La fiabilité de l'évaluation des compétences linguistiques pour des adultes non francophones : Présentation d'un protocole d'évaluation

DEMEUSE Marc, *Professeur, Institut d'Administration scolaire,* Université de Mons-Hainaut, Mons, Belgique, marc.demeuse@umh.ac.be

CRENDAL Alexandra, *Responsable pédagogique,* Chambre de Commerce et d'Industrie de Paris, Paris, France, acrendal@ccip.fr

DESROCHES Franck, *Responsable du TEF,* Chambre de Commerce et d'Industrie de Paris, Paris, France, fdesroches@ccip.fr

RENAUD François, *Responsable pédagogique,* Chambre de Commerce et d'Industrie de Paris, Paris, France, frenaud@ccip.fr

CASANOVA Dominique, *Responsable pédagogique,* Chambre de Commerce et d'Industrie de Paris, Paris, France, dcasanova@ccip.fr

ARTUS Frédérique, *Chercheuse, Institut d'Administration scolaire,* Université de Mons-Hainaut, Mons, Belgique, frederique.artus@umh.ac.be

Mots clés : français langue étrangère (FLE) ; compétences linguistiques ; adultes non francophones ; qualité (pédagogique, scientifique, ISO)

Résumé

Le champ de l'évaluation en français langue étrangère (FLE) s'est beaucoup développé ces quinze dernières années, et notamment en Europe. Plusieurs événements conjoncturels ont poussé ce domaine à évoluer : la convergence des politiques linguistiques et éducatives européennes, la validation des acquis de l'expérience (VAE) et les recherches en FLE, alliant désormais systématiquement analyses qualitatives et quantitatives. D'autres facteurs, socio-économiques, ont aussi amené ces changements. Face aux enjeux de mobilité, l'évaluation en français langue étrangère a mis en place des outils d'évaluation valides, fidèles, fiables, étalonnés sur des cadres de référence, permettant aux apprenants de voir reconnaître leur niveau, quel que soit le pays.

C'est dans ce contexte que la communication se situera, celle de l'opérationnalisation d'un test de français langue étrangère répondant aux critères de scientificité et se pliant aux contraintes du contrôle de sa qualité, suivi réalisé par une équipe universitaire indépendante et l'équipe pédagogique appartenant à la *Chambre de Commerce et d'industrie de Paris* (CCIP). L'exposé développe deux approches méthodologiques, l'une centrée sur les processus d'élaboration des tests (approche psychométrique et édumétrique, analyse des données issues de pré-tests et de situations réelles de test) et l'autre centrée sur le contrôle qualité des procédures (analyse des moments critiques), notamment liées au travail d'administration des tests dans des centres agréés. La réflexion autour des résultats soulignera la nécessité d'une quadruple lecture, à la fois scientifique, pédagogique, économique et éthique. L'échantillon de résultats mettra en évidence ces différentes approches et le choix des indicateurs nécessaires pour répondre aux différents interlocuteurs : les organismes de certification, les utilisateurs individuels et institutionnels, les développeurs et les chercheurs. Enfin, les auteurs montreront la complémentarité entre les approches, à travers l'analyse de données réelles, mais aussi les limites pouvant apparaître entre les intérêts de chacun des acteurs.

Introduction

Le domaine de l'évaluation en français langue étrangère s'est beaucoup développé ces quinze dernières années en Europe. Plusieurs événements conjoncturels en politiques linguistique et éducative ont poussé ce domaine à évoluer : la convergence des politiques linguistiques européennes, la valorisation de la reconnaissance et des acquis de l'expérience et les recherches dans le domaine de l'évaluation en français langue étrangère, alliant désormais systématiquement analyses qualitatives et quantitatives. Des référentiels déjà anciens, comme le *Français fondamental* (Gougenheim *et al.*, 1956) *,* en sont à l'origine ; le *Cadre européen commun de référence pour les langues* (CECR) du Conseil de l'Europe (2000) en est la continuité.

D'autres facteurs, socio-économiques, ont aussi amené ces changements. Face aux enjeux de mobilité, plusieurs opérateurs ont progressivement mis en place des outils d'évaluation valides, fidèles, fiables, étalonnés sur des cadres de référence tels que le CECR et les Standards linguistiques canadiens (Citoyenneté et Immigration Canada, 2002), permettant aux apprenants de se voir reconnaître leur niveau, quel que soit le pays d'accueil.

Ces outils d'évaluation, les tests de langue, sont aujourd'hui de plus en plus utilisés par les apprenants, les centres de langues, les entreprises et des institutions telles que les services canadiens de l'immigration (Citoyenneté et Immigration Canada). Ils leur permettent d'avoir une évaluation fiable, comparable et ne requièrent pas des candidats de s'engager au préalable dans un parcours de formation, comme dans le cas des diplômes.

Comme le souligne Eric Delamotte (1999), « *la temporalité est renouvelée* ». Deux niveaux temporels peuvent ainsi être dégagés : l'un que nous nommerons chronométrique, l'autre chronologique. Au niveau chronométrique, les apprenants recherchent de plus en plus une évaluation instantanée, attestant rapidement de leurs acquis. En dehors des institutions de formation, les autres usagers (ministères, entreprises…) se tournent aussi de plus en plus vers les tests de langues pour évaluer le niveau du candidat au moment de l'instruction de son dossier, sans nécessairement s'intéresser à la manière dont celui-ci a acquis ses compétences (formation initiale et/ou continuée, apprentissage formel ou non, immersion...). Au niveau chronologique, la mise en place du portfolio européen et l'incitation à la "formation tout au long de la vie" amènent aujourd'hui les candidats à capitaliser leurs expériences au fil du temps. Pour les centres de langues, on constate également que le test est un outil utile au positionnement des apprenants dans des groupes de niveaux. Il leur permet par ailleurs, comme aux apprenants, d'évaluer les progrès réalisés (Engle & Engle, 2003).

La présente communication abordera les méthodes mises en œuvre dans le cadre du contrôle de la qualité du *Test d'Evaluation de Français* (TEF) de la *Chambre de Commerce et d'Industrie de Paris*. Deux approches méthodologiques seront développées, l'une centrée sur le processus d'élaboration des tests (approche psychométrique et édumétrique, analyse des données issues de pré-tests et de situations réelles de test) et l'autre sur le contrôle qualité des procédures (analyse des moments critiques), notamment liées au travail d'administration des tests dans des centres agréés.

1. Le processus d'élaboration des tests : le cas du TEF

On procédera, dans cette première partie, à une brève présentation de la procédure de validation du processus de conception du TEF, depuis la conception pédagogique des items jusqu'à leur validation psychométrique. Cette présentation repose notamment sur l'analyse d'échantillons de données collectées lors de pré-tests servant à la mise au point des items des différentes formes parallèles et de l'exploitation *a posteriori* des données acquises auprès de plusieurs dizaines de milliers de candidats en situation réelle de test.

L'indexation sur des cadres de référence : CECR et SLC

Le TEF évalue les compétences langagières d'un candidat : la compréhension écrite, la compréhension orale et le lexique/structure. Il se compose de 150 questions à choix multiples, passées par le candidat sur une durée de deux heures dix.

Ce test s'appuie, entre autres, sur les travaux menés par le Conseil de l'Europe (2000) dans le cadre de l'élaboration du *Cadre européen commun de référence pour les langues (CECR)*. Ce cadre comporte six niveaux auxquels correspondent les niveaux TEF. Les résultats des candidats sont exprimés en termes de scores, puis rapportés en niveaux, tels que définis par le *CECR* et les *Standards linguistiques canadiens*[1], pour les candidats qui souhaitent faire valoir leurs résultats auprès des autorités canadiennes chargées de l'immigration économique dans ce pays.

La conception du test repose également sur le CECR qui sert de référent et permet à l'équipe de concepteurs, préalablement formée aux spécifications du test, de pré-calibrer les items.

La conception et la validation pédagogique des items

Plusieurs référents concourent au pré-calibrage des items : les cadres de référence, cités dans le paragraphe précédent, et la table de spécification. D'autres outils comme le guide du concepteur, le référentiel de validation, renforcent également la maîtrise des activités de conception et de validation. Les concepteurs d'items concevant à distance, un dispositif de formation à ces différents outils a été mis en place pour garantir la qualité en conception et en validation d'items.

Dès que les items sont conçus et validés pédagogiquement, ou rejetés en cas de non conformité, ils intègrent la banque de données dans l'attente d'être sélectionnés pour l'édition de jeux de pré-test.

[1] Ces derniers standards comportent 12 niveaux. La mise en correspondance théorique et empirique des standards canadiens et du CECR a été présentée lors d'un précédent colloque de l'ADMEE-Europe (Demeuse *et al.*, à paraître) et dans la revue Points Communs (Crendal, 2005). Ce travail s'articule autour de la réflexion menée par la Division des politiques linguistiques du Conseil de l'Europe (2003).

Le pré-testage des items

Le calibrage des items est effectué à partir de formes complètes (situations réelles d'examen) qui comportent des items d'ancrage et des nouveaux items à valider. Plusieurs procédures de validation existent : certaines reposent sur la passation, par des candidats volontaires, d'une seule forme ; d'autres comportent une double passation à quelques jours d'intervalle, sans qu'aucun apprentissage significatif dans la langue cible ne soit réalisé dans cet intervalle. Dans cette dernière procédure, les mêmes candidats sont soumis à deux formes différentes, mais réputées parallèles. Cette procédure présente naturellement un avantage indéniable en termes économiques puisqu'un plus grand nombre d'items peut être validé et que la procédure d'ancrage peut être optimisée. Ces procédures, recourant à des candidats volontaires, placés dans des situations réelles – souvent quelques jours avant une véritable passation amenant à une attestation – semblent, d'un point de vue éthique et pratique, préférable à la solution qui consiste à placer de nouveaux items au sein d'une forme validée et présentée en situation réelle, sans que le candidat, à la limite, n'en soit informé.

Le caractère éthique de cette décision est assez facile à mettre en évidence : 100% du temps que le candidat accorde à la passation est réellement consacré à son évaluation et non à un "détournement" au profit des développeurs. Le candidat n'est donc pas utilisé à son insu et son évaluation ne repose que sur des items validés au préalable. Dans une procédure de validation qui repose sur des situations réelles dans lesquelles de nouveaux items sont introduits, la tentation peut également être forte d'utiliser les résultats aux items non encore validés, notamment si les développeurs souhaitent ne pas perturber la forme du test, dans son ensemble, alors même qu'ils veulent aussi couvrir tout le spectre du test, notamment pour fournir des résultats diagnostiques en termes de compétences. Cette solution a, naturellement, été écartée dans le cas du TEF, principalement pour cette raison : la structure du TEF est stable et permet un diagnostic fin. Cette option des concepteurs du TEF peut être maintenue notamment parce que toutes les informations sont exploitées au bénéfice du candidat et parce que le TEF comporte un nombre élevé d'items, supérieur à d'autres tests, ce qui améliore à la fois sa consistance interne et sa couverture. Naturellement, cette option est bien plus coûteuse en termes de développement puisqu'elle impose le recours à l'organisation de pré-tests, mais c'est le prix à payer de cette approche éthique. Par contre, l'exploitation *a posteriori*, consistant à ré-exploiter les résultats acquis par les candidats en situation réelle, permet d'assurer la stabilité et la comparabilité des différents jeux utilisés et la ré-estimation, sur des effectifs beaucoup plus conséquents[2], des paramètres d'ancrage, notamment. Le travail de validation est rendu possible par le recours au modèle de Rasch (Analyse MRI), notamment son implémentation dans les logiciels *Quest* (Adams & Khoo, 1993) et *ConQuest* (Wu, Adams & Wilson, 1998) de l'*Australian Council for Educational Research*.

[2] Pour des raisons de sécurité, chaque forme est utilisée, sans aucune variante, un nombre limité de fois. L'ordre de grandeur est de 1 500 candidats par forme, ce qui assure des possibilités importantes de validation *a posteriori.*

Le traitement et l'analyse des données issues de pré-tests

Les données des candidats aux pré-tests (deux jeux complets à quelques jours d'intervalle) permettent d'établir la fidélité, selon une méthode dite des "formes parallèles" (cf. tableau 1). Dans ce type de pré-test, les candidats sont répartis en deux groupes auxquels deux formes, réputées équivalentes, sont proposées dans un ordre différent, de manière à neutraliser les effets d'apprentissage et les effets d'ordre. A titre d'exemple, le tableau suivant présente l'indice de stabilité (corrélation de Bravais-Pearson), calculé lors de pré-tests sur différentes paires de jeux[3].

Tableau 1 – *Corrélations entre deux jeux lors des pré-tests et avant ajustement des différentes formes.*

N° des jeux	Corrélation	Nombre de candidats
50009 – 60010	0,88	57
30113 – 40114	0,84	68
50115 – 60116	0,87	81
30219 – 40220	0,91	77
50221 – 60222	0,89	77

La fidélité par consistance interne du TEF est également estimée lors des pré-tests et, de manière systématique, après la tenue d'une session TEF dans un centre agréé, de manière à contrôler l'unidimensionnalité du test et la possibilité d'établir un score. A titre d'exemple, l'analyse de huit formes parallèles du TEF, du numéro 30113 au 40220, est présentée dans le tableau 2. L'alpha de Cronbach permet de mettre en évidence une fidélité constante sur l'ensemble des 150 items des trois premières épreuves :

Tableau 2 – *Fidélité (consistance interne) mesurée à l'aide de l'Alpha de Cronbach sur 8 jeux (items 1 à 150). La consistance interne est satisfaisante si l'indice est compris entre 0,90 et 1,00 (valeur maximale).*

N° du jeu	30113	40114	50115	60116	10217	20218	30219	40220
Alpha de Cronbach	0,97	0,96	0,94	0,96	0,90	0,95	0,95	0,95

Une analyse identique, réalisée séparément sur les trois premières épreuves (compréhension écrite, compréhension orale, lexique/structure) de ces 8 jeux, permet généralement de mettre en évidence une fidélité très élevée, comme l'indique le tableau 3.

[3] Dans l'exemple ci-dessous, ce sont des corrélations entre formes non validées (pré-test) qui sont présentées, alors même qu'il s'agit d'une situation moins favorable que lorsqu'il s'agit de formes validées et dont on peut avoir éliminé certains items. Cette présentation a pour objectif de mettre en évidence la qualité des jeux fournis par les concepteurs, avant toute analyse psychométrique.

Tableau 3 – *Fidélité (Consistance interne) mesurée à l'aide de l'Alpha de Cronbach sur 8 jeux disponibles pour chacune des trois premières épreuves.*

N° du jeu	30113	40114	50115	60116	10217	20218	30219	40220
Compréhension écrite (CE) 50 items	0,91	0,90	0,81	0,90	0,69	0,88	0,89	0,89
Compréhension orale (CO) 60 items	0,93	0,92	0,89	0,92	0,82	0,90	0,91	0,92
Lexique et structure (LS) 40 items	0,87	0,87	0,80	0,87	0,75	0,81	0,80	0,80

L'application du Modèle de Réponse à l'Item[4] (MRI) entreprise sur l'ensemble des jeux, permet également d'améliorer la stabilité et l'équivalence des jeux à travers la sélection qui est effectuée dans la banque d'items validés. Cette banque est progressivement enrichie au fil de la création de nouveaux items et de leur validation lors des pré-tests. Le système d'ancrage entre les jeux constitue une garantie supplémentaire de stabilité à travers le temps. Périodiquement, certaines formes sont constituées exclusivement d'items appartenant à un ensemble réduit de formes (maximum 4) déjà validées de manière à contrôler au mieux toute dérive.

Une autre approche peut s'intéresser à la stabilité des erreurs standards de mesure à travers différentes formes parallèles. Cette estimation, effectuée pour huit formes parallèles, analysées *a posteriori* à partir de données obtenues sur des échantillons de candidats en situation réelle, est particulièrement stable à travers les différentes formes considérées puisque le calcul des coefficients de variation relatifs aux séries de huit erreurs de mesure conduit, respectivement à 0,025 pour la compréhension écrite, à 0,039 pour la compréhension orale et à 0,023 pour lexique/structure[5]. La taille des erreurs standards de mesure s'explique naturellement par la fidélité élevée qui est observée pour chacune des épreuves. Elle est par contre affectée négativement par la dispersion des résultats qui est particulièrement large, ce qui est normal dans la perspective de tests qui visent à discriminer au mieux entre les niveaux de compétences à travers une population très hétérogène. Il faut en effet rappeler que les estimations qui sont présentées ci-dessus ont été calculées sur des candidats en situation réelle et issus de centres très différents : aussi bien en Asie qu'en Europe, dans des pays francophones ou non, auprès de candidats déjà en situation d'immersion et d'étudiants qui n'ont eu que très peu d'heures de cours.

[4] Marc DEMEUSE : Les échelles de mesure : Thurstone, Likert, Guttman, le modèle de Rasch, Service de Pédagogie Expérimentale – Université de Liège, mars 2000.

[5] Ces erreurs standards de mesure, associées à l'estimation individuelle du niveau de compétence, ont été estimées à l'aide du logiciel ConQuest. Elles ne sont pas calculées sur la base de la formule classique (par exemple, Bernier & Pietrulewicz, 2000), mais au départ de l'*analyse MRI* (Wu, Adams & Wilson, 1998). La méthode utilisée à travers le logiciel ConQuest permet d'estimer l'erreur standard de mesure en tenant compte de l'information fournie par chaque item, ce que ne permet pas la méthode classique de calcul. Cela explique la stabilité de ces valeurs, alors que les écarts type de distribution qui sont fournis fluctuent en fonction de la plus ou moins grande disparité de compétences entre les candidats testés face à chacune des formes parallèles.

Le traitement et l'analyse des données issues de tests a posteriori

Une autre manière d'analyser les données peut encore être proposée. Ainsi, afin d'utiliser l'ensemble des données disponibles dans un autre échantillon de candidats qui ont passé le TEF en situation réelle et qui est actuellement analysé, on a considéré qu'il ne s'agissait que d'une seule forme de test, présentée aux 16.526 candidats et non de plusieurs, comme c'est effectivement le cas dans la réalité. Cette solution est naturellement défavorable au TEF, dans la mesure où la fidélité par consistance interne de l'ensemble est à la fois tributaire de la fidélité par consistance interne de chaque forme et de la variabilité qui peut exister entre chacune des formes parallèles. Mais, comme nous l'avons montré, la fidélité par consistance interne de chacune des formes parallèles, la stabilité et la comparabilité de celles-ci entre elles sont excellentes, ce qui conduit, lors de l'examen de cet échantillon de plus de seize mille candidats, à observer une fidélité globale de 0,9635 (alpha de Cronbach) pour l'ensemble des 150 items des deux épreuves de compréhension (orale et écrite) et de lexique/structure. Cette fidélité, toujours calculée sur le même échantillon, s'élève à 0,9128 (compréhension écrite), 0,9192 (compréhension orale) et 0,8665 (lexique/structure) si on considère séparément chaque épreuve.

Le TEF étant un test à référence critériée, il ne suppose aucune distribution *a priori* des compétences des candidats. Cette approche critériée est naturellement beaucoup plus complexe à mener que la construction d'une épreuve privilégiant une approche normative, surtout si celle-ci prend la forme d'un concours unique. En vue de répondre aux interrogations de Citoyenneté et Immigration Canada (CIC), un sous-échantillon a été constitué de manière aléatoire à partir des 16.526 candidats de manière à disposer de 800 candidats, tirés au hasard, pour chacun des niveaux du référentiel canadien (SLC) permettant d'acquérir 0, 1, 2 ou 4 points, soit 4 groupes et un total de 3.200 candidats. La répartition du niveau des candidats selon cette méthode, contrairement à la distribution observée sur l'ensemble de l'échantillon initial de seize mille candidats, était uniforme dans le sous-échantillon de 3 200. La fidélité, toujours calculée à partir de l'alpha de Cronbach[6], s'élève alors à 0,9287 (compréhension écrite), 0,9279 (compréhension orale) et 0,8964 (lexique/structure), ce qui est comparable aux données et aux résultats calculés sur l'échantillon de base. Cette stabilité de la fidélité, indépendante de la nature de la distribution des scores est naturellement importante, notamment parce que contrairement au développement d'un concours qui ne serait construit que pour une seule occasion, il est nécessaire de se préoccuper de la stabilité des scores à travers le temps : éviter toute dérive qui ne serait pas liée à une évolution des compétences réelles au sein de la population constitue une préoccupation essentielle des développeurs du TEF. Pour ce faire, il est donc essentiel de disposer de tables de spécification précises de manière à définir, puis à contrôler la production des items, avant même de les soumettre aux analyses psychométriques que nous venons d'évoquer. C'est le contrôle de ces procédures qui fait l'objet de la seconde partie de ce texte.

[6] En réalité, trois échantillons différents ont été sélectionnés puisque le niveau d'un candidat peut différer selon la compétence considérée.

2. Le contrôle de la qualité des procédures du test

Les activités de conception du TEF (conception et validation des items, éditions des jeux et sélection et évaluation des concepteurs, CTEF) et de réalisation du TEF (agrément des centres, organisation des sessions depuis l'envoi du matériel dans les centres agréés jusqu'à la délivrance des résultats, RTEF) ont été certifiées en janvier 2005 Iso 9001, version 2000, par l'Association française pour le management et l'amélioration de la qualité (AFAQ).

La préparation à la certification a demandé un remodelage des activités du TEF (cf. figures 1 et 2), donnant lieu à la naissance de deux grand "processus" : la conception (CTEF) et la réalisation (RTEF), la conception devant répondre aux besoins de RTEF, lui-même interface directe avec le client. En amont de la mise en place de ces processus, des analyses risques ont été menées par l'ensemble de l'équipe TEF afin d'envisager les points critiques de l'ensemble du système (CTEF/RTEF).

Figure 1 – *Cycle du Test d'Evaluation de Français et mesure de sa qualité.*

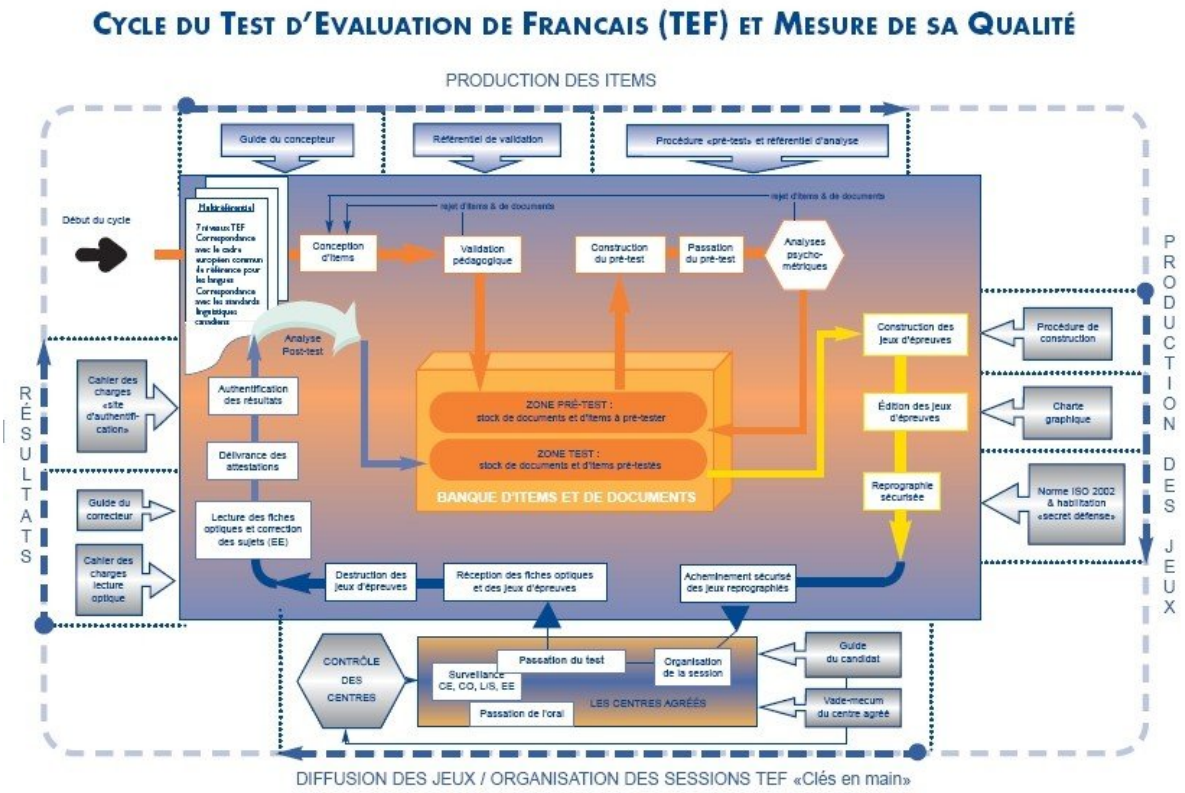
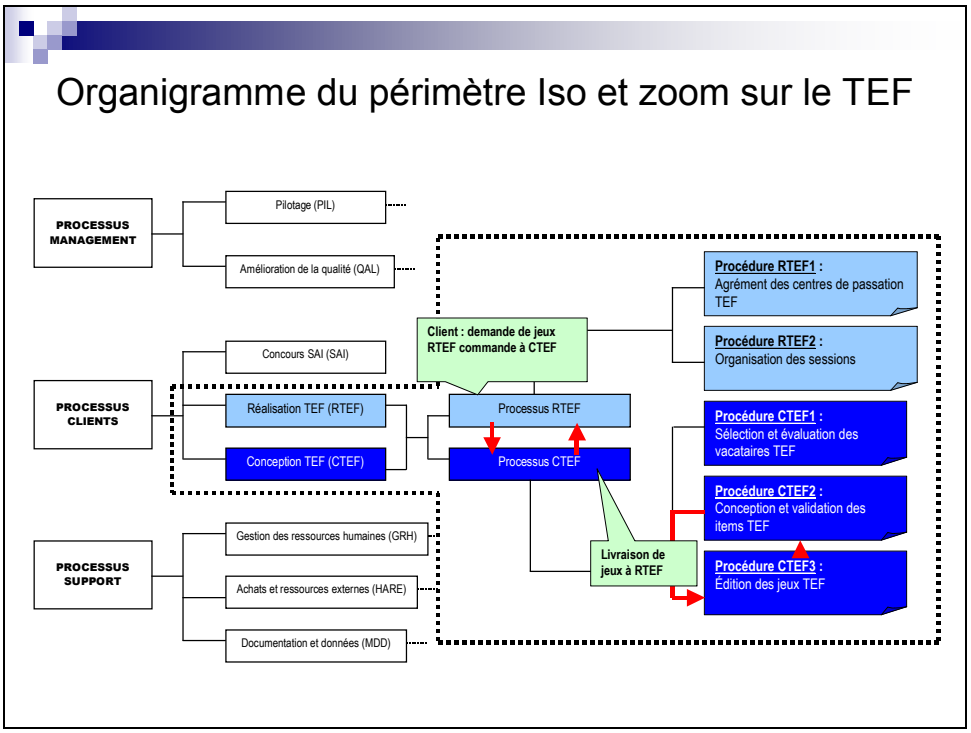


Figure 2 – *Organigramme du périmètre Iso et zoom sur le TEF.*



Les analyses risques partagées par l'ensemble de l'équipe du TEF, aussi bien administrative que pédagogique, font ressortir quelques indices exprimant le degré de gravité (de 1 à 4) et de fréquence (de 1 à 4) de cas de non-conformité sur le "processus CTEF". Comme l'indique le tableau 4, ces indices sont plus ou moins élevés et demandent à mener des actions préventives et/ou correctives. En effet, si la fiabilité du test est vérifiée en amont, il n'en reste pas moins que l'intensité de l'activité de production et les nombreux contrôles qualité au niveau des items demandent une planification précise : produire des jeux en nombre suffisant et de qualité. La variété des jeux est d'autant plus importante à assurer pour des centres qui organiseraient vingt-quatre sessions par an et pour des candidats qui se présenteraient à plusieurs sessions TEF par an. Le tableau 4 ci-dessous énumère quelques uns de ces points critiques.

Tableau 4 – *Points critiques CTEF. L'indice 16 se décompose en l'indice 4 pour la fréquence et 4 pour la gravité, la base de l'indice étant de 16.*

Points critiques	Indice "risques"
Défaut d'identification des besoins de RTEF	2
Défaut de concepteurs	6
Procédure de recrutement des concepteurs trop longue pour un besoin immédiat	9
Manque de nouveaux items car processus de validation trop long	16

Pour renforcer ce contrôle de la qualité, des évaluations "mixtes", comme celle présentée dans le tableau 5, sont menées sur la capacité de la conception TEF à répondre aux besoins de RTEF qui doit lui-même répondre à la commande des centres de passation du TEF. Par exemple, si RTEF commande 24 jeux d'épreuves à CTEF, alors pour répondre à la qualité attendue des items, aux respects des délais de livraison et à la quantité demandée, ce dernier doit anticiper la production de 3600 items, faire valider vingt-quatre pré-tests, recruter dix concepteurs et éditer vingt-quatre jeux.

Tableau 5 – *Evaluer la capacité de la conception TEF à répondre à la réalisation TEF.*

Évaluer la capacité de CTEF à répondre à RTEF

COMMANDES RTEF : HYPOTHESES			REPONSE DE CTEF EN TERMES DE PRODUCTION		
Nombre de jeux	Qualité	Délais	Qualité + + Délais + + Nombre + +	Qualité + + Délais ↘ Nombre + +	Qualité ↘ Délais + + Nombre + +
24	+ +	+ +	**Attendue** ↗ 3600 items produits ↗ 24 pré-tests validés ↗ 10 concepteurs recrutés ↗ 8 surveillants recrutés ↗ 24 jeux produits		
24	+ +	↘		**Altérée** ↗ priorité au module édition de la BDTEF ↗ 10 concepteurs à recruter ↗ 3 personnes à recruter validation ↗ Si délai d'édition, proposer e-tef	
24	↘	+ +			**Altérée** ↗ 4 ans de pré-tests (2880 items à valider) – 24 centres de 200 candidats à trouver immédiatement ↗ BDTEF à optimiser pour variété repérable des jeux ↗ 8 surveillants à recruter

Afin de surveiller et de maîtriser le processus CTEF, des indicateurs qualité ont été définis dans la "fiche analyse processus Iso", comme le nombre d'anomalies repérées sur les items par les centres. Bien que les pré-tests et les analyses *a posteriori* soient réalisées de façon à ce que les items soient bien calibrés, il arrive que des centres agréés signalent des anomalies sur ces items (faute de frappe, erreur de numérotation). Un des indicateurs porte par exemple sur le nombre de ces items à revalider ; un autre sur la durée de l'élaboration du test (depuis la conception des items jusqu'à leur validation) avec des cibles très précises à atteindre.

Ce contrôle de la qualité au niveau du macro-processus permet donc de renforcer la fiabilité de l'évaluation des compétences linguistiques. Il correspond par ailleurs au paragraphe 7.6. de la norme Iso qui demande qu'il soit mis en œuvre "*des processus efficaces et efficients de mesure et de surveillance, y compris des méthodes et des dispositifs de vérification et de validation des produits et processus pour assurer la satisfaction des clients et des autres parties intéressées. Ces processus incluent des enquêtes, simulations et autres activités de mesure et de surveillance.*" Les enjeux sur le respect de ces normes sont forts puisque l'audit de suivi peut signaler des écarts par rapport à la norme et, tous les trois ans, l'audit de certification peut révéler tant de dysfonctionnements que la certification n'est pas reconduite.

Ainsi, la norme Iso met bien en avant le critère économique ainsi que le critère temps qui font partie intégrante de la qualité perçue par le candidat, les centres agréés, les organismes certificateurs qui se doivent d'avoir un service irréprochable, voire une sélection équitable des candidats.

Conclusion

La réflexion autour de l'élaboration d'un test et de l'analyse des résultats met en évidence la nécessité d'une quadruple lecture, à la fois scientifique, pédagogique, économique et éthique. L'analyse d'échantillons de résultats, extraits de la base de données qui comporte actuellement plusieurs dizaines de milliers d'épreuves, présentées par des candidats qui s'inscrivent dans des contextes très variés (choix personnel, immigration économique, poursuite d'études à l'étranger…), met en évidence ces différentes approches et le choix des indicateurs nécessaires pour répondre aux différents interlocuteurs que sont les organismes de certification, les utilisateurs individuels, les utilisateurs institutionnels, les développeurs et les chercheurs. Cette analyse illustre, de manière concrète, la complémentarité entre les approches, à travers l'analyse de données réelles, mais aussi les différences qui peuvent apparaître entre les intérêts de chacun des acteurs.

Bibliographie

ADAMS, R.J., & KHOO, S.T. (1993). *QUEST. The Interactive Test Analysis System.* Melbourne : Australian Council for Educational Research.

CITOYENNETE ET IMMIGRATION CANADA (2002). *Standards Linguistiques Canadiens 2002*.

CONSEIL DE L'EUROPE (2000). *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer,* Paris : Didier.

CONSEIL DE L'EUROPE, DIVISION DES POLITIQUES LINGUISTIQUES (2003). *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.

DELAMOTTE, E. (1999). *Le commerce des langues*. Collection CREDIF Essais. Paris : Editions Didier.

DEMEUSE, M., CRENDAL, A., DESROCHES, F., OSTER, P., RENAUD, F., & LEROUX, X. (à paraître). « L'évaluation des compétences linguistiques des adultes en français langue étrangère dans une perspective de multiréférentialisation. L'exemple du Test d'Evaluation de Français (TEF) de la Chambre de Commerce et d'Industrie de Paris (CCIP) », In *Actes du 17e colloque international de l'Association pour le Développement des Méthodologies d'Evaluation en Education (ADMEE-Europe), Lisbonne, 18-20 novembre 2004*.

ENGLE, L., & ENGLE, J. (2003). « Assessing Language Acquisition and Intercultural Sensitivity Development in Relation to Study Abroad Program Design », *Frontiers : The Interdisciplinary Journal of Study Abroad, Vol. IX(Fall),* 219-236.

GOUGENHEIM, G., MICHEA, R., RIVENC, P., & SAUVAGEOT, A. (1956). *L'élaboration du français fondamental (premier degré)*, Paris : Didier.

Norme européenne – norme française, 1er tirage 2000-12-P

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