

Positionnement du *TFI*TM Test sur le Cadre Européen Commun de Référence pour les Langues

***Patricia A. Baron
Richard J. Tannenbaum***

Septembre 2010

ETS RM-10-12

Positionnement du TFI™
sur le Cadre Européen Commun De Référence Pour Les Langues

Patricia A. Baron et Richard J. Tannenbaum
ETS, Princeton, New Jersey

Septembre 2010

Dans le cadre de sa mission à but non lucratif, ETS publie et diffuse les résultats de ses programmes de recherche afin de faire progresser la qualité et l'équité dans le domaine de l'éducation et de l'évaluation pour le bénéfice d'ETS en particulier, et du secteur de l'éducation en général.

Pour télécharger une version PDF ou obtenir une version imprimée d'un rapport, rendez-vous sur :

<http://www.ets.org/research/contact.html>

Rédacteur technique principal : Daniel Eignor

Rédacteurs techniques : Irvin Katz et Michael Zieky

Copyright © 2010 par Educational Testing Service. Tous droits réservés.

ETS, le logo ETS, LISTENING. LEARNING. LEADING, et TOEIC sont des marques déposées appartenant à Educational Testing Service (ETS).

TFI est une marque déposée d'ETS.

Résumé

Le Cadre Européen Commun de Référence pour les Langues (CECRL) décrit six niveaux de compétence linguistique organisés en trois fourchettes de scores : A1 et A2 (utilisateur de base), B1 et B2 (utilisateur indépendant), C1 et C2 (utilisateur expérimenté). « *[Le CECRL] vise à fournir une base commune pour l'élaboration de programmes linguistiques, ainsi que des règles générales pour préparer des cursus, des examens, des livres de cours, etc. à travers l'Europe. Il décrit les éléments que les étudiants en langue doivent acquérir afin de pouvoir utiliser une langue à des fins de communication, ainsi que les connaissances et les compétences qu'ils doivent développer pour devenir efficaces.* » (CECRL, Council of Europe, 2001, p.1).

Cette étude met en lien les scores au TFI™ avec quatre des niveaux du CECRL : A2, B1, B2, et C1. Le TFI mesure les compétences linguistiques en Français dans les domaines de la compréhension écrite et orale, et se présente sous la forme d'un questionnaire à choix multiples. La méthode dite Angoff modifiée, associée à des évaluations holistiques, a été utilisée afin de faire correspondre les scores TFI aux niveaux du CECRL. Le panel responsable de l'étude de positionnement était composé de seize experts en langues originaires de sept pays différents.

Mots clés : CECRL, TFI, étude de positionnement, scores seuils

Remerciements

Nous souhaitons tout particulièrement remercier Laure Mounier, notre collègue du bureau ETS Global à Paris, pour son rôle dans l'organisation des réunions pour l'étude de positionnement. Nous sommes également reconnaissants envers nos autres collègues d'ETS Global : Zeineb Mazouz et Françoise Azak. Avec patience et rapidité, Zeineb a assuré la traduction du français vers l'anglais des discussions qui ont eu lieu dans le cadre de cette étude. Françoise a pour sa part géré les aspects liés à l'hébergement, aux réunions et à l'équipement. Enfin, nous remercions notre collègue Craig Stief pour son travail sur les formulaires de notation, les programmes d'analyse et le scanning des documents sur place.

Sommaire

	Page
Résumé.....	i
Remerciements	ii
Sommaire	iii
Liste des tableaux.....	iv
Contexte	1
Méthode	2
Composition du panel.....	2
Travail préliminaire.....	3
Procédure d'évaluation.....	5
Résultats.....	7
Section compréhension écrite	7
Section compréhension orale	9
Enquête d'évaluation de fin d'étude	11
Conclusions.....	13
Elaboration des scores seuils définitifs.....	15
Bibliographie	18
Notes	20
Annexe.....	21

Liste des tableaux

	Page
Tableau 1. Informations démographiques sur les experts.....	4
Tableau 2. Compréhension écrite : Résultats de l'étude de positionnement pour B1 et B2	8
Tableau 3. Compréhension écrite : Résultats de l'étude de positionnement pour A2 et C1	9
Tableau 4. Compréhension orale : Résultats de l'étude de positionnement pour B1 et B2	10
Tableau 5. Compréhension orale : Résultats de l'étude de positionnement pour A2 et C1	10
Tableau 6. Retour sur le processus de l'étude de positionnement	11
Tableau 7. Niveau de confort envers les scores seuils définitifs recommandés.....	12
Tableau 8. Scores seuils étalonnés pour le TFI	13

Contexte

Le Cadre Européen Commun de Référence pour les Langues (CECRL) définit six niveaux de compétence linguistique qui sont organisés en trois fourchettes de scores : A1 et A2 (utilisateur de base), B1 et B2 (utilisateur indépendant), C1 et C2 (utilisateur expérimenté). « *[Le CECRL] vise à fournir une base commune pour l'élaboration de programmes linguistiques, ainsi que des règles générales pour préparer des cursus, des examens, des livres de cours, etc. à travers l'Europe. Il décrit les éléments que les étudiants en langue doivent acquérir afin de pouvoir utiliser une langue à des fins de communication, ainsi que les connaissances et les compétences qu'ils doivent développer pour devenir efficaces.* » (CECRL, Council of Europe, 2001, p.1). Le but du présent travail de recherche était de mener une étude de positionnement afin de pouvoir mettre en rapport les scores obtenus au TFI™ avec les niveaux définis par le CECRL.

Le TFI mesure les compétences linguistiques en compréhension orale et écrite en français. Il s'adresse à des candidats dont la langue maternelle n'est pas le français, c'est-à-dire à des personnes qui sont en train d'apprendre la langue française. Le test mesure les compétences générales en compréhension orale et écrite pour communiquer en français, dans la vie de tous les jours et dans diverses situations professionnelles (*TFI Manuel du candidat*, ETS, 2008). Chaque section du test (compréhension orale et compréhension écrite) comprend 90 questions à choix multiples. Le TFI n'a, à l'origine, pas été conçu pour mesurer la gamme des niveaux de compétences définis par le CECRL, ni pour mesurer la compréhension orale et les compétences en compréhension écrite tels qu'elles sont définies dans le CECRL. Il est impossible, pour un panel d'experts d'une étude de positionnement, de définir les scores seuils d'un test pour des niveaux de connaissance et de compétences qui ne sont pas représentés dans le test. Par conséquent, avant de commencer l'étude sur la définition des normes, les spécialistes en matière de tests et d'évaluations d'ETS ont identifié les niveaux CECRL spécifiques correspondant le plus clairement aux sections de la compréhension orale et de la compréhension écrite du TFI. Chaque section a été évaluée comme couvrant les niveaux A2 à C1. La procédure de positionnement s'est centrée uniquement sur ces niveaux, et un ensemble distinct de scores seuils a été établi pour chacune des deux sections que sont la compréhension orale et la compréhension écrite.

Méthode

Une approche dite Angoff modifiée de définition des normes (Cizek & Bunch, 2007 ; Zieky, Perie, Livingston, 2008) associée à une méthode de jugement holistique a été utilisée afin d'identifier les scores TFI correspondant aux niveaux A2 à C1 du CECRL. La mise en œuvre spécifique de cette approche suit les travaux de Tannenbaum et Wylie (2008). Dans leur étude, les scores seuils ont été définis pour permettre une correspondance entre le Test of English for International Communication™ (TOEIC®) et le CECRL. Dans la présente étude, l'approche Angoff modifiée a été utilisée pour la série 1. Pour les séries 2 et 3, un panel a émis des jugements holistiques sur les scores des différentes sections (compréhension orale et compréhension écrite. Une revue de la littérature récente sur les approches en matière de définition des normes confirme un certain nombre de principes fondamentaux concernant les meilleures pratiques à suivre : une sélection rigoureuse d'experts, en nombre suffisant, pour représenter différents points de vues, prendre le temps de développer une compréhension commune du domaine, une formation appropriée des experts, l'élaboration d'une description de chaque niveau de performance, procéder à plusieurs séries de jugements, et le cas échéant l'intégration de données supplémentaires pour préciser les jugements (Brandon, 2004 ; Cizek, 2006 ; Hambleton & Pitoniak, 2006). L'approche que nous avons utilisée dans la présente étude tient compte de tous ces principes.

La définition des normes TFI a été réalisée en français par un animateur bilingue (français/anglais) possédant une expérience de travail avec les formateurs français. Tous les supports utilisés dans le cadre de la définition des normes, ont été élaborés par les auteurs du présent rapport, puis traduits de l'anglais vers le français, et validés avec l'animateur bilingue avant d'être utilisés pour l'étude. Les deux auteurs, ainsi qu'un deuxième animateur bilingue, étaient présents à chaque étape de l'étude. Ceci a permis de traduire les questions techniques ou concernant la procédure qui ont pu se présenter lors de la mise en œuvre de l'étude, afin que les auteurs puissent y répondre, dans la mesure où la plupart des experts ne parlaient pas l'anglais.

Composition du panel

Le panel était composé de seize personnes originaires de sept pays différents. Toutes possédaient une expérience avérée dans l'enseignement de la langue française ou son évaluation. Neuf personnes enseignaient le français comme seconde langue et quatre avaient une activité en lien avec

l'enseignement du Français Langue Etrangère (FLE) (ex : directeurs du département des langues d'une école de langues). Douze personnes avaient au moins 10 ans d'expérience dans leur domaine. Le TFI mesurant les compétences linguistiques en langue française, les personnes originaires de France étaient majoritaires dans le panel (10 sur 16). Les experts connaissaient le CECRL et le TFI, et étaient familiarisés avec la population type de candidat susceptible d'avoir à passer le TFI. Le tableau 1 présente une description des informations démographiques fournies par les experts. (Voir l'annexe pour les affiliations des experts).

Travail préliminaire

Avant de commencer l'étude de positionnement, il a été demandé aux experts d'analyser des tableaux sélectionnés de la version française du CECRL concernant la compréhension écrite et la compréhension orale. Ceci dans le but de noter les caractéristiques ou les indicateurs clés permettant de décrire un étudiant en langue française (candidat) possédant les compétences minimales requises pour être considéré comme ayant atteint les différents niveaux cibles du CECRL (A2, B1, B2 et C1). Les tableaux ont été choisis de façon à donner aux spécialistes une vue d'ensemble du niveau qui est attendu des candidats pour chaque compétence linguistique. Il était demandé aux spécialistes de considérer ce qui différencie un candidat ayant le niveau minimum lui permettant d'atteindre un niveau CECRL déterminé, d'un candidat n'ayant tout juste pas atteint le niveau seuil requis pour être considéré comme ayant atteint ce même niveau CECRL.

Afin de les aider dans cette tâche, nous avons fourni aux experts des descriptions de candidats se situant au tout début de chaque niveau CECRL cible pour la compréhension orale et la compréhension écrite. Ces descriptions provenaient de l'étude de Tannenbaum et Wylie (2008). Les spécialistes avaient pour consigne d'analyser les deux sources d'information et de modifier, le cas échéant, les descriptions de l'étude de Tannenbaum et Wylie en fonction de leur propre interprétation des niveaux du CECRL et de leur expérience dans les domaines de l'enseignement et de l'évaluation de la langue française. Ce travail en amont de la réunion a été l'occasion pour les experts d'analyser les parties pertinentes du CECRL, et constituait une première étape dans l'étalonnage des experts vers une compréhension commune des exigences minimales inhérentes à chaque niveau cible du CECRL.

Tableau 1. Informations démographiques sur les experts

		Nombre
Sexe	Féminin	13
	Masculin	3
Profession	Enseignant	9
	Directeur du département des langues d'une école de langues	4
	Consultant en éducation	1
	Chef de projet	1
	Spécialiste de l'évaluation linguistique	1
Expérience	Moins de 10 ans	4
	Entre 10 et 20 ans	6
	Plus de 20 ans	6
Pays	Belgique	1
	Canada	1
	France	10
	Iran	1
	Roumanie	1
	Russie	1
	Venezuela	1

Avant de participer à cette étude, chaque spécialiste avait également la possibilité de passer le TFI. Tous les experts ont signé un formulaire de confidentialité avant d'avoir accès au test. Le fait de se soumettre au test à titre personnel était important pour permettre aux experts de comprendre exactement quels étaient les aspects mesurés par le test ainsi que la difficulté des questions.

Procédure d'évaluation

La tâche des experts dans le cadre de l'étude était de définir les compétences minimum requises pour atteindre chacun des niveaux cibles du CECRL (A2, B1, B2, C1). Les experts ont travaillé en deux petits groupes qui avaient pour mission de définir les compétences d'un candidat qui a tout juste atteint le niveau d'exigence requis pour les niveaux B1 et B2 ; ceci a été fait séparément pour la compréhension écrite et la compréhension orale. Ce candidat-type était désigné comme un « Candidat Juste Qualifié » (CJQ). Les experts ont pu se reporter aux travaux qu'ils avaient effectués avant la réunion et aux tableaux du CECRL pour chacun des domaines de compétence. Une discussion entre tous les membres du panel était organisée pour chaque niveau et une définition définitive était établie pour chacun de ces niveaux. Les définitions CJQ pour les niveaux A1 et C1 ont été obtenues après discussion entre tous les experts, en utilisant les descriptions de B1 et B2 comme indicateurs limites. Ces définitions ont servi de cadre de référence pour les jugements de positionnement. C'est-à-dire que les experts devaient traiter les questions du test en fonction de ces définitions.

Une approche Angoff modifiée a été utilisée en suivant les procédures de Tannenbaum et Wylie (2008). Les experts ont été formés à la procédure et ont eu l'occasion de s'entraîner à faire des jugements. Il leur a été demandé, après cela, de signer un formulaire d'évaluation de formation afin de confirmer qu'ils avaient compris ce que l'on attendait d'eux et qu'ils étaient prêts à continuer. Ils ont ensuite procédé à trois séries de jugements, avec « feedback » et discussions entre les séries, pour les niveaux B1 et B2. Pour la Série 1 : pour chaque question du test, les experts devaient déterminer le pourcentage de 100 candidats juste qualifiés (B1 et B2) qui seraient capables de fournir la bonne réponse. Pour ce faire, ils ont utilisé l'échelle d'évaluation suivante (exprimée en pourcentage) : 0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100. Les experts devaient se concentrer uniquement sur la correspondance entre la compétence exigée par la question et les compétences que possédaient les CJQ, et ne devaient pas deviner pour effectuer leurs jugements. Les

experts procédaient au jugement concernant une question pour chacun des deux niveaux CECRL (B1 et B2) avant de passer à la question suivante.

La somme des jugements inter-question de chacun des experts, divisé par 100, représente le score seuil préconisé par ce groupe d'experts, en d'autres termes le nombre correct sur 90 questions.

Le score seuil préconisé par chaque membre est communiqué aux experts. La moyenne du panel (score seuil recommandé par le panel), ainsi que les scores seuils les plus élevés et les plus bas (anonymes) ont été calculés et présentés au panel pour discussion. Les experts devaient ensuite partager la logique qu'ils avaient suivie pour faire leurs jugements. Dans le cadre du « feedback » et de la discussion, les valeurs P+ (pourcentage de candidats récents¹ qui ont correctement répondu à toutes les questions) ont été communiquées. De plus, les valeurs P+ ont été calculées pour les candidats dont le score se situe sur ou au-dessus du 75^{ème} centile pour cette section (c'est-à-dire les 25% meilleurs candidats) et pour les candidats dont le score se situe sur ou en-dessous du 25^{ème} centile (c'est-à-dire les 25% des candidats les moins performants). L'analyse de la difficulté des questions pour les 25% des meilleurs candidats et les 25% des candidats les moins performants visait à donner aux experts une meilleure compréhension de la relation qui existe entre les compétences linguistiques globales pour cette section du TFI et chacune des questions. Le fractionnement a par exemple permis aux experts de voir les cas où une question n'était pas assez discriminatoire, ou lorsqu'une question était particulièrement difficile ou facile pour les candidats ayant des niveaux de compétences différents.

Dans la série 2, les jugements étaient donnés non pas au niveau de la question, mais au niveau de la section en général. Les experts devaient dire s'ils recommandaient un score de section-niveau différent (ex : compréhension orale) pour B1 et/ou B2. La transition vers un jugement de section-niveau s'intéresse plus à la construction d'ensemble (compréhension orale et compréhension écrite) qu'à la déconstruction des concepts au travers d'une autre série de jugements sur les questions-niveaux. Cette variation a déjà été utilisée dans le cadre d'autres études de mise en correspondance (Tannenbaum & Wylie 2005 ; 2008) et n'a posé aucune difficulté pour les experts faisant partie du panel TFI. Après la seconde série de jugements, un retour d'information similaire à celui de la série 1 est donné, avec en plus une présentation et une discussion du pourcentage de candidats issus d'une session récente qui seraient catégorisés dans chacun des deux niveaux (B1 et B2). Les experts devaient ensuite effectuer un jugement définitif (série 3) sur les sections-niveaux.

Les jugements définitifs (série 3) ont été analysés et communiqués aux experts. Ces derniers devaient alors recommander des scores seuils pour les niveaux A2 et C1. En particulier ils devaient analyser les descriptions A2, B1, B2 et C1 de candidats juste qualifiés afin d'identifier les scores de section-niveaux minimums pour des candidats CJQ ayant juste atteint les niveaux A2 et C1 (Tannenbaum & Wylie, 2008). Les jugements étaient encadrés par les scores seuils B1 et B2 établis précédemment. Les experts ont discuté ensemble de la position des scores seuils pour A2 et C1 ; ensuite chaque expert a émis un jugement individuel concernant ces scores seuils A2 et C1. La moyenne des recommandations individuelles (A2 et C1) a été calculée.

Résultats

Le premier ensemble de résultats résume les jugements du panel de cette étude de positionnement pour les sections compréhension écrite et orale du TFI. Les résultats sont présentés en scores bruts, ce qui correspond à l'unité de mesure utilisée par les experts. L'erreur standard de jugement (ESJ) est également mentionnée. Ceci donne une indication quant à la probabilité qu'un score seuil soit proche du score seuil actuel défini par d'autres panels de spécialistes de composition similaire et ayant reçu la même méthode de formation aux études de positionnement. Un résumé des réponses à un questionnaire d'évaluation de fin de recherche est également présenté, apportant des éléments attestant de la validité et de la pertinence de la procédure utilisée (les scores seuils étalonnés sont présentés dans la conclusion).

Section compréhension écrite

Le tableau 2 présente les résultats de l'étude de positionnement pour les niveaux B1 et B2 pour chaque série de jugements. Le score seuil moyen pour B1 a diminué dans la série 2, et a augmenté dans la série 3, pour atteindre un score plus cohérent avec le score seuil de la série 1. Le score seuil pour B2 a diminué dans la série 2 et a augmenté quelque peu dans la série 3, mais sans atteindre le score seuil recommandé dans la série 1. Pour B1 et B2 la variabilité, entre experts, a diminué au cours des trois séries, comme en témoigne la réduction de l'Ecart Type (ET). L'ESJ, liée à la variance, a également diminué au cours des séries. L'interprétation de l'ESJ indique que le score seuil d'un panel comparable se situerait à moins d'un ESJ du score seuil actuel dans 68 % des cas, et à moins de deux ESJ dans 95% des cas. L'ESJ pour la compréhension écrite dans la série 3 est de moins de deux points pour les niveaux B1 et B2, ce qui est relativement bas et apporte des éléments de confiance

dans le fait que les scores seuils seraient similaires avec un autre panel de composition et de caractéristiques similaires.

Tableau 2. Compréhension écrite : Résultats de l'étude de positionnement pour B1 et B2

	Série 1		Série 2		Série 3	
Niveaux	B1	B2	B1	B2	B1	B2
Moyenne	28.2	49.5	26.8	47.2	28.8	48.4
Médiane	26.5	49.6	25.5	48.4	28.0	46.7
Minimum	17.2	32.6	17.7	34.5	20.0	38.0
Maximum	53.0	69.8	41.2	60.0	35.0	63.0
ET	10.0	10.8	7.3	7.7	5.6	7.2
ESJ	2.5	2.7	1.8	1.9	1.4	1.8

Le tableau 3 présente les résultats de l'étude de positionnement pour la compréhension écrite pour les niveaux A2 et C1. Ces jugements ont été effectués après la présentation des scores seuils pour B1 et B2 de la série 3. Le score seuil recommandé pour A2 est inférieur d'environ 14 points bruts à celui pour B1, et le score seuil pour C1 est plus élevé d'environ 22 points bruts que le score seuil pour B2. L'ESJ pour les niveaux A2 et C1 est inférieur à deux points bruts.

Tableau 3. Compréhension écrite : Résultats de l'étude de positionnement pour A2 et C1

Niveaux	A2	C1
Moyenne	15.3	70.3
Médiane	15.0	70.0
Minimum	10.0	62.0
Maximum	20.0	82.0
ET	2.7	5.5
ESJ	0.7	1.4

Section compréhension orale

Le tableau 4 présente les résultats de l'étude de positionnement pour les niveaux B1 et B2 de chaque série de jugements. Le schéma des recommandations inter-séries, ainsi que le schéma des changements dans la variabilité, sont cohérents avec ce qui a été observé pour la compréhension écrite. Pour les scores seuils recommandés pour B1 et B2, dans la série 2, les experts ont diminué leur score seuil total, rendant l'accès à chaque niveau plus facile, puis ils les ont relevés quelque peu dans la série 3. Concernant la compréhension orale, l'augmentation constatée dans la série 3 n'a pas donné lieu à des recommandations aussi élevées que dans la série 1. Les jugements des experts ont convergé au cours des trois séries de jugements, comme en témoigne la réduction des écarts types. Les ESJ ont également diminué au cours des séries. L'ESJ de la série 3 pour les niveaux B1 et B2 est inférieur à deux points bruts.

Tableau 4. Compréhension orale : Résultats de l'étude de positionnement pour B1 et B2

	Série 1		Série 2		Série 3	
Niveaux	B1	B2	B1	B2	B1	B2
Moyenne	27.0	49.1	25.3	48.5	26.5	48.7
Médiane	26.1	49.4	25.0	48.0	25.9	47.5
Minimum	10.7	29.1	20.0	38.0	20.0	38.0
Maximum	45.3	66.5	38.0	66.0	38.0	66.0
ET	9.5	11.5	4.7	7.4	4.7	7.2
ESJ	2.4	2.9	1.2	1.8	1.2	1.9

Le tableau 5 présente les résultats de l'étude de positionnement pour la compréhension orale pour les niveaux A2 et C1. Le score seuil recommandé pour A2 est inférieur d'environ 13 points bruts à la recommandation pour B1 et le score seuil pour C1 est supérieur d'environ 18 points bruts au score seuil pour B2. L'ESJ pour les niveaux A2 et C1 est inférieur ou égal à un point brut.

Tableau 5. Compréhension orale : Résultats de l'étude de positionnement pour A2 et C1

Niveaux	A2	C1
Moyenne	13.8	66.6
Médiane	15.0	65.5
Minimum	8.0	60.0
Maximum	18.0	75.3
ET	2.6	4.1
ESJ	0.7	1.0

Enquête d'évaluation de fin d'étude

Les experts ont répondu à un questionnaire final sur la validité de la procédure (Kane, 1994) utilisée lors de l'étude de positionnement. Le tableau 6 présente le retour donné par le panel concernant la procédure dans son ensemble. La majorité des experts a émis un avis favorable ou très favorable sur le fait que le travail effectué en amont de la réunion avait été utile, qu'ils comprenaient le but de l'étude, que les instructions et les explications fournies étaient claires, que la formation était appropriée, que le temps consacré aux remarques et aux discussions était utile, et que la procédure de l'étude de positionnement était simple à suivre.

Tableau 6. Retour sur le processus de l'étude de positionnement

	Tout à fait d'accord		D'accord		Pas d'accord		Pas du tout d'accord	
	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%
Le travail préalable était utile pour la préparation de l'étude.	11	69%	5	31%	0	0%	0	0%
J'ai compris le but de cette étude.	14	88%	2	13%	0	0%	0	0%
Les instructions et les explications fournies par les animateurs étaient très claires.	8	53%	6	40%	1	7%	0	0%
La formation au positionnement était pertinente et m'a appris ce dont j'avais besoin pour accomplir ma tâche.	11	69%	5	31%	0	0%	0	0%
L'explication sur le calcul des scores seuils recommandés était claire.	4	27%	9	60%	2	13%	0	0%
Les remarques et les discussions entre les séries m'ont été utiles.	12	75%	4	25%	0	0%	0	0%
La procédure pour effectuer les évaluations étaient faciles à suivre.	4	25%	12	75%	0	0%	0	0%

Note : Les pourcentages sont calculés sur la base du nombre d'experts qui ont répondu.

Des questions complémentaires avaient pour but de déterminer l'importance de chacun des quatre facteurs suivants dans le processus de définition des normes : la définition des CJQ, les discussions entre les séries, les scores seuils des autres experts et leur propre expérience professionnelle. Tous les experts ont mentionné que leur expérience professionnelle avait une grande influence, et la grande majorité d'entre eux a également indiqué que chacun des trois autres facteurs avait également eu une grande influence sur leur travail. Cependant, près d'un tiers des experts ont également rapporté que les scores seuils des autres experts n'avaient que peu d'influence sur leurs propres jugements.

Enfin, il a été demandé à chaque membre du panel d'indiquer son niveau de confiance par rapport aux recommandations de score seuil définitives. Le tableau 7 reprend ces résultats. Quatorze des 16 experts se sont dits « très confortables » ou « assez confortables » avec les résultats de score seuil pour la compréhension orale. Deux experts ont indiqué qu'ils étaient « assez inconfortables ». Tous les experts se sont dits « très confortables » ou « plutôt confortables » avec les scores seuils pour la compréhension écrite, avec un peu plus de la moitié des experts indiquant qu'ils étaient « très confortables ».²

Tableau 7. Niveau de confort envers les scores seuils définitifs recommandés

	Très confortable		Assez confortable		Assez inconfortable		Très inconfortable	
	N	%	N	%	N	%	N	%
Compréhension écrite	8	53%	7	47%	0	0%	0	0%
Compréhension orale	8	50%	6	38%	2	13%	0	0%

Note : Les pourcentages sont calculés sur la base du nombre d'experts qui ont répondu.

Conclusions

Le but de la présente étude était de recommander des scores seuils (scores minimum) pour les sections de compréhension écrite et orale du TFI correspondants aux niveaux A2, B1, B2, et C1 du CECRL. Une approche Angoff modifiée de définition des normes associée à une partie holistique a été utilisée. Les experts ont utilisé des scores bruts comme unité de référence pour l'étude. Trois séries de jugements, avec des commentaires et des discussions, ont servi à construire les scores seuils pour les niveaux B1 et B2. Les commentaires comprenaient des informations sur les performances des candidats, à chacune des questions, et sur le pourcentage de candidats qui seraient catégorisés dans chacun des niveaux du CECRL. Les niveaux A2 et C1 ont été définis en utilisant comme référence les scores seuils définitifs (série 3) pour les niveaux B1 et B2. Le tableau 8 présente les recommandations définitives pour les scores étalonnés.

Tableau 8. Scores seuils étalonnés pour le TFI

	Compréhension écrite (max. 495 points)	Compréhension orale (max. 495 points)
Niveau CECRL		
A2	105	85
B1	185	160
B2	305	300
C1	430	395

Les réponses à l'évaluation de fin de recherche corroborent la qualité de la procédure employée pour l'étude de positionnement. La grande majorité des experts a indiqué qu'ils étaient favorables ou très favorables avec les points suivants : qu'ils avaient compris le but de l'étude, que les instructions et les explications étaient claires, que la formation était pertinente, que les remarques et les discussions leurs ont été utiles, et que la procédure de définition des normes était facile à suivre.

La moitié des experts s'est dite très confortable envers les scores seuils recommandés; les autres experts ont rapporté être assez confortables envers les recommandations de scores seuils pour la compréhension écrite, et une majorité s'est dite assez confortable envers les recommandations pour la compréhension orale. Les experts ont également eu l'occasion de proposer des commentaires ouverts par écrit concernant la procédure de définition des normes, ainsi que des commentaires sur les scores seuils recommandés. Douze des spécialistes ont rédigé de brefs commentaires.

Trois points ont émergé

Le premier étant que le TFI ne fournit pas une mesure complète des compétences en langue française car il ne tient pas compte de l'expression écrite et orale. Les experts ont fait remarquer que le TOEIC® inclut l'expression écrite et orale, et ont suggéré que ces compétences soient intégrées dans le TFI.

Un second point était que par moment les experts ont eu besoin de clarifications concernant leur tâche pour l'étude de positionnement. Ceci a nécessité une traduction du français à l'anglais puis la traduction de la réponse de l'anglais vers le français, ce qui a perturbé un peu le travail des experts et n'a pas forcément été fait au bon moment. La définition des normes pour le TFI a été réalisée en français. Tous les supports avaient été traduits de l'anglais vers le français, et les instructions ainsi que la formation concernant la procédure de définition des normes ont été présentés en français par un animateur bilingue (français/anglais). Les auteurs de l'étude étaient disponibles pour répondre aux questions techniques ou de procédure, mais ils ne parlaient que l'anglais et seule une minorité d'experts parlaient l'anglais, ce qui a nécessité le recours aux traductions.

Le dernier point avait déjà été abordé lors des discussions sur la définition de l'étude de positionnement. La question était que si seul un score total combiné était communiqué (somme des scores pour la compréhension écrite et la compréhension orale) cela risquait d'engendrer des confusions. Par exemple, le score étalonné recommandé pour A2 est de 105 pour la compréhension écrite et de 85 pour la compréhension orale (Cf. tableau 8). Cependant, le fait de conclure qu'un score combiné d'au moins 190 correspond à un niveau de compétence A2 n'est pas exact. Différentes combinaisons de scores en compréhension écrite et en compréhension orale peuvent donner un même score combiné de 190. Par exemple, un candidat pourrait obtenir 130 points étalonnés pour la

section compréhension écrite (au-dessus du score seuil recommandé), mais n'obtenir que 60 points pour la section compréhension orale (en-dessous du score seuil recommandé). Les experts ont donc suggéré que les scores seuils recommandés soient traités séparément pour la compréhension écrite et la compréhension orale.

Elaboration des scores seuils définitifs

Le panel de définition des normes a pour tâche de recommander des scores seuils. Les décideurs prennent en considération ces recommandations, mais leur tâche est de définir les scores seuils définitifs (Kane, 2002). Dans le cadre du TFI, les décideurs peuvent être des membres d'une institution académique ayant besoin d'un critère de décision pour, par exemple, l'admission à un programme d'étude en français. Les décideurs peuvent aussi être membres d'une entreprise qui souhaite définir des critères pour choisir quels seraient les employés qui devront suivre un programme de formation dispensé en français.

Les besoins et les attentes des décideurs sont variables et ne peuvent être entièrement représentés dans le cadre de la procédure de recommandation des scores seuils. Par conséquent, les décideurs ont le droit et le devoir de prendre en compte à la fois les scores seuils recommandés par le panel et d'autres sources d'information lors de la définition des scores seuils définitifs (Geisinger & McCormick, 2010). Les scores seuils recommandés peuvent être adoptés tels quels, ils peuvent être ajustés vers le haut pour satisfaire des attentes plus élevées, ou ajustés vers le bas pour refléter des attentes plus souples. Il n'y a pas de décision « correcte ». La pertinence des ajustements ne peut être évaluée que dans le contexte des attentes des décideurs. Deux sources d'information importantes à prendre en compte lors de la définition des scores seuils sont l'écart type de la mesure (ETM) et l'écart type des jugements (ETJ). La première concerne la fiabilité des scores du test TFI et la deuxième donne une indication de la fiabilité des recommandations des experts pour les scores seuils.

L'ETM permet aux décideurs de se rendre compte qu'un score obtenu à un test, quelque soit le score et le test, ne peut être parfaitement fiable. Un score n'est qu'une approximation de ce qu'un candidat connaît réellement et de ce qu'il/elle est capable de faire. Par conséquent, l'ETM répond à la question : « Dans quelle mesure le score obtenu à un test est-il une bonne approximation du *vrai score* ? » Le score obtenu par un candidat sera probablement à moins d'un ETM de son vrai score

dans 68% des cas et à moins de deux ETM dans 95% des cas. L'ETM du score étalonné pour la compréhension écrite du test TFI est de 22 points, il est également de 22 points pour la compréhension orale.

L'ETJ permet aux décideurs de se rendre compte de la probabilité que les scores seuils recommandés soient également recommandés par un autre panel de composition similaire et possédant la même expérience que le panel actuel. Plus l'ETJ est bas, plus il est probable qu'un autre panel recommanderait un score seuil proche du score seuil actuel. Plus l'ETJ est grand, moins il est probable qu'un autre panel proposerait un score seuil similaire au score seuil actuel. Par conséquent, l'ETJ peut être considéré comme une mesure de crédibilité, dans le sens où une recommandation apparaîtra comme étant plus crédible s'il est probable qu'un autre panel propose des recommandations similaires ou très proches.

Une valeur d'ETJ inférieure à la moitié de la valeur de l'ETM est souhaitable car l'ETJ est faible au regard de l'erreur de mesure globale pour le test (Cohen, Kane, & Crook, 1999). Dans la présente étude, les valeurs d'ETJ étaient inférieures de deux points bruts. En moyenne cela correspond à environ 11 points étalonnés, ou moins, soit moins de la moitié de la valeur des ETM étalonnés.

En plus des informations statistiques sur les mesures d'erreurs (ETM, EJT), les décideurs doivent également tenir compte de la possibilité d'erreurs de classification. C'est-à-dire que lorsqu'ils prennent la décision d'ajuster un score seuil, les décideurs doivent décider s'il est plus important de minimiser une décision de type faux-positif ou de minimiser une décision de type faux-négatif. Une décision de type faux-positif survient lorsque le score d'un candidat suggère un certain niveau de compétence, alors que le niveau de cette personne est plus faible en réalité (cette personne ne satisfait donc pas aux compétences requises). Une décision de type faux-négatif survient lorsque le score d'un candidat suggère qu'il ne possède pas les compétences requises, alors que dans les faits cette personne possède bien le niveau de compétence exigé. Par exemple, un score en compréhension écrite au TFI peut être utilisé par une société pour confier à un employé un poste qui nécessite une maîtrise de la langue au moins égale au niveau B2. La nature de ce poste pourrait induire que le fait de ne pas posséder un niveau de compétence au moins égal à B2 signifie que cette personne ne pourra pas assumer les fonctions clés inhérentes à son poste, ce qui pourrait avoir des conséquences

négligentes inacceptables. Dans ce cas de figure, les décideurs pourraient décider de minimiser les décisions de type faux-positif, et, pour ne pas prendre de risque, pourront relever le score seuil B2 pour la compréhension écrite. Relever ce score seuil réduit les chances d'avoir des décisions de type faux-positif dans la mesure où cela augmente le niveau d'exigence. Mais cela signifie également qu'un certain nombre d'employés qui possédaient bien un niveau B2 en compréhension écrite se verra refuser l'accès à ce poste. Les décideurs devront faire le choix quant au type d'erreur (faux-positif ou faux-négatif) qu'il est souhaitable de minimiser. Il n'est pas possible d'éliminer les deux types d'erreurs simultanément.

Bibliographie

- Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard setting topics. *Applied Measurement in Education*, 17, 59-88.
- Cizek, G. C. (2006). *Standard setting*. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of test development* (pp. 225-258). Mahwah, NJ: Lawrence Erlbaum Publishers.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: SAGE Publications.
- Cohen, A. S., Kane, M. T., & Crooks, T. J. (1999). *A generalized examinee-centered method for setting standards on achievement tests*. *Applied Measurement in Education*, 12(4), 343- 366.
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, England: Cambridge University Press.
- Geisinger, K. F., & McCormick, C.A. (2010). Adopting cut scores: Post-standard setting panel considerations for decision makers. *Educational and Psychological Measurement*, 29, 38-44.
- Hambleton, R. K., & Pitoniak, M. J. (2006). *Setting performance standards*. In R.L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 433-470). Westport, CT: Praeger.
- Kane, M. (1994). Validating performance standards associated with passing scores. *Review of Educational Research*, 64, 425-461.
- Kane, M. T. (2002). Conducting examinee-centered standard setting studies based on standards of practice. *The Bar Examiner*, 71, 6-13.
- Tannenbaum, R. J., & Wylie, E. C. (2008). *Linking English language test scores onto the Common European Framework of Reference: An application of standard setting methodology* (TOEFL iBT Series Rep. No. TOEFLibT-06, RR-08-34). Princeton, NJ: ETS.

- Tannenbaum R. J., & Wylie, E. C. (2005). *Mapping English language proficiency test scores onto the Common European Framework* (TOEFL Research Report No. RR-80). Princeton, NJ: ETS.
- Zieky, M. J., Perie, M., & Livingston, S.A. (2008). *Cutscores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: ETS.

Notes

¹ Les données P+ se basent sur 1568 personnes, qui ne sont pas de langue maternelle française, et qui ont passé le test entre le 18 Octobre et le 10 Novembre 2006. Les candidats étaient des adultes qui travaillaient dans un environnement de travail francophone ou qui étaient en train d'apprendre le français.

² L'un des membres du panel n'a pas répondu à la question sur le degré de confiance concernant les scores seuils pour la Compréhension écrite.

Annexe

Affiliations des membres du panel

Nom	Affiliation
Brigitte Ringot	FFBC/ Ecole des Mines de Douai
Patrick Goyvaerts	TOEIC BELNED - ToTaal Communicatie WIPAL bvba
Anne Lhopital	Institut National des Sciences Appliquées de Lyon (INSA)
Anna Le Verger	Université de Technologie de Compiègne
Aline Mariage	École des Ponts Paris Tech
Claudine Mela	Berlitz France
Geneviève Clinton	Arts et Métiers Paris Tech
Chantal Libert	Université Paris Ouest Nanterre La Défense
Alexandra Hull	INP-ENSEEIH
Călina-Christina Popa	Global English Inc./Groupe Renault - Automobile Dacia
Roxana Bauduin	Institut des langues et études internationales - Université de Versailles Saint-Quentin-en-Yvelines
Andrey Mikhalev	Université de langues de Pyatigorsk (Russie)
Christine Candide	Ministère de l'immigration, de l'intégration, de l'identité Nationale et du Développement Solidaire (MiiNDS)
Rokhsareh Heshmati	Ecrimed Formation et Université de Cergy-Pontoise
I. Thomas	Universidad del Lulia

Remarque : l'un des membres du jury n'a pas souhaité figurer dans le rapport final.