Validation empirique 'un test de français langue étrangère en regard du cadre européen commun de référence pour les langues

Marc Demeuse, Franck Desroches, Dominique Casanova, Alexandra Crendal, Alexandre Holle

VALIDATION EMPIRIQUE D'UN TEST DE FRANÇAIS LANGUE ETRANGEREEN REGARD DU *CADRE EUROPEEN COMMUM DE REFERENCE POUR LES LANGUES*

Marc Demeuse
Université de Mons, Belgique

Franck Desroches, *Dominique Casanova*, *Alexandra Crendal* & *Alexandre Holle*
Chambre de commerce et d'industrie de Paris, France

Résumé. Cette communication rend compte du travail mené par la Chambre de commerce et d'industrie de Paris (CCIP) pour compléter la mise en correspondance du Test d'évaluation de français (TEF) avec le *Cadre européen commun de référence pour les langues* (CECR), de plus en plus régulièrement utilisé pour le développement de matériel didactique et l'expression des niveaux de compétences en langue étrangère.
La méthodologie mise en œuvre s'appuie sur les recommandations du manuel *Relier les examens au Cadre européen commun de référence pour les langues* et s'inspire de la première approche développée par deux équipes, l'université de Mons et l'équipe du TEF, lors de la mise en correspondance empirique du test avec les *Standards linguistiques canadiens* (2002).
Le travail mené s'est articulé autour de quatre phases : premièrement, la traduction des spécifications (épreuves de compréhension écrite, compréhension orale et lexique/structure) du TEF pour répondre au format préconisé par le Conseil de l'Europe ; deuxièmement, l'approfondissement de la mise en correspondance théorique des spécifications du TEF avec le CECR ; troisièmement, après la familiarisation d'un groupe d'experts avec le matériel nécessaire à l'expérimentation, le classement individuel par ces derniers des objectifs spécifiques du référentiel TEF et d'un ensemble d'items (selon leur difficulté apparente) et, quatrièmement, l'analyse des résultats de l'expérimentation.
L'analyse des résultats montre que les jugements des évaluateurs-experts sont fréquemment concordants avec le classement initial opéré par les concepteurs du TEF, mais que des différences existent, notamment pour les items de compréhension orale.
Les analyses menées avec les experts ont été confrontées aux résultats obtenus lors de la passation de tests pour mettre en évidence, pour certains évaluateurs, la moins grande concordance entre leur classement et les niveaux de difficulté attendus et observés des items. Si la corrélation entre les classements portés par les évaluateurs et la difficulté empirique de ceux-ci est relativement élevée (0,87), c'est pour les items de niveau élémentaire que des variations importantes peuvent exister.

Mots clés: Référentiel; Test de français langue étrangère; Validation empirique.

Le *Cadre européen commun de référence pour les langues*[1] se veut une base commune pour l'élaboration de programmes de langues vivantes. Il décrit, en particulier, « les niveaux de compétence qui permettent de mesurer le progrès de l'apprenant à chaque étape de l'apprentissage et à tout moment de la vie ». Comme en

[1] Conseil de l'Europe (2e édition, 2005) : *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer*, Didier, Paris. Dans le présent document, l'acronyme CECR sera préféré pour faciliter la lecture du document.

témoigne le *Portfolio européen des langues* et l'*Europass mobilité*[2], les parcours de vie langagière sont multiples et peuvent être valorisés de différentes manières : test de français langue étrangère, diplômes de français professionnel, validation des acquis de l'expérience, etc.

Quel que soit le type d'évaluation choisie (test, examen, VAE...), les principaux concepteurs européens de tests se réfèrent désormais au CECR et fixent leurs objectifs à partir d'échelles de compétences et de niveaux partagés par toute la profession et tout utilisateur de tests. C'est ainsi que le Test d'évaluation de français (TEF) a opéré des choix (niveaux, compétences visées) dès sa création en 1998, alors même que la version du CECR n'était pas encore définitive. Tout score du candidat au TEF est ainsi exprimé en descripteurs de niveaux et de compétences qui peuvent être mis en correspondance avec le CECR.

Outre ce lien avec le CECR, le TEF officialise dès 2004 (Demeuse et al., admee 2004) sa mise en correspondance avec les *Standards linguistiques canadiens* pour les besoins de son utilisation dans le cadre des procédures d'immigration mises en place par Citoyenneté et Immigration Canada. Cette mise en correspondance a été réalisée non seulement au niveau théorique, mais aussi empirique.

À la lumière de ce travail de mise en correspondance et en s'appuyant sur les recommandations du manuel *Relier les examens au Cadre européen commun de référence pour les langues* (Conseil de l'Europe, 2003, rév. 2009), la Chambre de commerce et d'industrie de Paris (CCIP) a entrepris un travail de ré-interrogation du TEF par rapport à son cadre d'origine et notamment du CECR. Cette communication décrit l'expérimentation menée pour valider empiriquement l'alignement du TEF sur le CECR pour les épreuves sous forme de questionnaires à choix multiple.

1. Impact du Cadre européen commun de référence pour les langues sur le TEF

1.1. Les fondements théoriques du TEF

Le Test d'évaluation de français (TEF) a été créé en 1998 par la Direction des relations internationales de l'enseignement (DRI/E) de la Chambre de commerce et d'industrie de Paris (CCIP). Il est destiné à mesurer de façon précise, objective et fiable

[2] Consultable à l'adresse suivante : http://www.europass-france.org

le niveau en langue française des personnes dont la langue maternelle n'est pas le français. Il a été conçu pour répondre à la demande des grandes écoles de la CCIP et de centres de formation en français ou au français à l'étranger qui souhaitaient une évaluation standardisée des compétences en langue française pour leurs étudiants non francophones.

Fondé sur les approches communicatives, ce test repose plus particulièrement sur le modèle de compétence à communiquer de H.G. Widdowson (1996) et le cadre proposé par Bachman et Palmer (1996). S'inspirant du principe de transparence du Conseil de l'Europe, la structure du TEF est clairement décrite en termes de compétences (compréhension écrite, compréhension orale, lexique/structure, expression écrite et expression orale) et de niveaux (de 0+ à 6, soit 7 niveaux), comme en témoigne la figure 1.

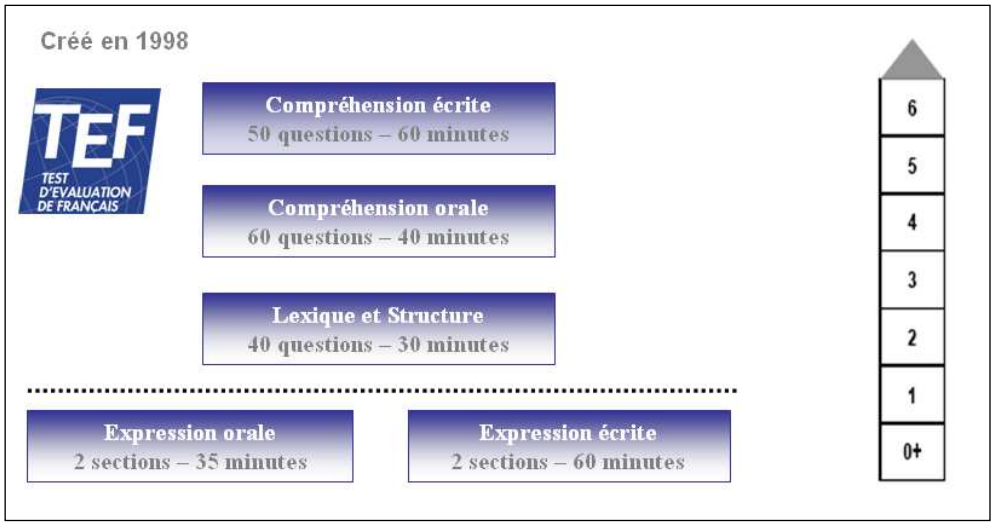


Figure 1. Présentation du Test d'évaluation de français.

Les spécifications décrivent de façon détaillée les caractéristiques du contenu du test (capacités évaluées, types de support, type de tâches demandées, etc.) ainsi que les caractéristiques techniques (durée du test, nombre de parties, nombre et type d'items, présentation et durée des tâches, etc.). La structure du TEF se compose de 50 unités-types. Ces unités sont les unités spécifiques du TEF, prototypées en termes de compétence et d'objectif visé.

Le référentiel d'évaluation du TEF s'appuie sur les mêmes fondements théoriques que le *Cadre européen commun de référence pour les langues* et notamment sur l'approche par les compétences et la perspective actionnelle. Il s'est aussi nourri d'autres référentiels comme les standards de l'ACTFL[3].

[3] American Council on the Teaching of Foreign Languages (ACTFL)

1.2. Le cadre européen commun de référence pour les langues ***et l'évaluation*** *(CECR)*

Publié en 2001 par le Conseil de l'Europe, le *Cadre européen commun de référence pour les langues* vise à proposer une référence commune pour tous ses utilisateurs que sont les administrateurs, les enseignants, les évaluateurs et les apprenants, facilitant ainsi la comparaison entre les différents systèmes de qualification européens. Devenu un instrument incontournable dans le champ de l'enseignement et de l'évaluation du français langue étrangère, il préside à l'élaboration de curricula et fonde la conception des tests et des diplômes.

Le CECR constitue une approche nouvelle dans l'enseignement, l'apprentissage et l'évaluation des langues étrangères. Non prescriptif, il présente de façon détaillée « *la description et l'étalonnage de l'utilisation de la langue et des différents types de connaissances et de compétences que cette utilisation requiert* »[4]. Doté d'une cinquantaine d'échelles de niveaux et de compétences qui s'échelonnent sur 6 niveaux, il décrit pour chaque niveau de compétence les activités langagières que tout utilisateur d'une langue est capable de réaliser pour accomplir une tâche. Cette description est ainsi appelée « descripteurs ». Le tableau 1 présente un exemple d'échelle : « l'échelle globale des niveaux communs de compétences ».

Tableau 1 – Niveaux communs de compétences – Échelle globale

ns effort pratiquement tout ce qu'il/elle lit ou ente
ts de diverses sources écrites et orales en les ré

[4] Ibid. Extrait de la 4e de couverture.

Si les descripteurs facilitent l'interprétation des niveaux, certains d'entre eux entretiennent un flou et ne permettent pas toujours aux utilisateurs d'apprécier le degré de distinction entre deux niveaux. Ces imprécisions sont remarquables entre et au sein des niveaux A2, B1 et B2.

1.3. Mise en relation du TEF avec le CECR : la mise en correspondance théorique

La mise en relation d'un test avec un référentiel de compétences langagières peut s'avérer complexe lorsqu'elle fait appel à des approches pédagogiques et linguistiques différentes. La mise en correspondance du TEF avec le *Cadre européen commun de référence pour les langues* a été en ce sens facilitée par leur proximité dans la conception de la langue axée sur l'approche par les compétences. Cette mise en relation a dû toutefois s'opérer en plusieurs fois pour assurer le lien permanent avec les versions successives du *Cadre européen commun de référence pour les langues* (1996, 2001, 2005).

La première période de mise en correspondance (1997-1998) correspond à la phase de développement et de lancement du TEF en lien avec les préconisations que met en avant le CECR dans sa version préliminaire de 1996. La deuxième période (1999-2002) voit l'approfondissement de la mise en relation du TEF avec le CECR, en particulier suite à la publication de la version officielle du CECR (2001). La troisième période (2003-2005) marque un nouveau tournant avec la révision de la mise en correspondance théorique du TEF avec le CECR. Cette révision a été effectuée à partir de la nouvelle version du CECR (2005) qui incluait des échantillons de production orale étalonnés.

Les travaux de mise en correspondance théorique ont été menés au niveau global (sur les échelles de niveaux et de compétence) et au niveau détaillé (sur les objectifs des questions du TEF). Ils ont porté sur toutes les compétences du TEF : compréhension écrite, compréhension orale, lexique/structure, expression écrite et expression orale. Au niveau global, les résultats de cette mise en correspondance théorique ont été satisfaisants : pour les échelles de niveaux relatives à la compréhension orale, à l'expression orale et au lexique/structure, on a constaté que 91 descripteurs sur 96 étaient en adéquation partielle (cas rares) ou en parfaite adéquation (cas les plus fréquents) avec ceux du CECR ; pour les échelles de niveaux relatives aux épreuves de compréhension et d'expression écrites, cette adéquation était particulièrement forte. Au niveau détaillé, les résultats ont souligné une bonne congruence entre les objectifs des

questions du TEF et les descripteurs de niveaux et de compétences du CECR, comme l'illustre le tableau 2.

Tableau 2 – Mise en correspondance théorique du TEF avec le CECR

Cadre théorique et questions du TEF			Cadre européen commun de référence pour les langues	
Code item TEF	Niveau TEF	Objectifs des questions du TEF pour le candidat	Niveau CECR	Descripteurs de compétence
CEAU1	TEF 0+	Reconnaître les situations élémentaires de la vie courante : - identifier la nature et la fonction d'un document très court de l'environnement quotidien (pancartes, étiquettes ...) ; - repérer des mots isolés. - 20 mots maximum.	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.
CEAU2	TEF 1	Comprendre les situations simples de la vie courante (sans complication) : - comprendre les informations essentielles d'un document court de l'environnement quotidien ; - textes plutôt injonctifs ou explicatifs (publicités, annonces, faire-part, notices ...) - 50 mots maximum.	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.
CEAU3	TEF 1	Comprendre les situations simples de la vie courante (sans complication) : - comprendre un document court de l'environnement quotidien ; - repérer les informations essentielles (textes plutôt injonctifs ou explicatifs : publicités, annonces, faire-part, notices ....). - 50 mots maximum.	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.

Face aux enjeux que constitue l'évaluation des compétences langagières dans l'accès au territoire, à l'université et à l'emploi en France et en Europe, il s'est avéré nécessaire de poursuivre ces travaux de mise en correspondance. De nouveaux chantiers ont donc été ouverts dès 2006 avec pour visée principale de valider empiriquement la mise en correspondance du TEF avec le CECR.

2. *Présentation de l'expérimentation*

L'objectif principal de l'expérimentation est de valider empiriquement la mise en correspondance du TEF avec le *Cadre européen commun de référence pour les langues* pour les épreuves au format de questionnaires à choix multiple. La validation[5] empirique du TEF en regard du CECR part des premiers constats établis lors de la mise en correspondance théorique[6] et vise à démontrer statistiquement la bonne corrélation entre les niveaux du TEF et ceux du CECR. La finalité de ces travaux est d'apporter des

[5] Nous nous appuierons ici sur la définition qu'en donne Jean-Pierre CUQ : « La validation est un processus par lequel on entérine, par des preuves chiffrées (notes), une évaluation, lui donnant ainsi un caractère officiel. La validation d'une évaluation ne doit pas être confondue avec sa validité ». *In* dictionnaire de didactique du français langue étrangère et seconde, Éditions Cle international, 2003.
[6] Se reporter *supra* 1.3

preuves de cet alignement conformes aux critères d'exigence du Conseil de l'Europe et de la profession[7].

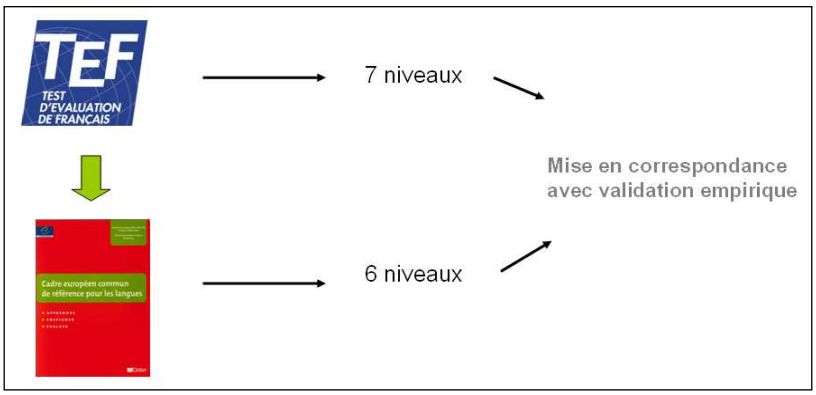


Figure 2. Présentation de l'expérimentation.

La mise en correspondance empirique du TEF avec le CECR s'appuie sur celle réalisée pour les *Standards linguistiques canadiens* (SLC). En effet, référencée dès 2002 puis homologuée par Citoyenneté et Immigration Canada pour attester avec le TEF du niveau en langue française des candidats à l'immigration, la CCIP avait défini un premier protocole en vue de la mise en correspondance théorique et empirique du TEF avec les SLC.

Le protocole suivi pour ces nouveaux travaux tient également compte des orientations méthodologiques préconisées par les experts du Conseil de l'Europe et définies dans un manuel intitulé *Relier les examens au Cadre européen commun de référence pour les langues* dont une version provisoire est proposée dès 2003. Ce manuel a été promu par des associations regroupant des professionnels de l'évaluation en langue comme ALTE[8] (Association of Language Testers in Europe) et EALTA (European Association for Language Testing and Assessment) et expérimenté par leurs membres. Les résultats de ces expérimentations ont conduit à la rédaction d'une nouvelle version du manuel en 2009, version sur laquelle les auteurs de cette communication se baseront pour le traitement des résultats.

Ce manuel propose trois types de procédures : la première est liée aux spécifications du contenu de l'examen, la seconde à la standardisation des évaluations et la troisième à

[7] La 5e norme minimal d'ALTE (Association of language testers in Europe) précise : « si vous déclarez que l'examen est relié à un système de référence externe (par exemple le Cadre européen commun de référence pour les langues), vous êtes en mesure de prouver l'alignement de l'examen sur ce système. » (Normes minimales consultables sur http://www.alte.org/standards/index.php

[8] Association dont la CCIP est membre de plein droit depuis 2010.

la validation empirique par l'analyse des données du test. Le tableau 3 présente de façon détaillée ces trois procédures.

Tableau 3 – Représentation graphique des procédures permettant de relier les examens au CECR

REPRÉSENTATION GRAPHIQUE DES PROCÉDURES PERMETTANT DE RELIER LES EXAMENS AU CECR

Extrait du Manuel *Relier les examens de langue au Cadre européen de référence pour les langues: apprendre, enseigner, évaluer* (p. 130). Conseil de l'Europe (2003)

SPÉCIFICATION DU CONTENU DE L'EXAMEN				STANDARDISATION DES ÉVALUATIONS				VALIDATION EMPIRIQUE PAR L'ANALYSE DES DONNÉES DU TEST			
Validité interne : description et analyse	1	2	3	Familiarisation	1	2	3	Validation interne	1	2	3
- du contenu général de l'examen ;				- Formation avec des échantillons normalisés de compétence de production				Analyse théorique classique des tests			
- du processus d'élaboration du test ;				- Formation avec des échantillons normalisés de compétence linguistique et de réception				Méthodes d'analyse qualitative			
- de la notation, du classement, des résultats ;								Théorie de la généralisabilité			
- de l'analyse du test et de la révision après passation.				Étalonnage d'échantillons de performances locales	1	2	3	Analyse factorielle			
Validité externe : mettre en relation	1	2	3					Analyse théorique de réponse aux items			
- la description générale de l'examen avec l'échelle du CECR ;				Définition de normes	1	2	3	Validation externe	1	2	3
- la description des activités de communication testées avec les échelles du CECR ;								Corrélation des corrections et des descripteurs du CECR			
- la description des aspects de la compétence de communication langagière testée avec les échelles du CECR.				Diffusion et mise en œuvre	1	2	3	Ancrage du test à un test déjà étalonné sur le CECR			
								Ancrage direct à la banque de descripteurs d'items qui sous-tend le CECR			

AFFIRMATION de mise en relation avec le **CECR** (en se fondant sur la spécification)

AFFIRMATION (renforcée en se fondant sur la spécification et la standardisation)

AFFIRMATION (confirmation en se fondant sur la validation empirique)

Suivant ce protocole, ce projet d'expérimentation s'est organisé autour de quatre phases :

1. la traduction des spécifications du TEF ;
2. l'approfondissement de la mise en correspondance théorique du TEF avec le CECR ;
3. la familiarisation des évaluateurs avec le CECR et le classement des questions du TEF ;
4. le traitement et l'analyse des résultats.

2.1. Phase 1 : traduction des spécifications du TEF

La première phase de ce travail a consisté en la traduction des spécifications du TEF pour répondre au format préconisé par le Conseil de l'Europe. Cela a concerné les

épreuves de compréhension écrite, de compréhension orale et de lexique/structure du test.

Parmi les vingt fiches annexées au manuel *Relier les examens au Cadre européen commun de référence pour les langues*, une dizaine ont été sélectionnées, puis complétées. Ces fiches portaient sur la description générale et détaillée de l'examen : le mode de correction et de notation retenu, la procédure de délivrance des résultats, les actions menées dans le cadre de l'analyse et de la révision de l'examen, la justification des décisions prises, etc.

Cette phase a permis de synthétiser la table de spécification du TEF en regard des recommandations spécifiques faites par le *Cadre européen commun de référence pour les langues*, mettant l'accent sur les stratégies que le candidat est amené à développer lors de la compréhension écrite par exemple.

2.2. Phase 2 : approfondissement de la mise en correspondance théorique du TEF avec le CECR

La seconde phase s'est appuyée sur la première mise en correspondance théorique du TEF avec le CECR, réalisée en 2005 par la CCIP à partir des échelles globales du TEF.

Pour chaque unité-type du TEF, la mise en corrélation effectuée avec le CECR avait pour objectif de répondre aux questions suivantes :

- Dans quelles situations attend-on des candidats qu'ils prouvent leur compétence ?
- Quels sont les thèmes de communication que les candidats doivent être capables de traiter ?
- Quelles activités communicatives les candidats doivent-ils être capables d'effectuer ?
- Quels types d'activités communicatives et quelles stratégies les candidats doivent-ils être capables de mettre en œuvre ?
- Quels types de textes et quelle longueur de texte attend-on que les candidats soient capables de traiter ?

2.3. Phase 3 : familiarisation des évaluateurs avec le CECR et classement des questions du TEF

L'objectif de cette phase était double : il s'agissait d'une part de s'assurer que les niveaux attribués aux unités-types et aux items par un groupe d'évaluateurs correspondaient aux niveaux de conception du TEF et, d'autre part, de voir dans quelle

mesure des évaluateurs étaient sensibles à des différences de niveau empirique d'items partageant les mêmes objectifs.

Cette troisième phase a été réalisée en deux étapes :

- une étape de familiarisation : formation des évaluateurs à l'utilisation du CECR ;
- une étape d'étalonnage : classement des items et réunion de consensus après identification des unités-types et des items pour lesquels les désaccords étaient les plus importants.

Pour l'étape de familiarisation, il a été fait appel à 15 évaluateurs-experts sélectionnés sur leur degré d'expertise (plus de cinq ans dans l'enseignement du français langue étrangère), leur formation (linguistique, didactique et évaluation des langues) ainsi que sur leur connaissance approfondie du CECR. Ce panel était composé de concepteurs d'items, de correcteurs de l'épreuve d'expression écrite du TEF (recrutés à l'issue d'une procédure de sélection très rigoureuse) et de formateurs en français langue étrangère.

Les évaluateurs ont, dans un premier temps, été familiarisés avec le matériel nécessaire à l'expérimentation (examen de l'ensemble des niveaux du CECR, tri des descripteurs d'une échelle du CECR, etc.). Ils ont ensuite été formés au classement d'items à l'aide d'échantillons standards du CECR (tri des descripteurs par compétence, analyse des niveaux des textes et des tâches des items calibrés, classement individuel puis mise en commun et débat jusqu'à ce qu'un consensus soit atteint). Pour déterminer la difficulté d'un item, les évaluateurs prennent en considération la complexité du support et les compétences langagières (linguistique, pragmatique, sociolinguistique) et stratégiques sollicitées, la compréhension de l'amorce et des options de l'item, le traitement du texte (compréhension générale, détaillée, etc.) et la compétence cognitive nécessaire à la réalisation de la tâche. Il faut donc se référer implicitement à un ensemble d'échelles. Or, chaque évaluateur accorde plus ou moins d'importance à telle ou telle échelle, ce qui conduit à des résultats différents. L'objectif de la standardisation est donc d'harmoniser les pratiques et de s'assurer de la capacité des personnes impliquées dans le processus de mise en relation à interpréter les niveaux du CECR de façon homogène pour aboutir à des résultats convergents.

Pour l'étape de classement, nous disposions des 50 unités-types du TEF (objectifs spécifiques du référentiel TEF) et de 185 items. Trois types d'items constituaient cet échantillon : ceux dont le niveau empirique est identique au niveau de conception, ceux dont le niveau diffère faiblement (1 niveau de différence) et ceux dont le niveau diffère fortement. Ces informations proviennent d'une part des analyses psychométriques

réalisées à la suite des passations réelles (analyse *a posteriori*) et d'autre part des analyses psychométriques réalisées à la suite des pré-tests (les items sont systématiquement pré-testés dans des conditions réelles d'examen auprès d'un échantillon représentatif de 200 candidats). Ces analyses permettent d'estimer le niveau empirique de chaque item et de le mettre en regard avec le niveau de conception pédagogique de l'item. La sélection des items est illustrée par le tableau suivant :

Tableau 4 – Standardisation des évaluations - sélection des items

Sélection des items (185 items) :

	Comparaison entre niveau empirique et niveau de conception	CE	CO	LS
Passations réelles →	Niveau empirique = niveau conception	31	29	17
	Ecart de 1 niveau	23	24	13
Pré-tests →	Ecart de plus de 1 niveau	18	18	12

Les 15 évaluateurs ont donc procédé, de manière individuelle et isolée, au classement de 50 unités-types et de 185 items, selon leur difficulté apparente, dans les six niveaux du *Cadre européen commun de référence pour les langues*. Chaque évaluateur a été convoqué quatre fois (Classement des unités-types puis - chaque compétence étant évaluée séparément - des items de compréhension écrite, de compréhension orale et de lexique/structure). La méthode adoptée pour le classement était de type semi-guidée : chaque évaluateur devait affecter un niveau à chaque items et disposait d'une indication du nombre d'items attendus pour chacun des niveaux, sous la forme d'une fourchette de valeurs.

Après l'identification des unités-types et des items pour lesquels les désaccords étaient les plus importants, les évaluateurs ont été regroupés pour une réunion de consensus. L'objectif de cette concertation était de tenter, par la discussion entre évaluateurs, d'améliorer la concordance des classements inter-évaluateurs.

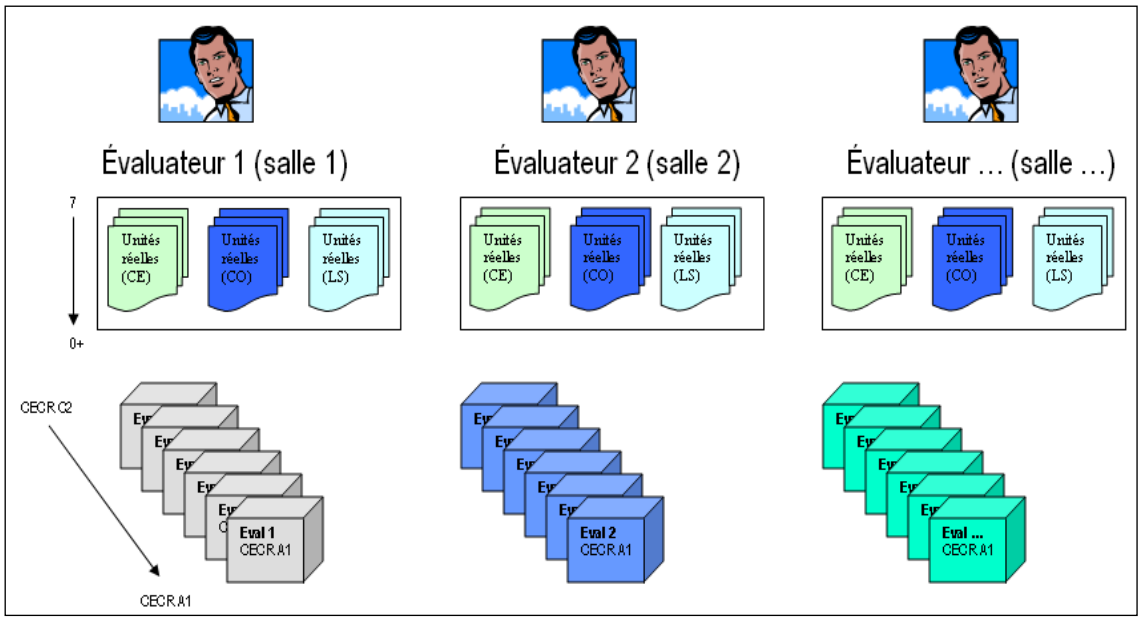


Figure 3. Standardisation des évaluations – classement des items.

3. Résultats

Les premiers classements réalisés par les évaluateurs concernaient les unités-types. Le tableau 5 présente le résultat de ces classements. Chaque ligne du tableau représente les résultats des classements individuels (soit 15 classements par unité-type), selon les niveaux du CECR, des unités-types d'un niveau TEF donné.

Tableau 5 – Classement des unités-types

		Classement selon le niveaux CECR					
		A1	**A2**	**B1**	**B2**	**C1**	**C2**
Niveau de conception TEF	**6**					4	**56**
	5				18	**69**	3
	4	1	2	29	**57**	12	
	3	10	32	**66**	23	4	
	2	49	**61**	18	4	3	
	0+/1	**164**	37	20	4		

Ainsi, lorsqu'on considère les unités-types de niveau de conception 0+ et 1, 164 jugements d'évaluateurs les classent au niveau A1, 37 au niveau A2, 20 au niveau B1 et 4 au niveau B2. La diagonale du tableau correspond donc aux classements qui accordent aux unités-types d'un niveau de conception donné du TEF le niveau CECR ciblé. Le

rapport entre la somme de ces valeurs et le nombre total de classement constitue le taux de concordance exact, qui s'élève à 63%. Lorsque l'on tient compte des cellules adjacentes (cellules sur fond blanc), on obtient un taux de concordance à un niveau près de 93%. Les cellules grisées correspondent aux unités-types pour lesquelles certains évaluateurs ont attribué un niveau CECR éloigné de plus d'un niveau du niveau ciblé.

On constate toutefois des différences notables de concordance selon les compétences considérées. Ainsi, le tableau 6 montre que c'est en compréhension orale que les classements sont les plus discordants.

Tableau 6 – Classement des unités-types

Compétence	Compréhension écrite	Compréhension orale	Lexique et structure
Taux moyen de concordance exact par évaluateur	70%	54%	77%

Ces discordances sont notamment dues à un manque de précision dans la description des objectifs et caractéristiques attendues des différents types d'items et à des interprétations différentes entre évaluateurs. Cela a été clairement mis en évidence lors de la réunion de consensus, où les évaluateurs ont été amenés à repositionner, après discussion et clarifications apportées par la CCIP, 6 unités-types pour lesquelles les écarts étaient particulièrement importants. Le taux de concordance exact pour ces unités est alors passé de 16% à 36%, et le taux de concordance à un niveau près de 57% à 94%.

Lorsque l'on considère les items dont le niveau empirique correspond au niveau de conception du TEF (cas de 76 items), on obtient des résultats comparables, avec toutefois des taux de concordance exacte plus faibles (Cf. tableau 7).

Tableau 7 – Classement des items

Compétence	Compréhension écrite	Compréhension orale	Lexique et structure
Taux moyen de concordance exact par évaluateur	61%	48%	64%

Compte-tenu de ces écarts, la médiane des classements individuels a été choisie pour exprimer le niveau attribué à une unité-type ou un item par le groupe d'évaluateurs.

Les tableaux 8 et 9 comparent les niveaux CECR ainsi attribués par le groupe d'évaluateurs au niveau TEF de conception, respectivement pour les unités-types et

pour les items. Dans les deux cas de figure, on constate un écart maximum de 1 niveau. Le taux de concordance exacte est de 80% pour les unités-types et de 71% pour les items.

Tableau 8 – Concordance entre le niveau CECR attribué par le groupe d'évaluateurs et le niveau TEF de conception pour les unités-types

		Niveau attribué par le groupe d'évaluateurs					
		A1	**A2**	**B1**	**B2**	**C1**	**C2**
Niveau de conception	**6**						**4**
	5					**6**	
	4			1	**6**		
	3		2	**6**	1		
	2	2	**6**	1			
	0+ / 1	**12**	3				

Pour les unités-types, les écarts de classement par rapport à la diagonale interviennent principalement pour les niveaux les plus faibles (A1 à B1), alors que l'accord est quasi parfait pour les niveaux B2 à C2. La situation est plus contrastée quand on considère les items où la proportion des écarts est plus importante dans les niveaux supérieurs.

Tableau 9 – Concordance entre le niveau CECR attribué par le groupe d'évaluateurs et le niveau TEF de conception pour les unités-types

		Niveau attribué par le groupe d'évaluateurs					
		A1	**A2**	**B1**	**B2**	**C1**	**C2**
Niveau de conception	**6**					1	**3**
	5				2	**2**	1
	4			3	**4**	2	
	3		1	**10**	1		
	2	4	**7**	3			
	0+ / 1	**28**	4				

Le Kappa de Cohen propose une correction au taux de concordance pour tenir compte de la possibilité de parvenir à un accord de classement par hasard (Conseil de l'Europe, 2003:99). Les valeurs du Kappa de Cohen obtenues respectivement pour les

unités-types et les items sont de 0,75 et de 0,61 ce qui, au regard des critères utilisés pour d'autres tests de langue (Shaw & Falvey, 2008:97-98), correspond à un bon niveau de concordance.

Le tableau 10 rend compte des taux de concordance exacte et du Kappa de Cohen pour les différentes compétences (l'écart de classement est de 1 niveau maximum, tant pour les unités-types que pour les items).

Tableau 10 – Taux de concordance exacte par compétence

		Compréhension écrite	**Compréhension orale**	**Lexique et structure**
Unités-types	Concordance exacte	62%	80%	100%
	Kappa de Cohen	0,53	0,75	1
	Qualité de la concordance	**Modérée**	**Bonne**	**Très bonne**
Items	Concordance exacte	68%	66%	88%
	Kappa de Cohen	0,55	0,54	0,83
	Qualité de la concordance	**Modérée**	**Modérée**	**Très bonne**

On constate que la concordance est particulièrement satisfaisante pour les unités-types et les items de lexique/structure, qui, pour la plupart, ne font pas référence à un document support et renvoient à la connaissance de structures grammaticales, à la connaissance/compréhension de mots ou d'expressions ou à la compréhension de phrases isolées.

Les items de compréhension écrite et de compréhension orale sont pour leur part construits autour d'un document support (image, texte, bande son), voire d'une combinaison d'images et de bandes son pour certains items de compréhension orale. Ceci rend probablement plus complexe la tâche d'attribution d'un niveau et peut expliquer en partie la moindre concordance entre les niveaux TEF et le classement du groupe d'évaluateurs.

Globalement, le classement des items par le groupe d'évaluateurs est proche des niveaux de conception attribués par la CCIP. Si la qualité de la concordance exacte est modérée pour les items des épreuves de compréhension écrite et orale, les écarts de classement sont au maximum de un niveau.

Par ailleurs, en considérant les évaluations individuelles et en appliquant le modèle de Rasch[9] (Penta et al., 2005), on peut obtenir une estimation de la difficulté des items, qui peut être comparée (figure 4) avec les valeurs de difficulté de la banque d'items, obtenues lors de la passation réelle des items par des candidats.

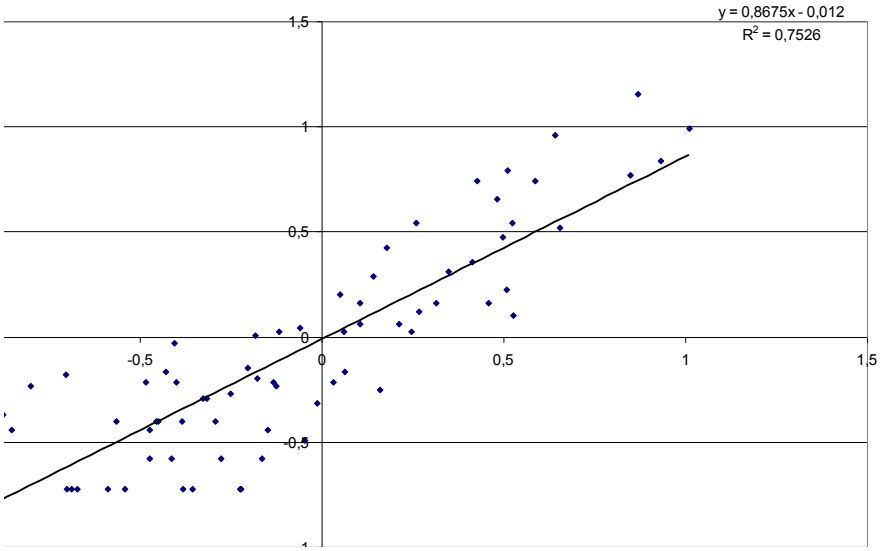


Figure 4. Comparaison de l'estimation de la difficulté des items à partir des classements individuels avec l'estimation empirique de leur difficulté obtenue lors de passations réelles.

Chaque point sur la figure représente un item avec, en abscisse, sa difficulté empirique et en ordonnée l'estimation de sa difficulté à partir du classement des évaluateurs. La corrélation entre ces deux estimations de la difficulté est élevée (0,87 à $p < 0,001$) et on constate que c'est pour les niveaux les plus faibles qu'il y a les plus fortes variations.

Cette relativement bonne correspondance doit cependant être nuancée par la relative faiblesse des accords entre évaluateurs. La notion de « niveau de groupe » a en effet d'autant plus de sens que les différents évaluateurs s'accordent entre eux pour attribuer un même niveau à l'item. La figure 5 rend compte de cette difficulté des évaluateurs à s'accorder entre eux dans le classement des items. Le taux d'accord par compétence

[9] La CCIP utilise le logiciel Conquest (Adams et al., 1998) pour la mise en œuvre du modèle de Rasch (modèle de réponse à l'item à un paramètre) afin d'estimer la difficulté des items du TEF.

entre un évaluateur et le reste du groupe tel qu'il est présenté correspond à la moyenne du taux d'accord exact entre l'évaluateur et chacun des autres évaluateurs du groupe.

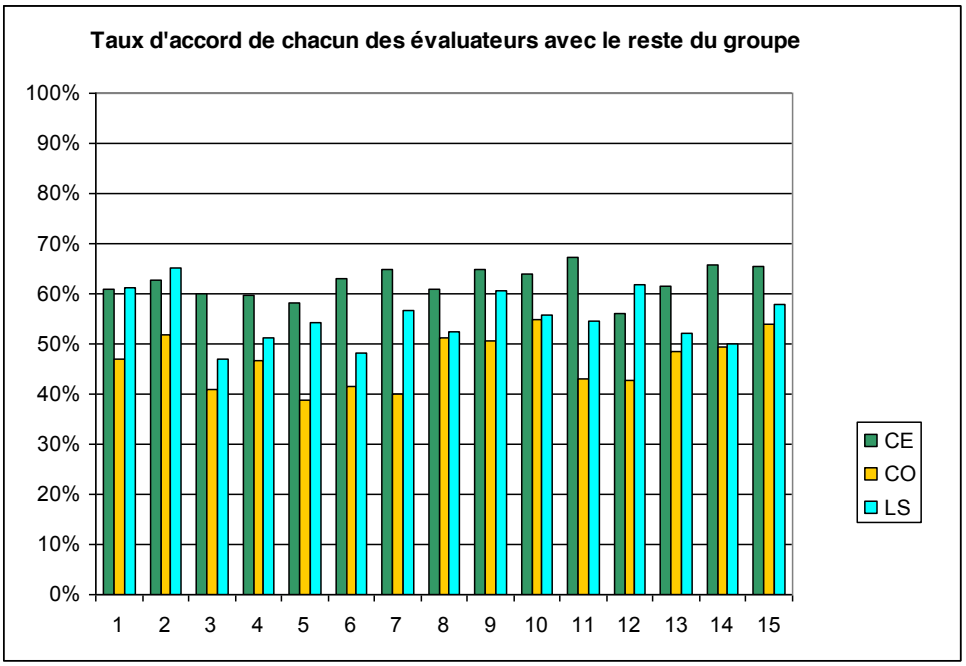


Figure 5. Taux d'accord, pour chaque compétence, entre chaque évaluateur et le reste du groupe (les évaluateurs sont représentés par les numéros 1 à 15).

Il est manifeste que les évaluateurs ont sensiblement plus de difficulté à s'accorder sur le niveau des items de compréhension orale et c'est en compréhension écrite que l'on constate le plus d'homogénéité dans les classements. Plusieurs facteurs contribuent à la relative faiblesse de ces taux d'accord.

- La nature du référentiel (CECR), tout d'abord, dont les niveaux sont décrits en faisant usage de qualificatifs relatifs ne permettant pas toujours de faire la différence entre certains aspects de la compétence pour deux niveaux adjacents (Cf. tableau 5).
- Le manque d'uniformité d'interprétation de ce référentiel par les évaluateurs qui peut subsister au-delà de la phase de familiarisation.
- La difficulté de la tâche qui consiste à classer des items selon les niveaux du CECR, qui est notamment liée au facteur précédent.
- Le fait d'avoir demandé aux évaluateurs de reclasser individuellement certains items, au début de la réunion de consensus, a notamment permis de constater un manque manifeste de consistance dans les classements des items. Comme le montre le tableau 11, la corrélation entre les résultats de ce classement et ceux du

classement initial sur les items sélectionnés pour la réunion de consensus, pour chaque évaluateur, est en effet souvent faible.

Tableau 11 – Corrélation, pour chaque évaluateur ayant participé à la réunion de consensus, entre les classements d'items effectués avant la réunion et les classements entrepris en début de réunion (items pour lesquels les classements des différents évaluateurs étaient les plus discordants)

Evaluateur	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Corrélation	0,59	0,66	0,58	0,54	0,58	0,65	0,40	0,51	0,57	0,48	0,33	0,57	0,69	0,51

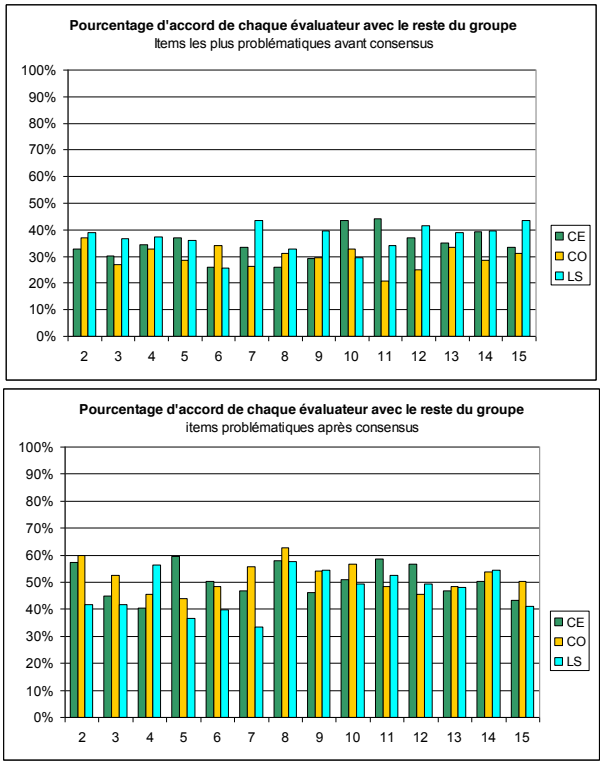
NB : *Toutes les corrélations sont significative à p < .001, à l'exception de l'évaluateur 8 (p = 0,007) et de l'évaluateur 12 (p=0,032).*

Ce reclassement d'items n'a toutefois concerné que les items pour lesquels les classements de différents évaluateurs étaient les plus discordants, dont potentiellement les items les plus difficiles à classer avec certitude à un niveau donné. Il aurait été intéressant de faire reclasser également des items faisant davantage consensus pour se faire une idée plus précise de la consistance des classements.

- Les différences de « sévérité » entre juges. La mise en œuvre du modèle de Rasch multi-facettes a permis d'estimer la sévérité relative avec laquelle les évaluateurs procèdent au classement des items. En dehors des deux évaluateurs extrêmes (qui surestiment ou sous-estiment le plus le niveau des items), les écarts de sévérité sont limités et l'indice de fidélité de la séparation des évaluateurs relativement peu élevé (0,613, p = 0,002). En revanche, les indices d'ajustement (Infit) sont souvent élevés, ce qui confirme l'existence de fluctuations intra-évaluateurs, qui expliquent probablement davantage les écarts de classement que la sévérité relative des évaluateurs.
- La difficulté à percevoir des différences de niveau entre des items du même type. Parmi les items classés par les évaluateurs, certains présentaient des estimations empiriques de difficulté (établies à partir de résultats de candidats à des pré-tests) les situant à un niveau différent du niveau requis. Lorsque cette différence n'était que de 1 niveau (cas de 60 items), les évaluateurs ont eu davantage tendance à positionner l'item au niveau requis par l'unité-type (42%) qu'au niveau constaté empiriquement (36%). Cela illustre la difficulté des évaluateurs à accorder un niveau précis à un item. Pour les items dont la différence entre le niveau empirique et le niveau requis était supérieur à 1 niveau (cas de 48 items), le taux de

correspondance avec le niveau requis par l'unité type était sensiblement plus faible mais restait non négligeable (32%).

Une façon de renforcer la convergence des classements des évaluateurs, et donc d'apporter davantage de crédit à la concordance entre les niveaux attribués par le groupe aux items et les niveaux ciblés par le test, est d'organiser des réunions de consensus, dans lesquelles les différents évaluateurs exposent les raisons pour lesquelles ils ont accordé un niveau particulier à un item donné avant de procéder à un nouveau classement. Cette confrontation des points de vue conduit en effet à harmoniser les représentations et renforce la standardisation des classements. C'est du moins ce qui a pu être mis en évidence pour les items les plus problématiques[10] de cette expérimentation, les seuls à avoir fait l'objet d'un traitement en réunion de consensus à laquelle ont participé 14 des 15 évaluateurs (figures 6 et 7).



Figures 6 et 7. Évolution du pourcentage d'accord entre évaluateurs après réunion de consensus.

On constate en effet une amélioration très sensible du pourcentage d'accord entre évaluateurs pour ces items, qui passe d'une moyenne de 34% avant consensus à une

[10] Ces items regroupaient 5 items dont le niveau empirique correspondait au niveau requis pour l'item et 38 items dont le niveau empirique était différent, ce qui explique en partie la difficulté des évaluateurs à s'accorder sur un niveau précis, même après consensus.

moyenne de 51% lors du classement individuel qui a suivi la réunion de consensus. Cette amélioration est notamment notable pour la compétence de compréhension orale. Elle est encore plus nette quand on considère le pourcentage moyen d'accord adjacent (pourcentage de classements d'évaluateurs concordants à un niveau près), comme le montre le tableau 12.

Tableau 12. Taux moyens d'accord adjacent entre évaluateurs avant et après réunion de consensus

Compétence	Taux moyen d'accord adjacent avant consensus	Taux moyen d'accord adjacent après consensus
Compréhension écrite	**80%**	**97%**
Compréhension orale	**72%**	**94%**
Lexique et structure	**81%**	**90%**

Cette convergence des points de vue à un niveau près montre l'intérêt de la réunion de consensus pour parvenir à une comparaison fondée entre les niveaux attribués aux items par le groupe d'évaluateurs et les niveaux ciblés par les concepteurs du test. Pour renforcer cette convergence, notamment lorsque le nombre d'évaluateurs est limité, un taux de concordance exacte cible peut être défini (par exemple lorsqu'au moins les deux tiers des évaluateurs attribuent un même niveau à l'item) et plusieurs réunions de consensus envisagées, en retenant pour la phase de consensus suivante les items pour lesquels le taux de concordance cible n'a pas été atteint.

4. Conclusion

Ce document rend compte du travail entrepris par la Chambre de commerce et d'industrie de Paris pour mettre en correspondance, de façon théorique et empirique, le Test d'évaluation de français avec le *Cadre européen commun de référence pour les langues*.

L'expérimentation menée avec les items du TEF montre une bonne concordance entre les niveaux attribués par le groupe d'évaluateurs et les niveaux ciblés par les unités-types et les items du TEF. Des disparités apparaissent toutefois selon les compétences, les unités-types et items se référant à un support (texte, image ou bande son) se voyant parfois attribuer un niveau différent. On constate également des écarts de

classements parfois importants entre évaluateurs, notamment pour les items de compréhension orale.

Il ressort de ces résultats qu'il est important, pour mener à bien la mise en correspondance empirique d'un test avec un référentiel comme le CECR, de :

- consacrer un temps suffisant à la familiarisation des évaluateurs avec le CECR, pour s'assurer qu'ils échangent des représentations partagées du référentiel et qu'ils prennent en considération les mêmes aspects dans le classement des items, ce qui renforce la validité des classements ;
- organiser des réunions de consensus pour parvenir à un taux de concordance satisfaisant entre les classements des évaluateurs (en d'autres termes, une meilleure fidélité des classements), qui permettra une comparaison fondée entre les niveaux attribués aux items par le groupe d'évaluateurs et les niveaux ciblés par les concepteurs du test ;
- faire appel à un panel relativement important d'évaluateurs, qui donne du sens à la notion de majorité ou à des indices statistiques comme la médiane lorsque des différences de classement subsistent entre évaluateurs.

L'entreprise de mise en correspondance d'un test et d'un référentiel est donc coûteuse, et soulève des questions de faisabilité, notamment lorsqu'on souhaite ainsi mettre en correspondance un nombre élevé d'items. Elle est toutefois essentielle dans le cas d'un test à fort enjeux comme le TEF, utilisé notamment à des fins de mobilité internationale.

Bibliographie

Bachman, L. F., Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Chambre de commerce et d'industrie de Paris (2008). *Le Test d'Evaluation de Français de la Chambre de Commerce et d'Industrie de Paris.* Paris : Chambre de Commerce et d'Industrie de Paris, Centre de Langue de la Direction des relations internationales de l'Enseignement.

Conseil de l'Europe (2007). *Portfolio européen des langues : 15 et +*. Scéren CRDP Basse-Normandie, Didier.

Conseil de l'Europe (2001, rév. 2005). *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer*, Didier.

Conseil de l'Europe (2003, rév. 2009). Manuel *Relier les examens de langue au Cadre européen de référence pour les langues : apprendre, enseigner, évaluer (CECR) – Manuel avant-projet*. Strasbourg : DGIV/EDU/LANG.

Crendal, A. (2005). « Vers la multiréférentialisation du TEF », *Points communs, 24,* 7-13.

Demeuse M., Desroches F., Crendal A., Oster P., Renaud F. et Leroux X (2004). L'évaluation des compétences linguistiques des adultes en français langue étrangère dans une perspective de multiréférentrialisation. In *Actes du XVII*^{e} *colloque de l'ADMEE*, Lisbonne, Portugal.

Penta M, Arnould C, Decruynaere C. (2005). *Développer et interpréter une échelle de mesure. Applications du modèle de Rasch*. Collection : Pratiques psychologiques: évaluation et diagnostic. Pierre Margada Editeur.

Shaw S., Falvey P. (2008). *The IELTS Writing Assessment Revision Project : Towards a revised rating scale*. Cambridge ESOL.

Widdowson, H.G. (1996). *Une approche communicative de l'enseignement des langues*. Didier.

Wu, M.L., Adams, R.J., & Wilson, M.R. (1998). *CONQUEST. Generalised Item Response Modelling Software.* Melbourne : Australian Council for Educational Research.