niveau cible (CECRL) de ce test	
Numéro/nom de la tâche	

Information générale - le test dans son ensemble

1. La durée du test ou de l'examen dans son ensemble	Minutes
2. L'objectif	Objectif spécifique (préciser)

3. Le contexte /l'arrière plan de l'examen
4. Les candidats
5. La structure du test ou de l'examen

Information générale - l'épreuve de production écrite

6.	Le nombre de tâches dans l'épreuve de production écrite	1	2	3	4 ou plus		
7.	La durée de l'ensemble de l'épreuve	Minutes					
8. L'intégration des capacités		Aucune			Réception écrite		
		Production orale			Réception orale		
		Une combinaison (préciser)					
9.	Le mode de présentation	Manuscrit		Saisie électronique		L'un ou l'autre	
10.	Le niveau CECRL de cette épreuve	A1	A2	B1	B2	C1	C2

11. Le format de l'épreuve de production écrite
12. L'information spécifique – exemple de tâche
13. La distribution des notes
14. L'évaluation de la tâche
15. Le niveau réel
16. Un échantillon de tâche :

Echantillon de tâche à mettre ici

i) Tâche proposée/déclencheur							
17	La langue de la tâche proposée/du déclencheur						
18	Le niveau CEFRL de la tâche proposée / du déclencheur	A1	A2	B1	B2	C1	C2
19	La durée permise ou suggérée pour cette tâche	minutes					
20	Les directives /Conseils	Directif		Semi-directif		Réponse ouverte	
21	Le contenu	Entièrement précisé		Partiellement précisé		Non précisé	

22	Le genre	(lettre domaine professionnel)	lettre (domaine personnel)
		revue	essai
		rédaction	rapport
		récit	projet
		article	fiche
		autre (préciser)	
23	La ou les fonctions rhétoriques de la tâche proposée	décrire (événements)	décrire (processus)
		raconter	commenter
		présenter	expliquer
		faire une démonstration	donner des instructions
		argumenter	persuader
		rapporter des événements	donner des opinions
		faire des réclamations	suggérer
		comparer et opposer	donner des exemples
		évaluer	exprimer des possibilités/probabilités
		exprimer des probabilités	résumer
		autres (préciser)	
24	**Le public attendu**	ami/connaissance	grand public
		employeur	employé(e)
		enseignant	étudiants
		comité	entreprise
		autres (préciser)	

25	Support de la tâche proposée/ du déclencheur	Oral		écrit
		Visuel		Une combinaison
26	**Le thème de la tâche**	identification personnelle		maison et foyer/environnement
		vie quotidienne		congés/loisirs
		voyages		relations avec les autres
		santé et bien-être		éducation
		achats		nourriture et boisson
		services		lieux
		langue étrangère		temps (météo)
		autre (préciser)		
27	L'intégration des capacités langagières dans la tâche proposée	Réception écrite	Réception orale	Une combinaison des deux

ii) Réponse (description de la réponse écrite suscitée par le déclencheur/la tâche proposée)				
28	Le nombre de mots attendus	0-50	51-100	101-150
		151-200	201-250	251-300
		301-350	351-400	Au-delà de 400
29	La fonction rhétorique attendue	décrire (événements)	décrire (processus)	
		raconter	commenter	
		présenter	expliquer	
		faire une démonstration	donner des instructions	
		argumenter	persuader	
		rapporter des événements	donner des opinions	
		faire des réclamations	suggérer	

		comparer et opposer			donner des exemples		
		évaluer			exprimer des possibilités		
		Exprimer des probabilités			résumer		
		autres (préciser)					
30	La ou les fonctions du texte	Référentielle			**Emotive**		
		Conative			Phatique		
		Métalinguistique			poétique		
31	Le registre	informel			sans marqueurs linguistiques à informel		
		sans marqueurs			sans marqueurs à formel		
		formel					
32	Le domaine	Personnel			Public		
		Professionnel			Educationnel		
33	La compétence grammaticale attendue	A1	A2	B1	B2	C1	C2
34	La compétence lexicale attendue	A1	A2	B1	B2	C1	C2
35	La compétence discursive attendue	A1	A2	B1	B2	C1	C2
36	L’authenticité : situationnelle	Faible		Moyenne		Forte	
37	L’authenticité : interactionnelle	Faible		Moyenne		Forte	
38	Le processus cognitif	reproduction d’idées connues					
		transformation des connaissances					
39	La connaissance du contenu requise	générale/non spécialisée			connaissance spécialisée		
		connaissance très spécialisée			large gamme de connaissances		

iii) Evaluation de la tâche		
40	Les critères connus	

41	La méthode d’évaluation de la tâche	impressionniste / holistique	échelle descriptive
		échelle analytique	Avec un système de compensation
		Autre (préciser)	
42	Les critères d’évaluation	étendue grammaticale	correction grammaticale
		étendue lexicale	correction lexicale
		cohésion et cohérence	accomplissement de la tâche/ contenu
		développement des idées	orthographe
		autres (préciser)	
43	Le nombre et la composition des évaluateurs	1	2
		3 ou plus	1 ou plus selon les cas
		2 ou plus selon les cas	Evaluation électronique

iv) Retours d’informations aux candidats			
44	Les retours d’informations quantitatifs	scores bruts	sous forme de pourcentage
		classement des candidats	niveau du CECRL
		note spécifique à l’examen	échec/réussite
		autres (préciser)	
45	Les retours d’informations qualitatifs	commentaires sur chaque critère d’évaluation	
		commentaires holistiques	
		autres (préciser)	

46 Exemple de réponse

47 Commentaire

48 Résultats attribués

Notes : On peut trouver toutes les références au CECRL sur le site de la division des Politiques linguistiques du Conseil de l'Europe.

Les numéros ci-dessous correspondent à ceux des items de la grille.

2. L'objectif du test peut être l'évaluation de la compétence générale, ou d'une compétence sur objectif spécifique. Indiquer l'objectif s'il est spécifique (français pour le droit, allemand pour des objectifs universitaires, etc.).

3. La description de l'arrière plan peut comprendre les raisons pour lesquelles ce test est conçu, une description de l'ensemble des tests dont fait partie ce test, ou d'autres détails de ce type.

4. Décrire le nombre et le profil des candidats (nationalités, âge…).

5. Décrire les autres épreuves du test ou de l'examen (par exemple l'épreuve de production orale, de réception écrite).

6. Au cas où le nombre de tâches dépend des options choisies, le spécifier dans l'introduction (point 5).

8. Les capacités, en plus de la production écrite, qui sont prises en compte dans cette tâche (indépendamment du fait qu'elles soient prises en compte de façon explicite au moment d'évaluer). Choisir entre : aucune, réception écrite, orale, production orale, une combinaison.

9. Sous quel format sont consignées les réponses du candidat. Choisir entre le format écrit, informatique ou les deux.

10. CECRL, chapitre 3

11. La description peut comprendre des informations sur le nombre de parties dans l'épreuve, le type de tâche dans chaque partie, la durée allouée à chaque partie.

12. Il est possible d'inclure une description courte de la tâche à ce niveau. La description peut comprendre les buts de la tâche, ce qu'on demande aux candidats de faire et ce qui est attendu pour pouvoir juger que la tâche est totalement accomplie.

13. Décrire comment les points sont répartis dans cette partie de la tâche et ce que les candidats devraient faire pour obtenir la totalité des points.

14. Expliquer comment la tâche est évaluée (par exemple manuellement, automatiquement), quels outils sont utilisés et quels sont les éléments pris en compte dans la décision du niveau.

15. Décrire les mesures prises pour s'assurer que les tâches de production écrite sont au niveau approprié. Cette description peut comprendre le processus de conception de l'épreuve et le pré-test.

16. Placer ici un échantillon de tâche, y compris la consigne et le document déclencheur.

18. Choisir un niveau du CECRL : A1, A2, B1, B2, C1, C2.

19. Si cela n'est pas précisé, la durée attendue.

20. Indiquer jusqu'à quel point la consigne, le document déclencheur ou la tâche proposée déterminent la nature et le contenu de la réponse. Choisir entre : directif, semi-directif ou réponse ouverte.

21. Le contenu de la réponse est-il précisé dans la consigne ? Choisir entre : précisé ou non précisé.

22. Choisir entre : **lettre (domaine professionnel), lettre (domaine personnel), revue, essai, rédaction, rapport, récit, projet, article, fiche**, autre (préciser).
23. Les fonctions attendues dans la réponse. Choisir entre : **décrire (événements), décrire (processus), raconter, commenter, présenter, expliquer, faire une démonstration, donner des instructions, argumenter, persuader, rapporter des événements, donner des opinions, faire des réclamations, suggérer, comparer et opposer, donner des exemples, évaluer, exprimer des possibilités/probabilités, résumer**, autres (préciser), CECRL, pages 98-101.
24. Le public auquel est censé s'adresser la tâche. Choisir entre : **ami/connaissance, enseignant, employeur, employé(e), comité, commission, entreprise, étudiants, grand public (par exemple des articles de journaux),** autres (préciser)
25. Choisir entre **oral, écrit ou visuel** ou une **combinaison**.
26. Le sujet ou le thème. Choisir entre : **identification personnelle, maison et foyer/environnement, vie quotidienne, congés/loisirs, voyages, relations avec les autres, santé et bien-être, éducation, achats, nourriture et boisson, services, lieux, langue étrangère, temps (météo)**, autre (préciser) CECRL page 45.
27. Les capacités langagières que le candidat doit avoir pour comprendre la consigne et le document déclencheur. Choisir entre : **réception écrite, orale ou les deux**.
29. Les fonctions attendues dans la réponse. Choisir entre : **décrire (événements), décrire (processus), raconter, commenter, présenter, expliquer, faire une démonstration, donner des instructions, argumenter, persuader, rapporter des événements, donner des opinions, faire des réclamations, suggérer, comparer et opposer, donner des exemples, évaluer, exprimer des possibilités/probabilités, résumer**, autres (préciser), CECRL, pages 98-101.
30. La ou les fonctions attendues de la réponse. Choisir entre : référentiel (pour donner des faits « objectifs » sur le monde), émotif (pour décrire l'état émotionnel de l'auteur), conatif (pour persuader le ou les lecteur(s), phatique (pour établir et maintenir un contact social avec le lecteur), métalinguistique (pour clarifier ou vérifier la compréhension), poétique (écrire avec des buts esthétiques).
31. Le registre que les candidats sont supposés adopter dans leur réponse. Choisir entre : **informel, sans marqueurs linguistiques à informel, sans marqueurs, sans marqueurs à formel, formel**. CECRL pages 93 à 96.
32. Le domaine auquel la réponse attendue est censée appartenir. Choisir entre : personnel, public, professionnel, éducationnel. CECRL page 41.
33. Choisir le niveau CECRL : **A1, A2, B1, B2, C1, C2 CECRL** pages 89 à 93
34. Choisir le niveau CECRL : **A1, A2, B1, B2, C1, C2 CECRL** pages 87 à 89
35. Choisir le niveau CECRL : **A1, A2, B1, B2, C1, C2 CECRL** pages 96 à 98
36. Jusqu'à quel point la tâche est un reflet d'une activité de la vie réelle qu'un candidat pourrait réaliser. Choisir entre : **faible, moyen, fort**.
37. Jusqu'à quel point les schémas d'interaction sont susceptibles de refléter ceux d'une tâche de la vie réelle. Choisir entre : **faible, moyen, fort**.
38. La difficulté de résoudre la tâche d'un point de vue non linguistique. Choisir entre : reproduction d'idées connues, transformation des connaissances.
39. Le type de connaissance extra-linguistique requise pour résoudre la tâche. Choisir entre : **domaines de connaissance personnel/de la vie quotidienne, domaines de connaissance**

générale/non spécialisée, domaines de connaissance spécialisée (scientifique, en rapport avec les études, etc.) **une large gamme de domaines de la connaissance.**

40. Décrire les critères d'évaluation portés à la connaissance des candidats, soit avant soit pendant l'examen. Si les critères ne sont pas donnés, indiquer où ils peuvent être consultés.
41. Choisir entre : **impressionniste / holistique, échelle descriptive, échelle analytique**.
42. Indiquer les critères utilisés pour la notation. Choisir entre : **étendue grammaticale, correction grammaticale, étendue lexicale, correction lexicale, cohésion et cohérence, accomplissement de la tâche/** contenu, développement des idées, **orthographe**, autres (préciser).
43. Si la correction est manuelle, il y aura un ou plusieurs correcteurs. Cependant, il se peut que dans certains cas, les réponses donnent lieu à une double ou triple correction par d'autres correcteurs de même statut ou de statut supérieur. Quand c'est le cas, ajouter la mention « + dans des cas particuliers » en fonction du nombre de correcteurs.
44. Les retours d'informations quantitatives transmises de façon régulière (pour l'épreuve de production écrite). Choisir entre : **scores bruts, sous forme de pourcentage, classement des candidats, niveau du CECRL, note spécifique à l'examen, échec/réussite,** autres (préciser).
45. Les retours d'informations qualitatives transmises de façon régulière (pour l'épreuve de production écrite). Choisir entre : **commentaires sur chaque critère d'évaluation, commentaires holistiques**, autres (préciser).
46. Proposer un échantillon de réponse.
47. Une explication ou justification du niveau attribué à l'échantillon.
48. Le niveau (ou note) attribué à cet échantillon.

La grille du CECRL pour les tâches de production orale v.3.1 (Présentation)

Cette grille a été conçue par un groupe de travail au sein de ALTE dont l'objectif est d'aider les concepteurs d'examen qui utilisent le *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer* et le *Manuel pour relier les examens au CECRL* disponibles auprès de la Division des Politiques linguistiques du conseil de l'Europe.

Deux versions sont disponibles : la grille d'analyse et la grille de présentation (version simplifiée)

La grille d'analyse est destinée à être utilisée dans des ateliers et des séminaires de calibrage.

- Si le but de l'atelier est d'analyser le contenu et les spécifications d'un test, l'étape adéquate est celle de la Spécification (chapitre 4).
- Si la grille est utilisée pour calibrer des échantillons locaux nouveaux, la partie adéquate du manuel est la partie 5.6.

La grille de présentation fournit un rapport descriptif de l'analyse des tâches d'un test, telle qu'elle a été faite dans un exercice de calibrage préalable. Si les grilles complétées sont utilisées pour la description d'échantillons représentatifs, elles peuvent être exploitées lors d'une formation à la standardisation (chapitre 5 de ce manuel).

1.	Rapport sur l'analyse de	
2.	Langue cible	

1. INFORMATION GENERALE (le test de production orale dans son ensemble)

3.	Le nombre de tâches dans l'épreuve de production orale	
4.	L'intégration des capacités	
5.	La durée de l'ensemble de l'épreuve	
6.	Le niveau cible de la performance	
7.	Le mode de présentation	
8.	L'objectif du test	

2. **TACHE PROPOSEE/DECLENCHEUR pour la tâche n° / nom**

9	La langue de la consigne	
10	Le mode de présentation	
11	Le niveau de langue de la consigne	
12	La durée de la tâche (minutes)	
13	Le nombre d'examinateurs présents	
14	Enregistré ?	
15	Directives / conseils par tâche	
16	Directives / conseils par interlocuteur	
17	**Spécification du contenu**	
18	Type d'interaction	
19	Type de discours	
20	Public (réel)	
21	Public imaginé (comme dans un jeu de rôle)	
22	Type de déclencheur	
23	Thème	
24	Organisation du temps	
25	Définition de la situation fictive	

3. REPONSE (la réponse orale suscitée par le déclencheur/la tâche proposée

26	Longueur de la réponse	
27	Type de texte	
28	Fonction rhétorique	
29	Registre	
30	Domaine	
31	Niveau grammatical	
32	Niveau lexical	
33	Compétence discursive attendue	
34	Authenticité : situationnelle	

35	Authenticité : interactionnelle	
36	Processus cognitif	
37	Connaissance du contenu	
38	Fonction de la tâche	

4. EVALUATION DE LA TACHE

39	Critères connus	
40	Méthode d'évaluation de la tâche	
41	Critères d'évaluation	
42	Nombre des évaluateurs	
43	Présence d'un modérateur	

5. RETOURS D'INFORMATIONS AUX CANDIDATS

44	Les retours d'informations quantitatifs	
45	Les retours d'informations qualitatifs	

La grille du CECRL pour les tâches de production orale v.3.1 (Analyse)

Cette grille a pour but de donner de l'information sur une seule tâche du test ou de l'examen étudié. Le tableau INFORMATION GENERALE (partie 1) traite du test de production orale dans son ensemble. Les autres parties se réfèrent à une seule tâche de ce test.

Pour les définitions (et les traductions) de la terminologie, les utilisateurs se réfèreront au *Glossaire multilingue des termes de l'évaluation* de ALTE (Cambridge University Press).

1. INFORMATION GENERALE (le test de production orale dans son ensemble)

0	Nom de l'organisme certificateur qui a conçu le test				
1	Intitulé du test ou de l'examen				
	Epreuves	**Epreuve de production orale**			
2	Langue cible				
3	Nombre de tâches dans l'épreuve de production orale	1	2	3	4 ou plus
4	Intégration des capacités[46] (entourer au moins une case)	Production orale (seule)	Réception écrite	Production écrite	Réception orale
	Commentaire				

[46] Degré auquel l'épreuve de production orale fait appel à une autre compétence langagière. Cette intégration est-elle explicite ou implicite ? N'oubliez pas que même un déclencheur écrit implique un degré d'intégration d'une compétence autre que la production orale qui peut ou non être prise en compte au moment de l'évaluation.

5	Durée totale de l'épreuve (y compris le temps de préparation	Environ …… minutes (dont …….. minutes de préparation)											
6	Niveau cible de performance. Production orale générale pp. 25 et 49 du CECRL et annexe D des « Être capable de » de ALTE, p. 244 (entourer au moins une case)	A1		A2		B1		**B2**		C1		C2	
7	Type de passation	Face à face	Téléphone	Ordinateur		Vidéo conférence		Magnétophone		Caméscope			
				Audio	Vidéo								
8	Objectif du test	Compétence générale				Objectif spécifique (langue sur objectif spécifique)							

Chaque tableau suivant (partie 2 – 6) doit être complété pour chacune des tâches du test (autant de tableaux que de tâches).

2. **TACHE PROPOSEE/DECLENCHEUR –** Consignes et déclencheurs (verbaux ou iconographiques) ou toute autre forme de tâche destinée à faire produire la réponse attendue dans la langue cible.

0	**Intitulé de la tâche dans l'épreuve de production orale**							
9	Langue de la consigne	Langue de l'organisme certificateur			Langue cible du test		Autre langue ?	
10	Mode de transmission de la consigne : oral ou écrit	Oral	Ecrit		Enregistré		Illustré	

11	Niveau de langue de la consigne	Bien plus facile que le niveau du test		Plus facile que le niveau du test	Même niveau que celui du test	Plus difficile que le niveau du test
12	Durée de la tâche (minutes)	Environ …… minutes				
13	Nombre d'examinateurs présents	0		1	2	
14	Enregistré ?	Oui – audio		Oui – vidéo	Non	
15	Directives / conseils par tâche (format de la tâche[47])	Très directif		Partiellement directif	Réponse ouverte	
16	Directives / conseils par interlocuteur (souplesse de l'examinateur[48])	Très directif (par exemple liste de questions à poser)		Partiellement directif (par exemple entretien sur un thème donné)	Réponse ouverte (par exemple un entretien ou une discussion non directifs)	
17	Spécification du contenu	Spécifique			Non spécifique	
18	Type d'interaction	Dialogue : 2 candidats	Dialogue : plusieurs candidats	Dialogue : candidat et examinateur	Dialogue : simulé, avec déclencheur enregistré	Monologue
		Répétition du déclencheur	Jeu de rôle	Lecture à voix haute	Réaction à un déclencheur	Autre :
19	Type de discours	Entretien		Narration (raconter une histoire)		
		Discours, exposé		Discussion / conversation		
20	Public (réel)	Examinateur	Autre candidat	Professeur	Aucun (magnéto-phone)	Autre :

[47] Degré auquel le format de la tâche guide ou limite la réponse du candidat.
[48] Degré auquel le candidat maîtrise le format de la tâche proposée par l'examinateur, ayant un impact sur la nature et le contenu de l'interaction. La production peut être en grande partie non dirigée, sous forme de conversation spontanée. Le contenu de la réponse attendue est-il spécifié par l'examinateur ?

21	Public imaginé (comme dans un jeu de rôle)	Employeur	Comité Commission	Entreprise, magasin, etc.	Professeur	Répondeur
		Grand public	Membre de la famille	Ami ou connaissance	Autre (préciser)	

22	Type de déclencheur (sélectionner au moins une case)	Uniquement oral (donné oralement par l'examinateur)		
		Texte (écrit)	Phrases, questions, instructions	
			Lettres	Par exemple à un correspondant
			Notes, messages, memos,	Exemple « post-it »
			publicités	
			Programmes	Exemple : théâtre, football, etc.
			Formulaires	Exemple : à remplir pour l'immigration
			Extraits	Livres/journaux/ magazines
		Iconographique	Graphique	Annotés ou non
			Tableau	
			Schéma	
			Diagramme	
			Carte	
			Suite de diagrammes	
		Illustrations (non verbal)	photos	
			dessins	
			Suite de dessins	

		Autre (préciser) :			
23	Thème CECRL p. 43 (sélectionner au moins une case)	Identification personnelle		Affaires courantes	
		Maison, foyer, environnement		Courses, achats	
		Vie quotidienne		Nourriture et boissons	
		Loisirs et divertissements		Services	
		Voyages		Lieux	
		Relations avec les autres		Langues	
		Santé et bien-être		Temps (météo)	
		Education		Célébrités	
		Sciences et environnement		Environnement professionnel	
		Autres (préciser) :			
24	Gestion du temps	30 secondes	1 minute	2 minutes	Sans objet
				Commentaire :	
25	Définition de la situation fictive	Lieu professionnel	Environnement social	Environnement éducationnel	Autre :

3. REPONSE (la réponse orale attendue, suscitée par le déclencheur/la tâche proposée)

26	Longueur de la réponse	30 sec	1 min	2 min	3 min	4 min	5 min	Au-delà de 5 min
27	Type de texte	Niveau de vocabulaire		Phrase		Niveau du discours		
28	Fonction rhétorique, CECRL p. 98	Décrire (événements) Décrire (processus) Décrire (données) Décrire (objets) Décrire (images) Raconter Commenter Exposer		Donner des instructions Argumenter Persuader Rapporter des événements Donner son opinion Se plaindre		Donner des exemples Faire une synthèse Analyser Evaluer Exprimer la possibilité / la probabilité Résumer Demander des		

		Expliquer Démontrer	Suggérer Comparer et opposer	informations autres : (préciser)	
29	Registre, CECRL p.94	Informel	Neutre	Formel	
30	Domaine, CECRL p.43	Personnel	Public	Professionnel	Educationnel
31	Niveau grammatical, CECRL p. 89	Uniquement structures simples	Essentiellement structures simples	Gamme réduite de structures complexes	Large gamme de structures complexes
32	Niveau lexical, CECRL p. 87	Uniquement vocabulaire fréquent	Essentiellement vocabulaire fréquent	Vocabulaire étendu	Large gamme de vocabulaire diversifié	Large gamme de vocabulaire diversifié et spécialisé
33	Compétence discursive (par exemple cohésion), CECRL p. 98	Usage extrêmement limité	Limité	Usage maîtrisé	Excellent usage
34	Authenticité : situationnelle[49]	Faible	Moyenne	Forte	
35	Authenticité : interactionnelle	Faible	Moyenne	Forte	
36	Processus cognitif[50]	Uniquement reproduction des idées connues	Transformation des connaissances		

[49] Degré auquel la tâche renvoie à une activité de la vie réelle que le candidat est susceptible d'accomplir.
[50] Quelle est la difficulté de la tâche, d'un point de vue non linguistique ? Exemple : la difficulté pour un candidat d'interpréter des déclencheurs présentés sous forme de graphique, s'il n'y est pas habitué.

37	Connaissance du contenu	Personnel / vie quotidienne / besoins liés à une communication de base	Commun, général, non spécifique	Large gamme de domaines de connaissances non spécifiques	Très large gamme de connaissances (sociales, scientifiques, éducationnelle et parfois spécifiques, etc.)
38	Fonction de la tâche	Référentielle (raconter)	Emotive (réagir)	Conative[51]	Phatique[52]

4. EVALUATION DE LA TACHE

39	Critères connus	Les critères d'évaluation sont-ils disponibles sur la feuille d'examen ? Les candidats sont-ils habitués à ces critères ? Oui / Non. Si « non », où peut-on les consulter ?					
40	Méthode d'évaluation de la tâche	Impressio-niste/holistique		Echelle descriptive (descripteurs par niveaux)		Méthode analytique	
41	Critères d'évaluation	Correction grammaticale	Cohésion et cohérence	Maîtrise du vocabulaire	Conte nus	Communi cation interactive	Développem ent des idées
		Prononciation (phonologie)		Prononciation (intonation et mélodie)		Autre :	
42	Nombre des évaluateurs	1	2	3	Evaluation électronique		
		Autre (expliquer) :					
43	Présence d'un modérateur[53]	Oui		Non			

[51] Conative : renvoie aux tâches qui supposent que le candidat argumente, persuade, discute du pour et du contre, etc.
[52] Phatique : qui a pour but de garder le contact avec l'interlocuteur.
[53] Le modérateur vérifie que les critères d'évaluation sont respectés de façon cohérente et s'assure que les notes sont attribuées de façon correcte et juste par les examinateurs.

5. RETOURS D'INFORMATIONS AUX CANDIDATS

44	Les retours d'information quantitatifs[54] (Cocher)	Score brut	Score en %	Classement (exemple : quartile)	Niveau du CECRL	Note spécifique à l'examen	Echec / réussite unique ment	Autre :
		☐	☐	☐	☐	☐	☐	☐
45	Les retours d'informations qualitatifs (Cocher)	Gram maire	Lexique	Cohésion cohérence	Contenu	Développe ment des idées	Pertinen ce de la tâche	Autre :
		☐	☐	☐	☐	☐	☐	☐

[54] Information sur leurs performances données aux candidats.

Annexe C

Fiches et échelles pour la standardisation et le calibrage (chapitre 5)

Fiche de rapport de formation		
Lieu		**Date :**
Coordinateur	Nom :	Institution/projet
Etape	Familiarisation Formation Calibrage	☐ ☐ ☐
Domaines	Echantillons d'évaluation de la production orale Echantillons d'évaluation de la production écrite Tâches/items du test Réception orale Réception écrite Compétence linguistique Autres :	☐ ☐ ☐ ☐ ☐
Participants	**Nombre :**	**Fonctions :**
Activités accomplies	Familiarisation Travail avec des exemples représentatifs Pratique libre/dirigée avec des exemples représentatifs Calibrage avec des échantillons de performances locales Formation avec des tâches représentatives Evaluation de la difficulté de l'item Retour d'information sur la difficulté réelle de l'item Autre	☐ ☐ ☐ ☐ ☐ ☐ ☐
Matériel utilisé	Echantillons d'exemples représentatifs du CECRL Outils d'évaluation du CECRL (tableaux 5.4,5.5,5.8) Echantillons de performances locales Outils d'évaluation adaptés (à joindre) Exemples du CECRL de tâches et d'items de tests Tâches et items de tests locaux Autres	☐ ☐ ☐ ☐ ☐ ☐
Informations sur les tâches et les items		
Commentaires complémentaires		
Diffusion des procédures prévues		

Fiche C1 Fiche de rapport de formation

NOM DE L'APPRENANT _____

Niveaux : A1, A2, A2+, B1, B1+, B2, B2+, C1, C2

1. Impression initiale

Classement - échelle globale

2. Analyse détaillée / Estimation à l'aide de la grille

ETENDUE	CORRECTION	AISANCE	INTERACTION	COHERENCE

3. Classement final

Fiche C2 : Fiche analytique d'évaluation
Eurocentres (North 1991/1992)/Projet suisse (Schneider and North 2000)

Capacité : _____	Niveau estimé	Commentaires
Echantillon / Tâche 1		
Echantillon / Tâche 2		
Echantillon / Tâche 3		
Echantillon / Tâche 4		
Echantillon / Tâche 5		
Echantillon / Tâche 6		
Echantillon / Tâche 7		
Echantillon / Tâche 8		
Exemple d'une fiche d'évaluation simple qui demande au participant d'évaluer globalement le niveau de chaque échantillon ou tâche. On peut l'utiliser pour évaluer des performances ou des items de tests.		

Fiche C3 : Fiche d'évaluation globale (DIALANG)

	Astérix	*Idéfix*	*Bécassine*	*Tintin*	*Henri IV*	*Hercule Poirot*	*Autre nom de code*	*Autre nom de code*	*Autre nom de code*
Item 1									
Item 2									
Item 3									
Item 4									

Fiche C4 : Fiche de synthèse de l'évaluation globale (DIALANG)

Capacité : _____	**Descripteur opérationnel** **(faire la liste des sous-échelles et les niveaux)**	Niveau CECRL estimé	**Commentaires**
Item 1			
Item 2			
Item 3			
Item 4			
Item 5			
Item 6			
etc			

Fiche C5 : Fiche d'évaluation des items (DIALANG)

TABLEAU C1: ECHELLE GLOBALE D'EVALUATION DE LA PRODUCTION ORALE

C2	***Transmet, avec naturel et précision des nuances de sens subtiles.*** Est capable de s'exprimer spontanément et avec beaucoup d'aisance, de communiquer facilement avec habileté et de discriminer avec précision des nuances de sens subtiles. Peut produire des descriptions claires, régulières et bien structurées.
C1	***S'exprime spontanément et avec aisance dans un discours clair et bien structuré.*** Est capable de s'exprimer spontanément et avec aisance, presque sans effort, dans un discours régulier. Peut faire des descriptions claires et détaillées de sujets complexes. Niveau élevé de correction ; les erreurs sont rares.
B2+	
B2	***Exprime ses opinions sans effort notable***. Est capable de communiquer sur une gamme étendue de sujets et de produire des énoncés sur un rythme assez régulier. Peut faire des descriptions claires et détaillées sur une vaste étendue de sujets relatifs à son centre d'intérêt. Ne commet pas de fautes qui provoquent des malentendus.
B1+	
B1	***Rapporte de façon compréhensible ce qu'il/elle tient à dire.*** Est capable de tenir un discours compréhensible même si les pauses pour rechercher des mots ou des phrases ainsi que la remédiation sont très évidentes. Peut relier des éléments discrets simples en un paragraphe articulé pour faire des descriptions simples sur des sujets familiers variés propres à son domaine. Utilisation assez juste d'un répertoire essentiel associé aux situations les plus prévisibles.
A2+	
A2	***Rapporte des informations de base sur, par exemple, le travail, la famille, les loisirs, etc.*** Est capable de communiquer dans un échange simple et direct d'informations sur des sujets courants. Peut se faire comprendre dans de très courts énoncés même si les pauses, les hésitations et la reformulation sont très évidentes. Peut décrire en termes simples ses conditions de vie, ses études, son dernier métier ou son métier actuel. Utilise correctement des structures simples mais peut commettre systématiquement des erreurs élémentaires.
A1	***S'exprime de façon simple sur des détails personnels et des sujets très familiers.*** Est capable de se faire comprendre de façon simple, de poser des questions sur des détails personnels et d'y répondre à condition que l'interlocuteur parle lentement et clairement et soit prêt à aider. Peut se débrouiller avec des énoncés très courts, isolés, le plus souvent stéréotypés. De nombreuses pauses pour chercher ses mots et prononcer les moins familiers.
Au-dessous de A1	N'atteint pas la norme A1
• ***Utiliser cette échelle pour les deux ou trois premières minutes d'un échantillon de production orale afin de décider approximativement du niveau auquel on pense que le locuteur se trouve.*** • ***Puis passer au Tableau C2 (Tableau 3 du CECRL) et évaluer plus en détail la performance par rapport aux descripteurs de ce niveau.***	

TABLEAU C2: GRILLE DES CRITERES D'EVALUATION DE L'ORAL
(CECRL Tableau 3)

	ÉTENDUE	CORRECTION	AISANCE	INTERACTION	COHÉRENCE
C2	Montre une grande souplesse dans la reformulation des idées sous des formes linguistiques différentes lui permettant de transmettre avec précision des nuances fines de sens afin d'insister, de discriminer ou de lever l'ambiguïté. A aussi une bonne maîtrise des expressions idiomatiques et familières.	Maintient constamment un haut degré de correction grammaticale dans une langue complexe, même lorsque l'attention est ailleurs (par exemple, la planification ou l'observation des réactions des autres).	Peut s'exprimer longuement, spontanément dans un discours naturel en évitant les difficultés ou en les rattrapant avec assez d'habileté pour que l'interlocuteur ne s'en rende presque pas compte.	Peut interagir avec aisance et habileté en relevant et utilisant les indices non verbaux et intonatifs sans effort apparent. Peut intervenir dans la construction de l'échange de façon tout à fait naturelle, que ce soit au plan des tours de parole, des références ou des allusions, etc.	Peut produire un discours soutenu cohérent en utilisant de manière complète et appropriée des structures organisationnelles variées ainsi qu'une gamme étendue de mots de liaisons et autres articulateurs.
C1	A une bonne maîtrise d'une grande gamme de discours parmi lesquels il peut choisir la formulation lui permettant de s'exprimer clairement et dans le registre convenable sur une grande variété de sujets d'ordre général, éducationnel, professionnel ou de loisirs, sans devoir restreindre ce qu'il/elle veut dire.	Maintient constamment un haut degré de correction grammaticale ; les erreurs sont rares, difficiles à repérer et généralement auto-corrigées quand elles surviennent.	Peut s'exprimer avec aisance et spontanéité presque sans effort. Seul un sujet conceptuellement difficile est susceptible de gêner le flot naturel et fluide du discours.	Peut choisir une expression adéquate dans un répertoire courant de fonctions discursives, en préambule à ses propos, pour obtenir la parole ou pour gagner du temps pour la garder pendant qu'il/elle réfléchit.	Peut produire un texte clair, fluide et bien structuré, démontrant un usage contrôlé de moyens linguistiques de structuration et d'articulation.
B2+					
B2	Possède une gamme assez étendue de langue pour pouvoir faire des descriptions claires, exprimer son point de vue et développer une argumentation sans chercher ses mots de manière évidente.	Montre un degré assez élevé de contrôle grammatical. Ne fait pas de fautes conduisant à des malentendus et peut le plus souvent les corriger lui/elle-même.	Peut parler relativement longtemps avec un débit assez régulier ; bien qu'il /elle puisse hésiter en cherchant structures ou expressions, l'on remarque peu de longues pauses.	Peut prendre l'initiative de la parole et son tour quand il convient et peut clore une conversation quand il le faut, encore qu'éventuellement sans élégance. Peut faciliter la poursuite d'une discussion sur un terrain familier en confirmant sa compréhension, en sollicitant les autres, etc.	Peut utiliser un nombre limité d'articulateurs pour lier ses phrases en un discours clair et cohérent bien qu'il puisse y avoir quelques "sauts" dans une longue intervention.
B1+					
B1	Possède assez de moyens linguistiques et un vocabulaire suffisant pour s'en sortir avec quelques hésitations et quelques périphrases sur des sujets tels que la famille, les loisirs et centres d'intérêt, le travail, les voyages et l'actualité	Utilise de façon assez exacte un répertoire de structures et "schémas" fréquents, courants dans des situations prévisibles.	Peut discourir de manière compréhensible, même si les pauses pour chercher ses mots et ses phrases et pour faire ses corrections sont très évidentes, particulièrement dans les séquences plus longues de production libre.	Peut engager, soutenir et clore une conversation simple en tête-à-tête sur des sujets familiers ou d'intérêt personnel. Peut répéter une partie de ce que quelqu'un a dit pour confirmer une compréhension mutuelle.	Peut relier une série d'éléments courts, simples et distincts en une suite linéaire de points qui s'enchaînent.
A2+					
A2	Utilise des structures élémentaires constituées d'expressions mémorisées, de groupes de quelques mots et d'expressions toutes faites afin de communiquer une information limitée dans des situations simples de la vie quotidienne actualité.	Utilise des structures simples correctement mais commet encore systématiquement deserreurs élémentaires.	Peut se faire comprendre dans une brève intervention même si la reformulation, les pauses et les faux démarrages sont évidents.	Peut répondre à des questions et réagir à des déclarations simples. Peut indiquer qu'il/elle suit mais est rarement capable de comprendre assez pour soutenir la conversation de son propre chef.	Peut relier des groupes de mots avec des connecteurs simples tels que "et", "mais" et "parce que".

A1	Possède un répertoire élémentaire de mots et d'expressions simples relatifs à des situations concrètes particulières	A un contrôle limité de quelques structures syntaxiques et de formes grammaticales simples appartenant à un répertoire mémorisé	Peut se débrouiller avec des énoncés très courts, isolés, généralement stéréotypés, avec de nombreuses pauses pour chercher ses mots, pour prononcer les moins familiers et pour remédier à la communication.	Peut répondre à des questions simples et en poser sur des détails personnels. Peut interagir de façon simple, mais la communication dépend totalement de la répétition avec un débit plus lent, de la reformulation et des corrections.	Peut relier des mots ou groupes de mots avec des connecteurs très élémentaires tels que "et" ou "alors".

TABLEAU C3 : GRILLE DES CRITERES SUPPLEMENTAIRES : NIVEAUX PLUS

	ÉTENDUE	CORRECTION	AISANCE	INTERACTION	COHÉRENCE
C2					
C1					
B2+	Peut s'exprimer clairement et sans donner l'impression d'avoir à restreindre ce qu'il/elle souhaite dire.	A un bon contrôle grammatical ; des bévues occasionnelles, des erreurs non systématiques et de petites fautes syntaxiques peuvent encore se produire mais elles sont rares et peuvent souvent être corrigées rétrospectivement.	Peut communiquer avec spontanéité, montrant souvent une remarquable aisance et une facilité d'expression même dans des énoncés complexes assez longs. Peut recourir à des circonlocutions et des paraphrases pour masquer des lacunes lexicales ou grammaticales.	Peut intervenir de manière adéquate dans une discussion, en utilisant des moyens d'expression appropriés et peut relier habilement sa propre contribution à celle d'autres interlocuteurs.	Peut utiliser avec efficacité une grande variété de mots de liaison pour marquer clairement les relations entre les idées.
B2				.	
B1+	Possède une gamme assez étendue de langue pour décrire des situations imprévisibles, expliquer le point principal d'un problème ou d'une idée avec assez de précision et exprimer sa pensée sur des sujets abstraits ou culturels tels que la musique ou le cinéma.	Communique avec une correction suffisante dans des contextes familiers ; en règle générale, a un bon contrôle grammatical malgré de nettes influences de la langue maternelle.	Peut s'exprimer avec une certaine aisance. Malgré quelques problèmes de formulation ayant pour conséquence pauses et impasses, est capable de continuer effectivement à parler sans aide.	Peut exploiter un répertoire élémentaire de langue et de stratégies pour faciliter la suite de la conversation ou de la discussion. Peut faire de brefs commentaires sur les points de vue des autres pendant une discussion. Peut intervenir pour vérifier et confirmer le détail d'une information.	Pas de descripteur disponible
B1		.			
A2 +	Possède un répertoire de langue élémentaire qui lui permet de se débrouiller dans des situations courantes au contenu prévisible, bien qu'il lui faille généralement chercher ses mots et trouver un compromis par rapport à ses intentions de communication.	Pas de descripteur disponible	Peut adapter des phrases simples répétées et mémorisées à des situations particulières avec suffisamment d'aisance pour se débrouiller dans des échanges de routine sans effort excessif, malgré des hésitations et des faux démarrages évidents.	Peut commencer, poursuivre et terminer une simple conversation en tête-à-tête sur des sujets familiers ou d'intérêt personnel, passe-temps, et activités passées. Peut interagir avec suffisamment d'aisance dans des situations structurées, à condition d'être aidé, mais la participation à une discussion libre est assez restreinte.	Peut utiliser les articulations les plus fréquentes pour relier des énoncés afin de raconter une histoire ou décrire quelque chose sous forme d'une simple liste de points.
A2					
A1					

TABLEAU C4 : GRILLE DES CRITERES D'EVALUATION DE LA PRODUCTION ECRITE

	Vue d'ensemble	Etendue	Cohérence	Correction	Description	Argumentation
C2	Est capable d'écrire des textes élaborés, limpides, fluides et parfaitement corrects dans un style personnel approprié et efficace et qui transmette des nuances fines de sens. Peut utiliser une structure logique qui aide le destinataire à remarquer les points importants.	Manifeste une grande souplesse pour formuler des idées sous des formes linguistiques différentes afin de transmettre avec précision des nuances fines de sens, pour insister et pour lever l'ambiguïté. Possède aussi une bonne maîtrise d'expressions idiomatiques et familières.	Est capable de créer des textes cohérents et articulés en faisant un usage complet et adéquat d'une variété de modèles d'organisation et un choix étendu de connecteurs et autres articulateurs.	Garde une maîtrise cohérente et extrêmement juste des formes de la langue même les plus complexes. Les fautes sont rares et portent sur des formes rarement utilisées.	Est capable de rédiger des histoires claires, fluides et très intéressantes ainsi que de décrire des expériences dans un style approprié avec le genre choisi.	Est capable de produire des comptes rendus, des articles et des essais élaborés, clairs et fluides pour présenter un cas ou donner une appréciation critique de propositions ou d'œuvres littéraires. Peut fournir une structure logique efficace et appropriée qui aide le lecteur à trouver les points importants.
C1	Est capable de rédiger des textes bien structurés et corrects dans l'ensemble sur des sujets complexes. Peut souligner les points pertinents les plus saillants, étendre et confirmer des points de vue de manière élaborée par l'intégration d'arguments secondaires, de justifications et d'exemples pertinents pour parvenir à une conclusion appropriée	A une bonne maîtrise d'une gamme étendue de langue qui lui permet de s'exprimer clairement dans un style approprié sur une vaste étendue de sujets généraux, académiques, professionnels ou de loisirs sans devoir limiter ce qu'il/elle veut dire. La souplesse de style et de ton est un peu limitée.	Est capable de produire des textes clairs et fluides, bien structurés, montrant un usage maîtrisé de modèles d'organisation, de connecteurs et autres articulateurs.	Garde constamment un niveau élevé de correction grammaticale ; fautes occasionnelles en grammaire, formes familières et idiomatiques.	Est capable de rédiger des descriptions et des textes créatifs clairs, détaillés et bien construits dans un style personnel, naturel et affirmé approprié au lecteur visé.	Est capable d'exposer clairement par écrit de façon bien structurée des sujets complexes en relevant les points saillants importants. Peut étendre et confirmer des points de vue de manière élaborée par l'intégration d'arguments secondaires, de justifications et d'exemples pertinents.
B2	Est capable de rédiger des textes détaillés officiels ou pas sur une gamme étendue de sujets relatifs à son domaine d'intérêt en faisant la synthèse et l'évaluation d'informations et d'arguments empruntés à des sources diverses. Peut faire la différence entre un discours formel ou pas avec de temps à autre des expressions moins appropriées.	Possède une étendue de langue suffisante pour pouvoir faire des descriptions claires, exprimer des opinions sur les sujets les plus généraux en utilisant des formes grammaticales complexes pour le faire. Néanmoins, le discours manque d'expressivité et l'utilisation de formes plus élaborées reste stéréotypée.	Est capable d'utiliser un nombre limité d'articulateurs pour relier ses phrases en un texte clair et cohérent bien qu'il puisse y avoir quelques « sauts » dans un texte un peu long.	Manifeste un degré relativement élevé de maîtrise de la grammaire. Ne commet pas de fautes qui causent des malentendus.	Est capable d'écrire des descriptions claires et détaillées d'événements réels ou imaginaires en établissant la relation entre des idées clairement articulées et en suivant les conventions en vigueur du genre en question. Peut faire des descriptions claires et détaillées sur un certain nombre de sujets relatifs à son centre d'intérêt. Peut écrire le compte rendu d'un film, d'un livre ou d'une pièce.	Est capable de rédiger un essai ou un rapport qui développe systématiquement une argumentation avec un éclairage approprié des points importants ainsi que des détails secondaires pertinents. Peut évaluer des idées ou des solutions différentes à un problème. Peut écrire un essai ou un rapport qui développe une argumentation, justifier ou rejeter une opinion particulière et expliquer les avantages et les inconvénients de choix variés. Peut faire la synthèse d'arguments et d'informations empruntés à des sources diverses.
B1	Est capable de rédiger des textes articulés simplement sur une gamme de sujets variés dans son domaine d'intérêt en liant une série d'éléments discrets en une séquence linéaire. Les textes sont compréhensibles bien que quelques expressions obscures et/ou des incohérences puissent provoquer une rupture de la lecture.	Possède une langue suffisante pour se débrouiller avec un vocabulaire suffisant pour s'exprimer avec quelques périphrases sur des sujets tels que la famille, les loisirs et les centres d'intérêt, le travail, les voyages et l'actualité.	Est capable de lier une série d'éléments discrets courts en un texte linéaire articulé.	Utilise de façon raisonnablement correcte un répertoire de clichés et d'expressions associés aux situations les plus courantes. Commet occasionnellement des erreurs que le lecteur peut habituellement interpréter correctement en s'appuyant sur le contexte.	Est capable de rendre compte d'expériences, de décrire des sentiments et des réactions dans des textes simplement articulés. Peut faire la description d'un événement, d'un voyage récent – réel ou imaginaire. Peut raconter une histoire. Peut faire des descriptions sur un certain nombre de sujets courants proches de son centre d'intérêt.	Est capable d'écrire de courts essais simples sur des sujets d'intérêt général. Peut résumer avec quelque assurance des informations factuelles nombreuses, en rendre compte et donner son opinion sur des sujets courants ou pas, dans son domaine. Peut écrire sous une forme classique des comptes rendus très courts pour transmettre des informations factuelles courantes et justifier des actions.
A2	Est capable de rédiger une série d'expressions et de phrases simples reliées par des connecteurs simples tels qu « et », « mais » et « parce que ». Des textes plus longs peuvent contenir des expressions ainsi que des problèmes de cohérence qui rendent le texte difficile à comprendre.	Utilise des formules de base d'expressions toutes faites mémorisées, de groupes de quelques mots et expressions afin de communiquer une information limitée dans des situations simples de la vie quotidienne.	Est capable de lier des mots avec des connecteurs simples tels qu « et », « mais » et « parce que ».	Utilise correctement des structures simples mais commet encore systématiquement des fautes élémentaires. Les erreurs peuvent quelquefois provoquer des malentendus.	Est capable d'écrire de brèves biographies simples et des poèmes simples sur les gens. Peut écrire des descriptions élémentaires très brèves d'événements, d'activités passées et d'expériences personnelles.	
A1	Est capable d'écrire des expressions et phrases simples isolées. Des textes plus longs peuvent contenir des expressions ainsi que des problèmes de cohérence qui rendent le texte très difficile, voire impossible à comprendre.	Possède un répertoire élémentaire de mots et d'expressions simples relatives à des questions personnelles et à des situations concrètes particulières.	Est capable de lier des mots ou groupes de mots avec des connecteurs très élémentaires tels que « et » et « alors ».	Ne montre qu'une maîtrise limitée de quelques structures grammaticales simples et de clichés mémorisés. Les erreurs peuvent provoquer des malentendus.	Est capable d'écrire des phrases et des expressions simples sur des gens réels ou imaginaires, où ils vivent et ce qu'ils font.	