Analyse des différentes facettes influant sur la fidélité de l'épreuve d'expression écrite d'un test de français langue étrangère

Dominique Casanova
Chambre de commerce et d'industrie de Paris

Marc Demeuse
Université de Mons

MOTS CLÉS : Tests linguistiques à forts enjeux, expression écrite, français langue étrangère, fidélité, théorie de la généralisabilité, modèle multi-facettes de Rasch, contrôle de la qualité

Le contrôle de la qualité des évaluations des productions écrites en français langue étrangère pose de nombreuses questions, encore amplifiées lorsqu'il s'agit de tests à forts enjeux pour les candidats. Cet article illustre comment ce contrôle peut s'appuyer à la fois sur la théorie de la généralisabilité et sur le modèle multi-facettes de Rasch pour relever les sources d'erreur (stabilité inter- et intra-correcteurs, variation des stimuli fournis aux candidats pour produire leurs textes…) et en estimer l'importance respective dans le cadre d'un monitorage de la qualité d'une épreuve de français langue étrangère.

KEY WORDS: High stakes language tests, written production, French as foreign language, reliability, generalizability theory, many-facets Rasch model, quality monitoring

Quality control of written productions in French as a foreign language is a tricky issue, especially when the test is used to make high stakes decisions. This paper shows the complementarity of the generalizability theory and the many-facets Rasch model in order to identify and to estimate the size of the different sources of error in a quality monitoring process.

Palavras-chave: Testes linguísticos de nível elevado, expressão escrita, Francês como língua estrangeira, fidelidade, teoria da generalizabilidade, modelo multi-facetas de Rasch, controlo de qualidade

O controlo de qualidade das avaliações das produções escritas em Francês enquanto língua estrangeira coloca muitas questões, ainda para mais porque se trata de testes decisivos para os candidatos. Este artigo ilustra como este controlo se pode apoiar, ao mesmo tempo, na teoria da generalizabilidade e no modelo multi-facetas de Rasch para identificar as fontes de erro (estabilidade inter e intra-correctores, variação dos estímulos fornecidos aos candidatos para produzirem o seus textos…) e prever a importância respectiva no quadro de uma monitorização da qualidade de uma prova de Francês como língua estrangeira.

Note des auteurs – Les auteurs souhaitent remercier Jean Cardinet et Daniel Bain pour leur relecture et pour les remarques pertinentes qu'ils ont formulées sur le manuscrit initial. Toute correspondance peut être adressée comme suit : Dominique Casanova, Chambre de commerce et d'industrie de Paris, Direction des relations internationales de l'enseignement, Centre de langue française, 28 rue de l'Abbé Grégoire, 75279 Paris Cedex 06, France, téléphone : +33 1 49 54 17 37, télécopieur : +33 1 49 54 28 90, ou Marc Demeuse, Université de Mons, Faculté de psychologie et des sciences de l'éducation, Institut d'administration scolaire, Place du Parc, 18, B – 7000 Mons, Belgique, téléphone : +32 65 37 31 95, télécopieur : +32 65 37 37 74, ou par courriel aux adresses suivantes : [dcasanova@ccip.fr] ou [marc.demeuse@umons.ac.be].

Introduction

Dans le domaine de l'évaluation en langue, de nombreuses études concernant la fidélité des dispositifs d'évaluation de l'expression écrite ou orale se sont concentrées sur les correcteurs, en raison de la subjectivité que comporte tout jugement humain et sa sensibilité possible à des variables extérieures au contexte de l'évaluation (Artus & Demeuse, 2008). Il est vrai que les correcteurs constituent un maillon essentiel du dispositif d'évaluation d'une épreuve d'expression, qu'ils peuvent être sujets à des variations dans leurs jugements (manque de consistance interne) ou évaluer différemment un même ensemble de productions (différence de sévérité, sensibilité à des effets parasites).

Les raisons à cela sont multiples, car l'acte d'évaluation est complexe. Il intervient à un moment particulier de la journée, au sein d'une série de corrections, dans un environnement donné et est entrepris par un évaluateur qui apporte avec lui des émotions, une sensibilité, une vie personnelle qui peuvent interagir avec la ou les productions évaluées de manière singulière dans ce contexte précis, en dépit des précautions prises et des recommandations. L'évaluateur entreprend par ailleurs ce travail en s'appuyant sur des critères d'évaluation explicites lorsqu'il s'agit d'un test, et en référence à des niveaux de compétence généralement décrits dans une grille, qui peut être holistique ou analytique. Sa compréhension et son expérience d'utilisation d'une telle grille ont un impact direct sur la validité de son jugement et il importe qu'il bénéficie d'une formation appropriée et participe à des séances de standardisation avec d'autres correcteurs pour garantir la cohérence d'ensemble des corrections.

Cette variabilité des correcteurs justifie que leurs décisions soient fréquemment l'objet d'analyse, que ce soit pour la mise en évidence de différents profils de correcteurs (McNamara et Adams, 1991), le suivi des écarts de sévérité (Eckes, 2005) ou le contrôle de la fidélité intra et interévaluateurs (Weigle, 1998). Lorsque les enjeux de l'évaluation sont élevés, il importe de mettre en œuvre des stratégies d'évaluation qui permettent de réduire l'impact

de ces différences de jugement, par exemple des corrections multiples (Bachman, Lynch & Mason, 1995) ou le recours à des corrections statistiques (Eckes, 2009), et de s'assurer de leur pertinence.

Cependant, les correcteurs ne sont qu'une des facettes qui interviennent dans de tels dispositifs. D'autres facettes tout aussi importantes sont les tâches que le candidat doit réaliser et les sujets de l'épreuve, notamment dans le cas où des sessions multiples sont organisées et pour lesquelles des sujets d'épreuves différents sont utilisés. Des modèles de mesure particuliers, comme les modèles de généralisabilité ou les modèles de Rasch multifacettes peuvent alors être mobilisés pour tenir compte des contributions des différentes facettes à l'erreur de mesure. C'est ce que nous tâcherons de montrer dans cet article en nous appuyant sur des données empiriques provenant du Test d'évaluation de français (TEF) de la Chambre de commerce et d'industrie de Paris (CCIP, 2010 ; Noël-Jothy & Sampsonis, 2006, p. 74-75).

L'analyse multifacettes des dispositifs d'évaluation de l'expression écrite

Les principales facettes à considérer

Une épreuve d'expression écrite à réponse construite consiste en général en une ou plusieurs tâches à réaliser, à travers lesquelles les candidats produisent des textes qui sont évalués par un ou plusieurs correcteurs humains[1]. Les trois facettes essentielles de telles épreuves sont donc les candidats, les correcteurs et les tâches, et chacune de ces facettes est une source potentielle de variation des scores.

La variation des scores due aux candidats est liée à la compétence que l'on désire mesurer, mais d'autres facteurs contribuent généralement de manière non souhaitée à cette variation et alimentent l'erreur de mesure. La standardisation de l'épreuve et les consignes d'organisation permettent de limiter cette part de variance non désirée en plaçant les candidats dans des conditions de test similaires, mais les variations individuelles dues à des facteurs tels que la motivation et l'état de fatigue ne peuvent être complètement éliminées.

La variation des scores due aux correcteurs a fait l'objet de nombreuses études et est souvent considérée comme la principale source d'erreur de mesure. Elle peut être due à des facteurs relativement systématiques, tels qu'une différence de sévérité entre correcteurs, un effet de tendance centrale

qui conduit certains correcteurs à éviter de situer les candidats dans les niveaux extrêmes de l'échelle ou une interprétation différente, par les correcteurs, de la grille d'évaluation, que ce soit sur le plan des aspects de la performance à évaluer (critères d'évaluation) ou de la différenciation entre niveaux de performance. Elle peut être due également à des facteurs plus aléatoires, comme la sensibilité de certains correcteurs à l'effet d'ordre, ce qui les conduit à attribuer à un candidat un score qui dépend des copies corrigées précédemment, ou à l'effet de halo lorsque, du fait d'une particularité dans la copie ou d'une impression d'ensemble, ils ont tendance à attribuer un score similaire aux différents aspects de la production écrite du candidat[2] (Engelhard, 1994). Enfin, des facteurs tels que la fatigue ou l'état émotionnel peuvent aussi avoir des répercussions sur les scores délivrés. Pour limiter l'importance de la variation due aux correcteurs, il est nécessaire d'organiser régulièrement des sessions de formation et de standardisation (Lumley & McNamara, 1995) qui, si elles permettent notamment d'améliorer la consistance individuelle[3] et la validité des évaluations, ne permettent toutefois pas de faire disparaître les différences de sévérité entre correcteurs[4] (Weigle, 1998 ; Eckes, 2005). Dès lors, une attention constante est portée à la facette Correcteurs, en négligeant parfois le rôle des autres facettes dans la variation des scores.

Or, dans un test, les différents candidats sont soumis à des tâches identiques, mais à travers la réalisation de productions portant sur des sujets différents d'une session à l'autre. La définition des tâches s'effectue au regard du construit du test. Elles doivent permettre de recueillir un échantillon de productions suffisant pour pouvoir généraliser les compétences des candidats au domaine évalué par le test (Weigle, 2002, p. 60-61). Les différences de sujets introduisent une troisième source de variation des scores. Deux sujets peuvent en effet être de difficulté différente et conduire à des scores différents pour un même groupe de candidats. Une telle tendance systématique peut être mise en évidence à l'occasion de prétests, pour éviter que des sujets « trop faciles » ou « trop difficiles » ne soient utilisés.

Mais ce que l'on constate le plus souvent, c'est que si deux sujets conduisent à des résultats en moyenne semblables, certains candidats sont plus à l'aise avec le premier sujet et que les autres sont plus à l'aise avec le second. On voit alors apparaître une nouvelle source de variation des scores, due à l'interaction entre les candidats et les sujets de l'épreuve. Or, il s'agit d'une variation non souhaitée puisqu'elle met en cause la capacité de généraliser les résultats à une épreuve donnée aux différentes situations du domaine cible. Si limiter cette variation peut se révéler complexe ou coûteux, il est néanmoins

primordial d'en évaluer l'importance et d'en tenir compte dans l'estimation globale de la fidélité de l'épreuve, qui ne saurait se résumer à la fidélité intercorrecteurs.

De même, l'interaction entre correcteurs et candidats (par exemple, si un correcteur a tendance à être systématiquement plus sévère face à des écritures moins lisibles) est une source potentielle de variation des scores, en partie incluse, sans distinction, dans l'analyse de fidélité intercorrecteurs au moyen de la théorie classique des tests. L'interaction entre correcteurs et sujets (par exemple si un correcteur a tendance à noter plus sévèrement les copies en raisons d'opinions très arrêtées et très personnelles sur le sujet) peut également conduire à des variations de scores non désirées.

Enfin, l'interaction entre l'ensemble de ces facettes est elle-même une source potentielle d'erreur, un correcteur donné pouvant se montrer plus indulgent, pour un sujet particulier, avec un candidat dont les propos sont au diapason de ses propres convictions.

Les outils d'analyse

L'incapacité de la théorie classique des tests à faire la distinction entre les différentes sources d'erreur de mesure constitue l'une de ses principales limites (Bachman, 2004, p. 174-175 ; Bertrand & Blais, 2004, p. 71). Pour prendre en considération différentes sources d'erreur et étudier leur contribution relative à l'erreur de mesure, il faut donc recourir à d'autres approches, telles que la théorie de la généralisabilité ou les modèles de Rasch multifacettes.

La théorie de la généralisabilité

La théorie de la généralisabilité peut, dans une certaine mesure, être considérée comme une extension de la théorie classique des tests par l'application de certaines procédures d'analyse de variance (Brennan, 2001, p. 2). Les traitements de généralisabilité permettent notamment, par rapport à un plan de mesure déterminé (qui précise les facettes de différenciation, objet de la mesure, et les facettes d'instrumentation, instruments de la mesure), d'identifier les composantes de la variance d'erreur et d'estimer la fidélité du dispositif. Une phase d'optimisation permet également d'estimer la fidélité et les erreurs de mesure que l'on obtiendrait en faisant varier le niveau de différentes facettes. Cardinet et Tourneur (1985, p. 34-35) présentent en détail cette démarche.

La fidélité du dispositif est estimée au moyen du coefficient de généralisabilité, qui correspond à la part de la variance vraie (ou variance de différenciation) sur la variance totale (variance vraie + variance d'erreur). Cependant, les composantes de la variance d'erreur (ou variance d'instrumentation) dépendent du type d'erreur que l'on souhaite considérer : erreur relative ou erreur absolue. Dans le cas d'une épreuve destinée simplement à classer les candidats les uns par rapport aux autres, on considérera l'erreur relative qui ne tient compte que des effets d'interaction entre facettes (à partir du moment où les candidats d'une session passent tous le même sujet, peu importe si ce sujet est plus difficile que celui d'une autre session). Mais si, comme c'est le cas pour le Test d'évaluation de français, l'épreuve est destinée à situer les candidats sur une échelle de niveaux et que des décisions sont prises selon le niveau obtenu, alors on considérera l'erreur absolue qui tient compte également des effets directs des facettes d'instrumentation (dans ce cas, les différences de difficulté des sujets peuvent avoir un impact direct sur le niveau attribué aux candidats et donc sur la prise de décision).

La théorie de la généralisabilité offre donc un cadre intéressant pour le traitement de dispositifs d'évaluation qui comportent plusieurs facettes, comme c'est le cas pour l'évaluation des compétences en expression écrite. Cependant, sa mise en œuvre nécessite généralement le recours à des plans équilibrés (il est alors nécessaire de disposer d'une mesure pour chaque combinaison de niveaux des facettes considérées) qui sont souvent difficiles à constituer en dehors de situations expérimentales. Par ailleurs, comme le font remarquer Bertrand et Blais (2004, p. 97), un nombre important d'observations est à recueillir si on veut que les erreurs-type des composantes de variance soient raisonnablement petites, ce qui peut rendre coûteuse la mise en œuvre d'expérimentations *ad hoc*.

Le modèle multifacettes de Rasch

En phase de production, on privilégiera donc le recours à des modèles qui, comme le modèle multifacettes de Rasch introduit par Linacre (1989), permettent l'utilisation de plans partiellement équilibrés.

Le modèle multifacettes de Rasch a été mis en œuvre à différentes occasions dans le domaine de l'évaluation en langue, pour mettre en évidence l'impact des facettes *Correcteurs* et *Sujets* dans l'évaluation (Bachman et al., 1995), pour analyser différentes caractéristiques des correcteurs (sévérité, exploitation de l'échelle de notation, différence d'application de l'échelle selon le critère d'évaluation) (McNamara & Adams, 1991) ou pour mesurer l'effet

de la formation des correcteurs (Weigle, 1994). Outre l'estimation des différences de sévérité ou de difficulté, il permet de contrôler également la consistance des évaluations au moyen d'indices d'ajustements[5].

Les conditions d'application de ce modèle ne sont cependant pas toujours aisées à réunir. Sa mise en œuvre nécessite avant tout de faire le choix d'un plan d'analyse des données approprié, qui s'appuie sur une identification des facettes à considérer (correcteurs, sujets, etc.) et qui tient compte ou non de l'interaction entre les différentes facettes. Plus le nombre de facettes et d'interactions entre facettes pris en compte sera élevé, plus il sera nécessaire d'accumuler des données pour obtenir des estimations précises. Il faut par ailleurs que ces données soient suffisamment interreliées pour pouvoir calibrer les différents éléments (correcteurs, sujets, etc.) sur des échelles de mesure propres à chaque facette (échelle de sévérité, échelle de difficulté, etc.).

Le cas de l'épreuve d'expression écrite du Test d'évaluation de français (TEF)

Le Test d'évaluation de français (TEF) est un test à forts enjeux. Il est notamment utilisé pour l'évaluation du niveau de langue française des candidats à l'immigration économique au Canada et est reconnu par le ministère français de l'Enseignement supérieur et de la Recherche pour dispense de l'examen officiel auquel doivent se soumettre les étudiants étrangers candidats à une première inscription en premier cycle d'études universitaires en France. Ainsi, toute erreur dans le positionnement d'un candidat sur l'échelle de niveaux du TEF pourra conduire à une acceptation ou à un rejet erronés d'un dossier d'immigration, raison pour laquelle la Chambre de commerce et d'industrie de Paris doit apporter des garanties suffisantes sur la qualité du dispositif d'évaluation (Demeuse, Desroches, Crendal, Renaud & Casanova, 2005 ; Holle, à paraître). C'est dans ce cadre qu'a été menée cette étude.

L'épreuve d'expression écrite du TEF est constituée de deux tâches indépendantes, qui placent les candidats dans deux situations de communication différentes. Dans la première de ces situations, le candidat doit *raconter une histoire* en imaginant la fin d'un article de presse insolite, alors que dans la seconde situation, le candidat doit *exposer son point de vue et argumenter* en réponse à une affirmation lue dans la presse. Un jeu d'épreuve comportera donc deux sujets (correspondant à chacune des deux tâches), chaque sujet étant constitué d'un stimulus et de consignes conduisant à une production de la part des candidats.

Les productions sont toutes deux évaluées par un même jury selon :

- trois critères communicatifs propres à chacune des deux tâches ;
- six critères linguistiques s'appliquant à l'ensemble du contenu des deux productions.

Le jury est constitué de deux correcteurs, qui évaluent individuellement les copies, et d'un arbitre, qui décide de la note finale à accorder au candidat pour chacun des critères[6]. Après arbitrage, les notes finales relatives à chacun des critères sont combinées, selon un système de pondération, pour aboutir à l'expression d'un score total et d'un niveau global de compétence en expression écrite, qui permet de situer le candidat sur l'échelle de niveaux du Cadre européen commun de référence pour les langues (Conseil de l'Europe, 2005) et sur l'échelle des Niveaux de compétence linguistique canadiens (Ministère de la Citoyenneté et de l'Immigration Canada, 2006), à la suite d'un double travail d'indexation (Demeuse, Desroches, Crendal, Renaud, Oster & Leroux, 2004 ; Casanova, Crendal, Demeuse, Desroches & Holle, 2010).

L'objectif de la présente étude étude est de distinguer les différentes sources d'erreur affectant le dispositif d'évaluation de l'épreuve d'expression écrite du TEF et leur contribution relative à l'erreur de mesure globale, et ce, dans plusieurs buts :

- estimer la précision avec laquelle le TEF permet d'attribuer un niveau aux candidats sur l'échelle de référence ;
- identifier des leviers pertinents pour l'amélioration de la fidélité du dispositif d'évaluation.

Hypothèses de recherche

Les hypothèses que nous avons cherché à vérifier dans le cadre de cette recherche sont les suivantes :

1. l'erreur de mesure de l'épreuve d'expression écrite permet de positionner les candidats sur l'échelle avec suffisamment de précision ;
2. s'il peut exister des différences de sévérité entre correcteurs, la différence de sévérité entre deux jurys distincts est non significative ;
3. la différence de difficulté entre deux jeux d'épreuve est non significative ;
4. le classement des candidats est identique d'un jeu d'épreuve à l'autre ;
5. les différents sujets se rapportant à une même tâche sont de difficulté identique ;

6. pour une même tâche, le classement des candidats est identique d'un sujet à l'autre ;
7. la sévérité relative des correcteurs est indépendante de la tâche évaluée ;
8. pour une même tâche, la sévérité relative des correcteurs est indépendante du sujet évalué.

Méthodologie

Le choix des modèles

Si le modèle de généralisablité semble *a priori* le plus approprié pour analyser la contribution de différentes sources d'erreur à l'erreur de mesure et estimer la fidélité du dispositif d'évaluation de l'expression écrite du TEF, l'organisation de cette épreuve et la procédure de correction associée ne permettent pas de mener aisément de telles études de manière routinière.

Pour mener une analyse multifacettes de l'épreuve d'expression écrite du TEF, au moyen de la théorie de la généralisabilité, qui permette de vérifier nos hypothèses, il nous a donc fallu recourir à une expérimentation *ad hoc*, dans le cadre de prétests. Différents plans d'étude de généralisabilité ont ainsi été mis en œuvre pour analyser les différentes contributions à l'erreur de mesure, repérer des différences de sévérité ou de difficulté, estimer la fidélité de l'épreuve et mettre en évidence des pistes éventuelles pour l'amélioration du dispositif.

Les analyses de généralisabilité ont été complétées par un recours au modèle de Rasch multifacettes de manière à obtenir une information synthétique sur les écarts de sévérité entre correcteurs (permettant de tenir à jour un panel de correcteurs calibrés) et sur la consistance de leurs évaluations.

Cependant, l'application du modèle de Rasch multifacettes dans le cadre de l'évaluation de l'épreuve d'expression écrite du TEF nécessite une transformation des données résultats qui conduit à une réduction de l'information. La mise en œuvre de ce modèle nécessite en effet que l'on dispose, pour chaque jugement, d'au moins une observation par degré de l'échelle de notation. Les scores totaux des candidats, délivrés sur une échelle de 450 points, ne pouvaient dès lors être utilisés tels quels, et il a fallu se résoudre de raisonner en termes de niveaux (l'échelle du TEF comporte sept niveaux).

Description de l'expérimentation

Échantillon

Pour cette étude, nous avons convié une cinquantaine d'étudiants non francophones d'une université parisienne à passer successivement deux épreuves d'expression écrite du TEF (avec une pause de 15 minutes entre chaque épreuve). Pour encourager leur participation à cette expérimentation, nous leur avons par ailleurs offert la possibilité de passer gratuitement l'ensemble des épreuves du TEF et d'obtenir ainsi une attestation officielle de leurs résultats. Cependant, seuls 36 étudiants se sont effectivement présentés le jour de l'expérimentation. Ils ont été répartis en deux groupes de 18 candidats, qui ont passé successivement, en ordre alterné, les deux épreuves d'expression écrite du TEF (avec une pause de 15 minutes entre chaque épreuve). Trois candidats n'ayant traité que l'un des deux sujets de la première épreuve ont été retirés de l'échantillon. L'échantillon global est donc constitué de 33 candidats, soit 66 copies. Si la taille réduite de cet échantillon peut limiter la portée des conclusions générales, elle permet néanmoins d'appréhender les risques en phase de mise au point et de déterminer les sources potentielles d'erreur de mesure.

Par ailleurs, la plupart des candidats se sont vu attribuer un niveau 3 ou 4 sur l'échelle, qui comporte sept niveaux. Cela nous a conduits, pour pouvoir mettre en œuvre les modèles de Rasch multifacettes à partir des niveaux des candidats et disposer également d'un échantillon de données suffisamment large, à compléter l'échantillon initial en faisant corriger par chacun des quatre correcteurs 60 nouvelles copies produites en situation réelle par des candidats de niveaux variés (30 pour chacun des deux jeux d'épreuves utilisés dans l'expérimentation) et à utiliser le plan de données partiellement connectées décrit dans la figure 1.

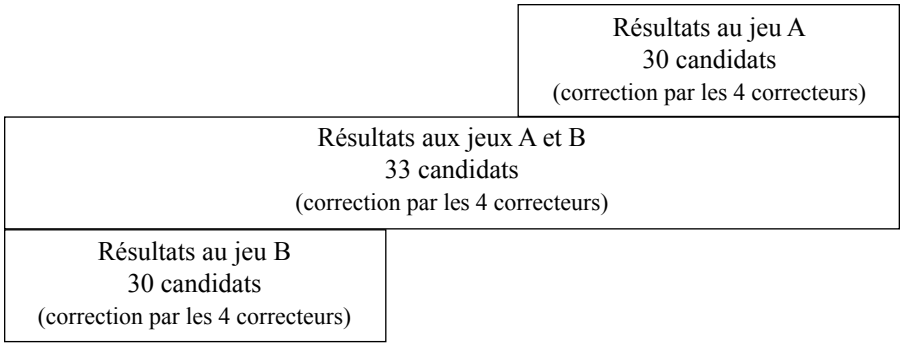

Figure 1. ***Nature de l'échantillon de données utilisé pour les analyses Rasch multifacettes***

Méthode

Chaque copie de l'expérimentation a été corrigée par quatre correcteurs, regroupés en deux jurys (un arbitre différent a été affecté à chacun des jurys). Les analyses de généralisabilité ont été menées au moyen du logiciel EduG 6.0[7] (IRDP, 2010), et les modèles de Rasch multifacettes ont été mis en œuvre au moyen du logiciel CONQUEST 2.0 (Wu, Adams, Wilson & Haldane, 2007).

Résultats

Statistiques descriptives

Le tableau 1 décrit la répartition des scores et des niveaux des 33 candidats ayant participé à l'expérimentation, selon les correcteurs, les jurys, les jeux ou l'ensemble des corrections.

Tableau 1
Répartition des scores et niveaux des candidats

	Scores (sur 450)		*Nombre de candidats par niveau*			
Copies considérées	*Moyenne*	*Écart type*	*Niveau 2*	*Niveau 3*	*Niveau 4*	*Niveau 5*
Toutes les copies arbitrées	267,0	51,5	2	12	19	0
Toutes sans arbitrage	264,1	50,1	2	13	18	0
Jeu A, arbitrées	262,0	55,5	3	13	17	0
Jeu B, arbitrées	271,9	51,7	3	9	21	0
Jury 1	268,0	56,3	2	11	20	0
Jury 2	265,9	47,9	3	9	21	0
Correcteur 1	298,1	54,5	1	9	20	3
Correcteur 2	262,5	49,3	3	12	18	0
Correcteur 3	233,5	54,5	10	13	10	0
Correcteur 4	262,4	48,4	2	13	18	0

On voit que les classements par niveau peuvent être très variables selon les correcteurs, le correcteur 3 étant sensiblement plus sévère et le correcteur 1 plus indulgent ; mais quand on constitue des jurys équilibrés (le jury 1 est composé des correcteurs 1 et 3 et le jury 2 des correcteurs 2 et 4), on obtient, conformément à l'hypothèse 2, des résultats proches en moyenne ($t = 0,674$, avec $p = 0,5054$), mais qui n'empêchent pas des différences de classement pour des candidats se situant à la frontière entre deux niveaux. Par ailleurs, si le jeu d'épreuve semble avoir une influence non négligeable sur le classement des candidats, l'hypothèse nulle selon laquelle la différence de moyenne serait due au hasard ne peut pas être rejetée ($t = -1,906$, avec $p = 0,0657$) au risque α de 5%.

Il faut aussi signaler que l'écart de moyenne entre les deux passations[8] est faible (3,818 points sur un total de 450) et non significatif ($t = -0,263$ avec $p = 0,7943$). Il ne semble donc pas y avoir eu d'effet d'ordre manifeste.

Estimation de la fidélité au moyen de la théorie classique

La fidélité interjurys, calculée sur les 66 copies, est très élevée (0,937 alors que l'amplitude des différences de scores de candidats est inférieure à la moitié de l'amplitude de l'échelle). Cependant, cette fidélité est établie en faisant corriger les mêmes copies par chacun des deux jurys. Or, en situation réelle, si un candidat était amené à repasser le test, il se verrait attribuer un jeu d'épreuve différent, qui serait très probablement corrigé par un jury

différent. Le prétest permet d'estimer la fidélité globale de l'épreuve en répartissant les résultats en deux sous-échantillons. Dans le premier cas, on considère :

- d'une part, les résultats que le jury 1 attribue aux candidats pour le jeu A ;
- et d'autre part, les résultats que le jury 2 leur attribue pour le jeu B,

et inversement, pour le second cas :

- d'une part, les résultats que le jury 2 attribue aux candidats pour le jeu A ;
- d'autre part, les résultats que le jury 1 leur attribue pour le jeu B.

On calcule ainsi, pour chacun des deux cas, la corrélation entre les scores obtenus lors de deux passations successives d'une épreuve différente d'expression écrite corrigées par des jurys différents, soit deux évaluations réellement indépendantes. Cette corrélation s'élève à 0,843 dans le premier cas et à 0,780 dans le second. Ces deux corrélations sont significatives ($p < 0{,}0001$).

On constate donc que l'estimation de la fidélité globale est sensiblement plus faible que la fidélité interjurys et, par conséquent, que les jeux d'épreuve ont un impact sur la fidélité de l'évaluation. Les analyses menées au moyen de la théorie de la généralisabilité permettent d'établir plus précisément les différentes sources d'erreur et de déterminer leur contribution relative à la variance d'erreur.

Analyse des différentes facettes au moyen de la théorie de la généralisabilité

Analyse des facettes principales

Le plan d'étude le plus classique consiste à croiser les facettes *Candidats*, *Correcteurs* et *Jeux d'épreuve*. Il modélise le cas où des candidats se voient attribuer aléatoirement un jeu pour une session, où leurs copies sont corrigées par des correcteurs indépendants, sélectionnés aléatoirement, et où le score délivré au candidat correspond à la moyenne des scores délivrés par chacun des correcteurs.

S'il est réaliste en ce qui concerne les candidats et les jeux d'épreuve (le choix des différents niveaux de ces facettes n'a rien de spécifique), ce plan d'étude ne reflète pas complètement la situation du TEF, où les jurys ne sont pas constitués aléatoirement mais en tenant compte de la sévérité relative des correcteurs, et où le score délivré au candidat résulte d'une phase d'arbitrage.

Néanmoins, il présente l'intérêt de montrer l'impact qu'aurait la facette *Correcteurs* si les jurys étaient constitués aléatoirement et d'estimer les effets de son interaction avec les autres facettes.

Le tableau 2 présente les plans d'observation et d'estimation utilisés et le tableau 3 les résultats de l'analyse de variance menée à partir des scores globaux des candidats, tels que produits par EduG. Le tableau 3 permet de montrer que, en dépit de la diversité modérée de compétence linguistique des candidats[9], les différences de compétence entre candidats expliquent tout de même 59,7 % de la variance totale.

Tableau 2
Plan d'observation et d'estimation

Facette	*Étiquette*	*Niveaux*	*Univers*
Candidats	C	33	INF
Jeux d'épreuve	S	2	INF
Correcteurs	E	4	INF

Tableau 3
Analyse de variance pour le plan croisé Candidats(C) X Jeux (S) X Correcteurs (E)

				Composantes				
Source	*Somme des carrés*	*Degrés de liberté*	*Carrés moyens*	*Aléatoires*	*Mixtes*	*Corrigées*	*%*	*Er. St.*
C	642668,25	32	20083,38	2252,21	2252,21	2252,21	59,7	611,70
S	7245,03	1	7245,03	26,73	26,73	26,73	0,7	46,06
E	138528,86	3	46176,29	665,55	665,55	665,55	17,6	442,95
CS	61228,34	32	1913,39	404,64	404,66	404,64	10,7	116,49
CE	42928,27	96	447,17	76,16	76,16	76,16	2,0	38,26
SE	6293,34	3	2097,78	54,63	54,63	54,63	1,4	40,22
CSE	28304,78	96	294,84	294,84	294,84	294,84	7,8	42,12
Total	927196,88	263					100%	

Comme nous nous intéressons à l'influence des différentes sources d'erreur sur les résultats des candidats, nous avons adopté un plan de mesure qui considère la facette *Candidats (C)* comme facette de différenciation et les facettes *Jeux d'épreuve (S)* et *Correcteurs (E)* comme facettes d'instrumentation. Le tableau 4 présente la répartition de la variance d'erreur absolue entre les différentes facettes et leurs interactions.

Tableau 4
Analyse de généralisabilité pour le plan de mesure C/SE

Sources de var.	*Variance de différ.*	*Sources de var.*	*Variance d'err. abs.*	*% abs.*
C (candidats)	2252,21	.....		
	.....	S (jeux)	13,37	3,0
	.....	E (correcteurs)	166,39	37,4
	.....	CS	202,32	45,5
	.....	CE	19,04	4,3
	.....	SE	6,83	1,5
	.....	CSE	36,86	8,3
Total des variances	2252,21		444,80	100%
Écarts types	47,46		Erreur type absolue : 21,09	
Coef_G absolu		0,84		

Ce tableau met en évidence la présence d'un effet *Correcteurs* important, qui explique à lui seul 37,4 % de la variance d'erreur absolue. On voit donc que les différences de sévérité entre correcteurs contribuent de manière appréciable à l'erreur de mesure, ce qui plaide pour la mise en œuvre d'un dispositif de monitorage permettant de tenir compte de cet aspect.

L'effet *Jeux* est pour sa part faible (3 % de la variance d'erreur), témoignant que dans l'ensemble, la différence apparente de difficulté des jeux ne joue qu'un rôle marginal, bien que significatif, sur l'erreur de mesure absolue (hypothèse 3). Cependant, l'effet d'interaction entre *Candidats* et *Jeux* (CS) explique la plus grande partie de la variance d'erreur (45,5%), ce qui réfute clairement l'hypothèse 4, selon laquelle « le classement des candidats est identique d'un jeu d'épreuve à l'autre »[10]. Cela montre probablement que les candidats peuvent être plus ou moins à l'aise selon les thématiques des sujets

(les tâches sont identiques d'un jeu à l'autre, mais réalisées à partir de sujets différents), mais l'interaction entre *Candidats* et *Jeux* englobe également les variations dues à d'éventuels changements de stratégie des candidats entre les deux passations successives, ce qui conduit sans doute à une surestimation de l'effet d'interaction.

Il est par ailleurs intéressant de constater que l'effet d'interaction entre les facettes *Candidats* et *Correcteurs* (*CE*) n'explique que 4,3 % de la variance d'erreur et l'effet d'interaction entre les facettes *Jeux d'épreuve* et *Correcteurs* (*SE*) seulement 1,5%. Ainsi, les correcteurs montrent globalement une bonne stabilité dans leurs évaluations, indépendamment du jeu traité et du candidat concerné.

Le coefficient absolu de généralisabilité (IRDP, 2010, p. 38-39) s'élève à 0,84, mais il est établi pour l'évaluation d'un candidat au moyen de deux jeux d'épreuve et de quatre correcteurs, ce qui ne reflète pas les conditions réelles de passation du TEF. Par ailleurs, sa valeur est tributaire de la répartition des scores des candidats et il convient davantage de considérer l'erreur type absolue, comme le recommande Cronbach:

> I am convinced that the standard error of measurement […] is the most important single piece of information to report regarding an instrument, and not a coefficient (Cronbach & Shavelson, 2004, p. 413).

C'est à partir de cette erreur type qu'on pourra déterminer un intervalle de confiance autour des points de césure entre niveaux. Compte tenu des décisions qui sont prises par les utilisateurs institutionnels, à partir des niveaux obtenus par les candidats au TEF, il importe en effet de limiter le risque qu'un candidat ne se voit classé dans un niveau différent de son niveau réel. Ce risque est maximal pour les candidats se situant à la césure entre deux niveaux adjacents. La précision de l'outil sera jugée acceptable si l'écart entre le score attribué au candidat et son score réel (score univers dans la terminologie de la théorie de la généralisabilité) ne diffère pas de plus de un niveau dans plus de 5 % des cas.

Une étude de décision menée avec un plan d'optimisation à un jeu d'épreuve et à deux correcteurs conduit à une estimation de l'erreur type absolue de 31,3 points (et à un coefficient de généralisabilité absolue de 0,697), soit 6,96 % de l'amplitude totale de l'échelle de scores (de 0 à 450 points). Cette erreur de mesure étant inférieure à la moitié de l'amplitude de chacun des niveaux du TEF (l'amplitude minimale est de 67 points pour les niveaux concernés[11]), le risque d'une erreur importante de classement des candidats

(plus de un niveau d'écart entre le niveau attribué aux candidats et leur niveau réel) reste inférieur à 5 %. Il serait toutefois souhaitable de disposer d'un échantillon plus important pour en garantir la stabilité.

Ces résultats montrent cependant que les différences de sévérité entre correcteurs ont un impact sensible sur la fidélité de l'épreuve et qu'il importe également de tenir compte de l'interaction entre candidats et jeux d'épreuve.

Prise en considération de l'organisation en jurys dans l'estimation de la fidélité

Afin de tenir compte de l'arbitrage et de la stratégie d'appariement des correcteurs mise en œuvre pour parvenir à des jurys équilibrés, une étude de généralisabilité a été menée à partir du plan croisé *Candidats (C) X Jeux (S) X Jurys (J)*, chaque facette étant considérée comme aléatoire infinie, reflétant le fait que les jurys sont tirés aléatoirement parmi l'ensemble des jurys équilibrés envisageables pour la correction du TEF. Elle s'appuie sur le score final délivré par chacun des deux jurys à chacun des 33 candidats pour chacun des deux jeux d'épreuve et ne tient donc pas compte des différences de scores entre correcteurs au sein d'un même jury.

L'analyse de variance montre que les différences de compétence entre candidats expliquent alors 78,6 % de la variance totale. Le tableau 5 permet quant à lui d'établir clairement la principale source d'erreur, à savoir l'effet d'interaction entre *Candidats* et *Jeux*, qui explique 65,4 % de la variance d'erreur absolue.

Tableau 5
Analyse de généralisabilité avec le plan croisé Candidats(C) X Jeux (S) X Jurys (J)

Sources de var.	*Variance de différ.*	*Sources de var.*	*Variance d'err. abs.*	*% abs.*
C (Candidats)	2381,95		.....	
	.....	S (jeux)	16,72	5,8
	.....	J (jurys)	(0.00000)	0,0
	.....	CS	188,54	65,4
	.....	CJ	47,37	16,4
	.....	SJ	1,07	0,4
	.....	CSJ	34,44	12,0
Total des variances	2381,95		288,15	100%
Écarts types	48,81		Erreur type absolue : 16,97	
Coef_G absolu		0,89		

Par ailleurs, alors que l'effet d'interaction entre *Candidats* et *Correcteurs* était faible (*cf*. tableau 4), l'effet d'interaction entre *Candidats* et *Jurys* (CJ) est sensiblement plus important (16,4%). Ainsi, si les jurys délivrent en moyenne des résultats comparables aux candidats (aucun effet *Jurys* ne peut être mis en évidence), ils ne les classent pas toujours de la même manière (la corrélation est toutefois très élevée : 0,937). La facette *Jeux d'épreuve* n'explique quant à elle directement que 5,8 % de la variance d'erreur absolue. Cependant, si cela indique que les deux jeux utilisés sont globalement de difficulté semblable, il n'est pas possible à ce stade de déterminer, pour chacune des deux tâches qui composent l'épreuve, si les sujets utilisés dans les jeux d'épreuve sont de difficulté comparable.

Une étude de décision pour un plan d'optimisation à un jury et un jeu d'épreuve permet d'obtenir une estimation de la fidélité du dispositif. Le coefficient de généralisabilité absolue est alors de 0,786 et l'erreur type absolue de 25,4 points (soit 5,6 % du nombre maximum de points accordés pour cette épreuve, c'est-à-dire 450 points). Ainsi, si l'interaction entre *Candidats* et *Jeux d'épreuve* explique les deux tiers de la variance d'erreur, l'erreur type absolue correspondante est plutôt faible et en tout cas plus favorable que lorsqu'on prend en compte isolément les évaluations des correcteurs.

Difficulté relative des sujets, sur le plan communicatif

Afin d'analyser, pour chacune des deux tâches, la difficulté relative des sujets sur le plan communicatif, on met en œuvre les deux études de généralisabilité suivantes :

- la première considère, en plan croisé, les notes (sur 20) accordées pour chacun des trois premiers critères (se rapportant à la première tâche) à chacune des copies d'un candidat par chacun des correcteurs ;
- la seconde considère, en plan croisé, les notes accordées aux critères 4 à 6 (se rapportant à la seconde tâche) à chacune des copies d'un candidat par chacun des correcteurs.

Le tableau 6 présente les plans d'observation et d'estimation correspondants (les mêmes plans sont utilisés pour chacune des deux tâches), où les critères sont considérés comme une facette fixe (une même tâche est toujours évaluée à partir des mêmes trois critères communicatifs). Les sujets sont, pour leur part, tirés au hasard parmi une banque de sujets potentiellement infinie pour la tâche considérée et les correcteurs sont considérés comme choisis aléatoirement parmi un ensemble potentiellement infini.

Tableau 6
Plans d'observation et d'estimation

Facette	*Étiquette*	*Niveaux*	*Univers*
Candidats	C	33	INF
Sujets	S	2	INF
Critères	A	3	3
Correcteurs	E	4	INF

L'effet *Sujets* qui peut ainsi être mis en évidence est très limité : il n'explique que 1,2 % de la variance d'erreur absolue dans le premier cas et 2,3 % dans le second cas. Ainsi, aucun des sujets ne semble présenter une difficulté intrinsèque supérieure à l'autre sur le plan communicatif, tant pour la première que pour la seconde tâche, ce qui confirme l'hypothèse selon laquelle « les différents sujets se rapportant à une même tâche sont de difficulté identique » (hypothèse 5). En revanche, l'effet d'interaction entre *Candidats* et *Sujets* explique la plus grande part de la variance d'erreur (56,7 % dans un cas, 54,9 % dans l'autre)[12], certains candidats étant manifestement plus à l'aise que

d'autres pour certaines thématiques, résultat qui contredit l'hypothèse 6 selon laquelle « pour une même tâche, le classement des candidats est identique d'un sujet à l'autre».

La seconde source de variation est liée à l'effet *Correcteurs*, qui explique respectivement 27 % et 23,9 % de la variance d'erreur, l'effet d'interaction entre *Candidats* et *Correcteurs* étant limité (3,2 % et 6,2 % de la variance d'erreur) et l'interaction entre *Correcteurs* et *Sujets* quasi nulle, ce qui confirme l'hypothèse 8 selon laquelle « pour une même tâche, la sévérité relative des correcteurs est indépendante du sujet évalué».

Il ne semble donc pas y avoir de sujets plus difficiles que d'autres (du moins pour la réalisation de la tâche sur le plan communicatif) et les correcteurs montrent une bonne stabilité dans leurs évaluations, indépendamment du sujet traité et du candidat concerné. Cependant, certains candidats semblent plus à l'aise que d'autres pour certaines thématiques. Cela plaide en faveur de l'introduction d'une tâche supplémentaire, qui exposerait les candidats à davantage de thématiques et améliorerait probablement la fidélité de l'épreuve d'expression écrite du TEF, mais l'allongerait sensiblement.

Amélioration de la fidélité par l'ajout d'un correcteur ou d'une tâche supplémentaire

On peut s'attendre à ce que les deux tâches de l'épreuve d'expression écrite classent les candidats de manière sensiblement différente. Cependant, nous avons montré la présence d'un effet important d'interaction entre *Candidats* et *Sujets*. Aussi, lorsqu'on mène une étude de généralisabilité par jeu d'épreuve en croisant les facettes *Candidats*, *Correcteurs* (ou *Jurys*) et *Tâches* (facette nichante de la facette *Critères*) et en se basant sur les scores délivrés à chacun des critères communicatifs, l'effet d'interaction qui peut être mis en évidence entre les facettes *Candidats* et *Tâches* (45,5 % et 46,9 % selon le jeu d'épreuve considéré) englobe-t-il l'interaction entre les candidats et la thématique des sujets (instances des tâches pour le jeu considéré) ?

À partir des résultats de l'expérimentation, on peut comparer les gains de fidélité que l'on pourrait espérer en considérant, d'une part, trois correcteurs ou, d'autre part, trois tâches (et donc trois thématiques). Compte tenu de la procédure de correction du TEF (l'évaluation des critères linguistiques s'effectue à partir de l'ensemble des deux productions), ce traitement ne peut être mené qu'en considérant les critères communicatifs (différents pour les deux tâches) et en faisant abstraction de la pondération de chaque critère (noté sur 20).

Le résultat des études de décision, qui prennent en considération, pour chacun des jeux pris séparément, les facettes *Candidats*, *Correcteurs, Tâches* et *Critères* (cette dernière étant nichée dans la facette *Tâches*) montrent un gain prévisible de fidélité appréciable (de 0,73 à 0,76) si l'on recourait à trois tâches, mais inférieur au gain que l'on pourrait espérer en faisant corriger les copies par trois correcteurs.

Tableau 7
Études de décision en faisant varier le nombre de tâches et de correcteurs (1er jeu)

	Plan original de l'expérimentation		*2 correcteurs et 2 tâches*		*2 correcteurs et 3 tâches*		*3 correcteurs et 2 tâches*	
	Niv.	*Univ.*	*Niv.*	*Univ.*	*Niv.*	*Univ.*	*Niv.*	*Univ.*
C (Candidats)	33	INF	33	INF	33	INF	33	INF
E (Correcteurs)	4	INF	2	INF	2	INF	3	INF
T (Tâches)	2	INF	2	INF	3	INF	2	INF
A:T (Critères)	3	3	3	3	3	3	3	3
Coef_G abs.	0,80777		0,73118		0,76064		0,78052	
Erreur type absolue	1,10028		1,36759		1,26525		1,19604	

Cela justifie le fait de se préoccuper en priorité de la problématique des correcteurs. Toutefois, une fois les correcteurs appariés en jurys équilibrés, l'ajout d'une tâche supplémentaire devient un des meilleurs moyens d'améliorer la fidélité. Ainsi, lorsqu'on considère la facette *Jurys* et non la facette *Correcteurs* et qu'on mène des études de décision en faisant varier le nombre de tâches, on obtient, pour l'un des deux jeux, une meilleure estimation de la fidélité lorsqu'on rajoute une troisième tâche plutôt qu'un second jury.

La sévérité des correcteurs

Nous avons vu que la constitution de jurys équilibrés permettait de renforcer la fidélité du dispositif d'évaluation. Il est pour cela nécessaire de mettre en place un suivi de la sévérité relative des correcteurs. Cependant, si un tel suivi peut être envisagé aisément lorsque les sessions sont organisées à date fixe et avec le même sujet d'épreuve, en faisant corriger par l'ensemble des correcteurs d'un même groupe un premier échantillon de copies, avant de constituer des paires de correcteurs, il est plus complexe à organiser quand les sessions sont à dates multiples et utilisent des jeux d'épreuves différents.

Nous avons vu que l'application de la théorie de la généralisabilité permettait de mettre en évidence des écarts de sévérité (présence d'un effet *Correcteurs*) ou éventuellement de difficulté entre jeux d'épreuves. Un logiciel comme *EduG* permet également, à travers le calcul de moyennes assorties d'écarts types sur la base de plans équilibrés (souvent difficiles à constituer), d'identifier quels sont les correcteurs les plus sévères ou les plus indulgents, et quels sont les jeux pour lesquels les candidats ont tendance à obtenir de meilleurs résultats.

L'utilisation du modèle de Rasch multifacettes à des fins directes d'évaluation dans le cas du TEF, pour lequel de nombreux sujets d'épreuves sont utilisés en parallèle au gré des sessions organisées à la demande par les centres agréés, et qui doit permettre la délivrance rapide de résultats, serait également problématique en production. Elle nécessiterait en effet la constitution d'une banque de données suffisamment interreliée pour permettre la prise en compte des variations dues aux différentes facettes d'instrumentation et à leurs interactions dans l'expression de scores ajustés, ce qui relève de la gageure. Elle nécessiterait par ailleurs de revoir la grille d'évaluation afin que les échelles de notations soit limitées (actuellement, chaque critère est évalué sur une échelle de 0 à 20) et que les critères d'évaluation soient d'importance comparable (abandon du système de pondération).

L'utilisation du modèle multifacettes de Rasch à partir d'échantillons partiellement équilibrés présente néanmoins un intérêt particulier, dans le cadre du Test d'évaluation de français, pour contrôler les différences de sévérité entre correcteurs (ainsi que la consistance de leurs évaluations), et en tenir compte dans la constitution des jurys.

Les données recueillies dans le cadre de l'expérimentation permettent de prendre également en considération les facettes *Jeux d'épreuves* (ou *Sujets*) et *Critères*, ainsi que des interactions entre facettes d'instrumentation, dans la mise en œuvre de modèles Rasch multifacettes. Cependant, le nombre de paramètres à évaluer évolue avec le nombre de facettes et d'interactions prises en considération par le modèle et plus le modèle est complexe, plus il est nécessaire de disposer d'un nombre important d'observations pour obtenir des estimations précises. Les résultats de l'analyse de généralisabilité, en informant sur l'importance relative des différentes sources d'erreur, permettent d'identifier des plans d'analyse pertinents et limités à la prise en considération des facettes à l'origine des principales variations des scores.

Identification de plans d'analyse pertinents pour le modèle multifacettes de Rasch

L'étude de généralisabilité menée pour les facettes *Candidats*, *Jeux* et *Correcteurs* a montré que les principales sources de variance d'erreur absolue correspondaient à un effet *Correcteurs* et à un effet d'interaction entre *Candidats* et *Jeux*, et qu'un léger effet *Jeux* était présent. Ces trois effets expliquent à eux seuls 86 % de la variance d'erreur absolue. L'effet d'interaction entre *Correcteurs* et *Jeux* n'explique pour sa part qu'une part très faible de cette variance (1,5 %). Ces résultats justifient l'adoption d'un plan d'analyse simplifié pour le modèle multifacettes de Rasch, qui fait abstraction des interactions entre les facettes *Correcteurs* et *Jeux*, tout en restant attentif à l'ajustement du modèle.

Résultats de l'application du modèle multifacettes de Rasch

Les résultats permettent de confirmer des différences de sévérité entre correcteurs. L'amplitude de la différence de sévérité entre les correcteurs, exprimée sur l'échelle de Rasch, est de 3,249, alors que l'écart type de l'estimation de la capacité des candidats est de 5,56. Compte tenu de la valeur du test du χ^2 (115,05 pour trois degrés de liberté, avec un niveau de signification < 0,0001), l'hypothèse nulle selon laquelle ces différences seraient dues à un biais d'échantillonnage peut être rejetée, et la valeur de l'indice de séparabilité qui exprime, sur une échelle de 0 à 1, la fidélité avec laquelle l'échantillon de données permet de différencier les correcteurs selon leur sévérité, est élevée (0,983). Ces résultats sont cohérents avec ceux de l'étude de généralisabilité. Ils permettent, par ailleurs, de situer les correcteurs les uns par rapport aux autres sur un continuum de sévérité.

Les résultats montrent également une légère différence de difficulté entre les deux jeux d'épreuve. L'amplitude de cette différence, exprimée sur l'échelle de Rasch (0,632), est toutefois faible comparée à l'écart type de l'estimation de la compétence des candidats, qui est de 5,56.

Les indices d'ajustement des correcteurs sont satisfaisants, ce qui témoigne d'une bonne consistance des résultats délivrés par les différents correcteurs. On peut s'étonner que les indices d'ajustement soient également satisfaisants en ce qui concerne les jeux d'épreuve, puisque les études de généralisabilité avaient mis en évidence un effet important d'interaction entre les facettes *Candidats* et *Jeux*. Cela est probablement dû à la réduction d'information nécessaire à la mise en œuvre des analyses Rasch multifacettes (menée à partir

des niveaux et non des scores), les variations de niveau étant sensiblement moins fréquentes que les variations de scores. En effet, dans 71 % des cas de l'expérimentation, les candidats se sont vu délivrer le même niveau TEF pour chacune des deux passations.

L'ajustement des données au modèle est d'ailleurs nettement moins bon lorsque l'analyse descend au niveau des critères communicatifs de chacune des tâches, mettant en évidence les interactions fortes entre *Candidats* et *Sujets* au niveau des évaluations par critère. Cela montre qu'il convient de faire un usage raisonné du modèle multifacettes de Rasch dans le cadre de l'épreuve d'expression écrite du TEF. La CCIP utilise ce modèle pour une estimation globale de la sévérité des correcteurs et, plus épisodiquement, pour analyser en détail les profils des correcteurs. C'est à partir de ce suivi de la sévérité des correcteurs et de l'information issue de l'arbitrage que la CCIP procède à la constitution de jurys équilibrés de correcteurs pour l'épreuve d'expression écrite du TEF.

Conclusion

L'évaluation des compétences en expression écrite est un système complexe faisant intervenir diverses facettes. Dès lors, l'analyse de la fidélité des épreuves doit tenir compte des contributions à l'erreur de mesure de ces différentes facettes, mais aussi de leurs interactions.

La théorie de la généralisabilité fournit un cadre théorique adapté à de telles analyses. Son application à l'épreuve d'expression écrite du TEF a toutefois nécessité la mise en œuvre d'expérimentations *ad hoc,* de manière à collecter des données selon un plan équilibré. Elle a permis de mettre en évidence, en plus d'un effet *Correcteurs* indéniable, un effet d'interaction important entre *Candidats* et *Jeux d'épreuves* qui rappelle l'intérêt de collecter des productions écrites à partir de plusieurs tâches et de thématiques variées. Il s'agit alors de trouver un juste équilibre entre les exigences en matière de fidélité, de validité et de faisabilité.

L'importance de l'effet *Correcteurs* traduit des écarts de sévérité. La faiblesse des effets d'interaction entre, d'une part, *Correcteurs* et *Candidats*, et, d'autre part, entre *Correcteurs* et *Sujets* ou *Jeux*, atteste la présence régulière de tels écarts. L'analyse Rasch multifacettes a confirmé ce constat et a permis de situer les différents correcteurs sur une échelle de sévérité. Cette différence de sévérité peut toutefois être prise en considération pour mettre en œuvre

une stratégie de constitution de jurys de correcteurs équilibrés, qui permet d'améliorer sensiblement la fidélité d'un dispositif d'évaluation prévoyant des corrections multiples.

Certains auteurs préconisent aussi l'utilisation du modèle multifacettes de Rasch pour délivrer aux candidats un score « ajusté » établi sur la base de leur score Rasch, qui tient compte de la sévérité des correcteurs et, éventuellement, de la difficulté des jeux d'épreuve. Un tel usage nécessite toutefois que l'outil d'évaluation et son contexte d'utilisation se prêtent à la mise en œuvre de ce modèle et que les conditions d'application du modèle de Rasch soient clairement réunies[13], notamment l'ajustement des données au modèle, qui suppose un effet d'interaction limité entre candidats et jeux d'épreuve. Ces deux exigences sont malheureusement rarement rencontrées lors de l'administration de tests au cours de sessions multiples (ce qui est naturellement différent dans le cas d'examens réunissant de nombreux candidats lors d'une même session et qui sont soumis à un même jeu d'épreuve).

NOTES

1. Des systèmes de correction automatisés font cependant leur apparition (Laurier & Diarra, 2009).
2. Ce cas est susceptible de se produire lorsqu'une grille analytique, constituée de critères évaluant séparément différents aspects de la production écrite, est utilisée pour évaluer les candidats.
3. À savoir, la capacité d'un correcteur à attribuer des scores semblables à deux moments différents pour une même série de copies ou à attribuer des scores proches à des copies réputées de niveau équivalent.
4. Pour McNamara (1996, p. 27), il est naturel d'observer une diversité dans les jugements, qui renvoient à des expériences de lecture individuelles.
5. La consistance des évaluations d'un correcteur reflète la régularité de ses caractéristiques d'évaluation (sévérité, manière d'exploiter l'échelle de notation, etc.). Un correcteur qui serait tantôt sévère, tantôt indulgent pourra avoir une estimation de sévérité proche de 0, mais sera probablement caractérisé par un indice d'ajustement peu satisfaisant, qui reflètera sa versatilité. L'étude de Weigle (1994) montre que, si la formation des correcteurs ne permet pas de supprimer les différences de sévérité, elle a un impact généralement positif sur la consistance des évaluations.
6. L'arbitre joue également un rôle dans le pilotage de la qualité en transmettant, au responsable pédagogique du TEF en charge du suivi des correcteurs, les copies pour lesquelles des écarts de plus d'un niveau entre correcteurs sont constatés sur un ou plusieurs critères, ce qui permet d'alimenter un tableau de suivi des correcteurs.
7. La version française du logiciel EduG 6.0 peut être téléchargée gratuitement à l'adresse [http://www.irdp.ch/edumetrie/logicielfrancais.htm].
8. Calculé en considérant, pour chaque candidat, la moyenne des scores des deux jurys.
9. Parallèlement aux épreuves d'expression écrite, les candidats de l'expérimentation ont été soumis aux épreuves de réception du TEF, sous la forme d'un questionnaire à choix multiple, qui montrent que si leurs compétences linguistiques s'échelonnent du niveau 2 au niveau 5 (ce qui correspond aux niveaux A2 à C1 du Cadre européen commun de référence pour les langues – CECR), la plupart des candidats sont de niveau 3 ou 4.
10. La corrélation interjeux, établie à partir de la moyenne des scores délivrés par les quatre correcteurs, est de 0,81 ($p < 0,0001$) sous condition d'égalité de moyenne et de variance.
11. L'amplitude des niveaux extrêmes est moindre (34 points), mais il n'y a pas de risque de voir le score vrai d'un candidat correspondre à un niveau inférieur au niveau minimum ou supérieur au niveau maximum.
12. Mais s'agissant de deux passations successives, cet effet inclut les modifications de stratégie des candidats entre les deux épreuves. Il est donc probablement surestimé.
13. Lorsque des correcteurs évaluent, pour un même candidat, une même copie, nous ne sommes pas face à des évaluations réellement indépendantes, alors qu'il s'agit d'une des conditions d'application du modèle de Rasch. Selon Linacre (1997), ce défaut d'indépendance locale conduit à une surestimation de la précision de la mesure, mais ne remet pas en cause la nécessité d'un ajustement tenant compte des différences de sévérité, et l'expression à partir du score Rasch des candidats, d'un score « ajusté ».

RÉFÉRENCES

Artus, F., & Demeuse, M. (2008). Évaluer les productions orales en français langue étrangère (FLE) en situation de test. Étude de la fidélité inter-juges de l'épreuve d'expression orale du Test d'évaluation de français (TEF) de la Chambre de commerce et d'industrie de Paris. *Les cahiers des sciences de l'éducation, 25 et 26,* 131-151.

Bachman L. F. (2004). *Statistical analyses for language assessment*. Cambridge: CUP.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12,* 238-257.

Bertrand, R., & Blais, J. G. (2004). *Modèles de mesure. L'apport de la théorie des réponses aux items*. Sainte-Foy (Canada) : Presses de l'Université du Québec.

Brennan, R. (2001). *Generalizability theory*. New York: Springer-Verlag.

Cardinet, J., & Tourneur, Y. (1985). *Assurer la mesure*. Berne: Peter Lang.

Casanova, D., Crendal, A., Demeuse, M., Desroches, F., & Holle, A. (2010, janvier). Validation empirique d'un test de français langue étrangère en regard du Cadre européen commun de référence pour les langues. *Actes du 22e colloque international de l'Association pour le développement des méthodologies d'évaluation en éducation (ADMEE-Europe)*, Braga, Portugal.

Chambre de commerce et d'industrie de Paris (2010). *TEF, le Test d'évaluation de français de la Chambre de commerce et d'industrie de Paris*. Paris : CCIP.

Conseil de l'Europe (2005). *Cadre européen commun de référence pour les langues*. Paris : Didier.

Cronbach, L. J., & Shavelson, R. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*, 391-418.

Demeuse, M., Desroches, F., Crendal, A., Renaud, F., Oster, P., & Leroux X. (2004, novembre). L'évaluation des compétences linguistiques des adultes en français langue étrangère dans une perspective de multiréférentialisation. *Actes du 17e colloque international de l'Association pour le développement des méthodologies d'évaluation en éducation (ADMEE-Europe)*. Lisbonne, Portugal.

Demeuse, M., Desroches, F., Crendal, A., Renaud, F., & Casanova, D. (2005, octobre). La fiabilité de l'évaluation des compétences linguistiques pour des adultes non francophones : présentation d'un protocole d'évaluation. *Actes du 18e colloque international de l'Association pour le développement des méthodologies d'évaluation en éducation (ADMEE-Europe)*. Reims, France.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2,* 197–221.

Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (éd.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (Section H)*. Strasbourg, France: Conseil de l'Europe/Division des politiques linguistiques.

Engelhard, G. Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*(2), 93–112.

Holle, A. (*à paraître*). Garantir la qualité d'un outil d'évaluation à forts enjeux : le cas du TEF. In O. Soutet, G. Maratier-Declety & M. Demeuse (dir.), *Assurer la qualité des épreuves d'évaluation en langues. Quels enjeux ?* Paris : Champion.

IRDP (2010). *Guide pour EduG*. Neuchatel (Suisse) : IRDP.

Laurier, M. D., & Diarra L. (2009). L'apport des technologies dans l'évaluation de la compétence à écrire. In J. G. Blais (dir.), *Évaluation des apprentissages et technologies de l'information et de la communication* (p. 77-104). Laval : PUL.

Linacre, J. M. (1989). *Many-facet Rasch Measurement*. Chicago: MESA Press.

Linacre, J. M. (1997) Investigating Judge Local Independence. *Rasch Measurement Transactions, 11*(1), 546-547.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training*, Language Testing, 12*, 54–71.

McNamara, T.F (1996). *Measuring second language performance*. New York: Longman.

McNamara, T. F., & Adams, R. J. (1991). *Exploring rater behavior with Rasch techniques*. Communication présentée au 13th Language Testing Research Colloquium, Educational Testing Service. Princeton, N.J.

Ministère de la Citoyenneté et de l'Immigration Canada (2006). *Niveaux de compétence linguistique canadiens*, Ottawa : Ministère de la Citoyenneté et Immigration Canada.

Noël-Jothy, F., & Sampsonis, B. (2006). *Certifications et outils d'évaluation en FLE.* Paris: Hachette.

Weigle, S. C. (1994). Effect of training on raters of ESL compositions. *Language Testing 11*, 197-223.

Weigle, S. C. (1998). Using Facets to model rater training effects. *Language Testing, 15*(2), 263-287.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: CUP.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). ACER Conquest Version 2.0. *Generalised Item Response Modelling Software*. ACER Press.

Date de réception : 26 juillet 2010

Date de réception de la version finale : 15 juillet 2011

Date d'acceptation : 20 juillet 2011