Mesure et évaluation en éducation

Une comparaison de l'étendue intra- et interindividuelle du niveau de sévérité d'examinateurs en français langue étrangère

Christophe Chénier

Volume 44, numéro 3, 2021

L'évaluation des compétences langagières : enjeux et perspectives

Réception : 30 novembre 2021
Version finale : 02 mai 2022
Acceptation : 03 mai 2022

URI : https://id.erudit.org/iderudit/1093066ar
DOI : https://doi.org/10.7202/1093066ar

Aller au sommaire du numéro

Éditeur(s)

ADMEE-Canada

ISSN

0823-3993 (imprimé)
2368-2000 (numérique)

Découvrir la revue

Citer cet article

Chénier, C. (2021). Une comparaison de l'étendue intra- et interindividuelle du niveau de sévérité d'examinateurs en français langue étrangère. *Mesure et évaluation en éducation*, *44*(3), 59–85. https://doi.org/10.7202/1093066ar

Résumé de l'article

De nombreuses recherches ont tenté de quantifier les écarts entre les niveaux de sévérité de différents examinateurs travaillant pour les mêmes évaluations. Leurs résultats montrent que les écarts interindividuels de niveaux de sévérité sont souvent importants, peu importe le contexte évaluatif. Toutefois, peu de recherches ont modélisé l'évolution temporelle intra-individuelle du niveau de sévérité et encore moins ont comparé, sur une période donnée, le rapport entre les étendues intra-individuelles et interindividuelles des niveaux de sévérité. Cette étude vise à combler ce manque en comparant les rapports entre les écarts intra- et interindividuels de six examinateurs ayant travaillé de septembre 2011 à avril 2014 pour l'épreuve d'expression orale du Test d'évaluation du français adapté au Québec (TEFAQ). Ces six examinateurs ont évalué la performance de 4083 candidats au test et leur niveau de sévérité a été estimé à l'aide du modèle de Rasch à multifacettes. Cinq dyades d'examinateurs ont été suivies durant cinq périodes distinctes, totalisant de 11 à 38 temps de mesure. Le niveau de sévérité a été estimé d'une à quatre fois par mois, ce qui a permis de calculer, pour chaque période, une étendue intra-individuelle du niveau de sévérité ainsi qu'une étendue interindividuelle. Ces étendues ont ensuite été mises en rapport, pour obtenir un ratio permettant de voir si le niveau de sévérité fluctue autant d'un examinateur à lui-même et d'un examinateur à l'autre. Les résultats montrent que, globalement, les écarts intra-individuels sont aussi élevés que les écarts interindividuels (rapport médian de 0,97), et ce, malgré le faible nombre d'examinateurs impliqués dans les modélisations. Finalement, les considérations pratiques, les limites méthodologiques et conceptuelles de l'étude sont discutées.

Une comparaison de l'étendue intra- et interindividuelle du niveau de sévérité d'examinateurs en français langue étrangère

Christophe Chénier
Université de Montréal

Mots-clés : sévérité des examinateurs, dérive temporelle de la sévérité, français langue étrangère (L2), effets de l'examinateur

De nombreuses recherches ont tenté de quantifier les écarts entre les niveaux de sévérité de différents examinateurs travaillant pour les mêmes évaluations. Leurs résultats montrent que les écarts interindividuels de niveaux de sévérité sont souvent importants, peu importe le contexte évaluatif. Toutefois, peu de recherches ont modélisé l'évolution temporelle intra-individuelle du niveau de sévérité et encore moins ont comparé, sur une période donnée, le rapport entre les étendues intra-individuelles et interindividuelles des niveaux de sévérité. Cette étude vise à combler ce manque en comparant les rapports entre les écarts intra- et interindividuels de six examinateurs ayant travaillé de septembre 2011 à avril 2014 pour l'épreuve d'expression orale du Test d'évaluation du français adapté au Québec (TEFAQ). Ces six examinateurs ont évalué la performance de 4083 candidats au test et leur niveau de sévérité a été estimé à l'aide du modèle de Rasch à multifacettes. Cinq dyades d'examinateurs ont été suivies durant cinq périodes distinctes, totalisant de 11 à 38 temps de mesure. Le niveau de sévérité a été estimé d'une à quatre fois par mois, ce qui a permis de calculer, pour chaque période, une étendue intra-individuelle du niveau de sévérité ainsi qu'une étendue interindividuelle. Ces étendues ont ensuite été mises en rapport, pour obtenir un ratio permettant de voir si le niveau de sévérité fluctue autant d'un examinateur à lui-même et d'un examinateur à l'autre. Les résultats montrent que, globalement, les écarts intra-individuels sont aussi élevés que les écarts interindividuels (rapport médian de 0,97), et ce, malgré le faible nombre d'examinateurs impliqués dans les modélisations. Finalement, les considérations pratiques, les limites méthodologiques et conceptuelles de l'étude sont discutées.

Key words: rater severity, severity time drift, French as a second language (L2), rater effects

Several studies have tried to quantify the differences in severity levels between raters working for the same assessments. Their results show that interindividual differences in severity levels are often important, regardless of the assessment situations. However, few studies have modeled the longitudinal evolution of intra-individual severity levels, and even fewer have compared the ratio between the intra- and interindividual differences. This paper seeks to remedy this lack of knowledge by comparing the ratio between the intra- and interindividual severity levels of six raters, who worked together, from September 2011 to April 2014, as raters for the oral expression test of the Test d'évaluation du français adapté au Québec *(TEFAQ). Those six raters assessed the performance of 4,083 candidates and their severity levels were estimated using the multi-facet Rasch model. Five raters dyads were modeled during five distinct periods, totaling from 11 to 38 time points, and their severity levels were estimated from once to four times per month. This allowed us to calculate, for each period, an intra-individual and interindividual severity range and these ranges were then compared to obtain a ratio showing whether a given rater's severity level fluctuates as much over time as it does when compared to the severity level of their peer. Results show that, overall, the intra-individual differences are as high as the interindividual ones, with a median ratio of 0.97, despite the small number of raters modeled. The practical impacts of those results are then discussed, as well as the methodological and conceptual limits of this study.*

Palavras-chave: severidade dos examinadores, desvio temporal na severidade, francês como língua estrangeira (L2), efeitos do examinador

Numerosas investigações tentaram quantificar as diferenças entre os níveis de severidade de diferentes examinadores que trabalharam as mesmas avaliações. Os resultados mostram que as diferenças interindividuais nos níveis de severidade são muitas vezes significativas, independentemente do contexto avaliativo. No entanto, são poucas as investigações que modelizaram a evolução temporal intraindividual do nível de severidade, e menos ainda que tenham comparado, ao longo de um determinado período, a relação entre as extensões intraindividual e interindividual dos níveis de severidade. Este estudo visa preencher esta lacuna comparando as razões entre as lacunas intra e interindividuais de seis examinadores que trabalharam de setembro de 2011 a abril de 2014 para a prova de expressão oral do Teste de avaliação do francês adaptado para o Québec (TEFAQ). Estes seis examinadores avaliaram o desempenho de 4.083 candidatos ao teste e o seu nível de severidade foi estimado usando o modelo Rasch multifacetado. Cinco duplas de examinadores foram monitorizadas durante cinco períodos distintos, totalizando de 11 a 38 momentos de aferição, e o grau de severidade foi estimado de uma a quatro vezes por mês, o que possibilitou calcular, para cada período, uma extensão intraindividual do nível de severidade, bem como uma extensão interindividual. Estes intervalos foram então comparados para obter um rácio que mostrasse se o nível de severidade flutua tanto de um examinador para consigo próprio quanto de um examinador para outro. Os resultados mostram que, no geral, as diferenças intraindividuais são tão altas quanto as diferenças interindividuais, com uma razão mediana de 0,97, apesar do pequeno número de examinadores envolvidos na modelização. Por fim, são discutidas considerações práticas, limites metodológicos e conceituais do estudo.

Problématique

La « révolution » communicationnelle, à laquelle a succédé l'approche actionnelle, a apporté de profonds changements aux manières d'évaluer dans le domaine des langues seconde ou étrangère (L2) (Spolsky, 2000 ; Uyaniker, 2017). L'accent y est mis sur l'atteinte de buts communicationnels ou sur l'accomplissement de tâches pour lesquelles la personne évaluée doit mobiliser sa compétence langagière (Shehadeh, 2012). Cela mène à inclure comme objet d'évaluation des performances complexes où l'on doit juger de l'atteinte de buts ou de la réussite d'actions, ainsi que la qualité avec laquelle ces buts sont atteints ou ces actions sont posées. Cette évaluation de performances complexes fait appel au jugement évaluatif et à la subjectivité qui en est constitutive. Cette question de la subjectivité est au cœur des recherches et des réflexions en évaluation, les plus anciennes sur le sujet datant de la fin du 19e siècle (Edgeworth, 1888, 1890). Divers auteurs ont analysé et conceptualisé les rapports entre subjectivité et évaluation, et ce, de multiples manières, mais tous s'entendent sur un principe fondamental pour l'exercice du jugement évaluatif : celui-ci ne doit pas être arbitraire (Cardinet, 1987 ; Gerard, 2002 ; Hadji, 1992 ; Linacre, 1994). Puisqu'évaluer, c'est poser un jugement en partie subjectif, il est donc normal, voire attendu, que les jugements exprimés diffèrent d'un examinateur[1] à l'autre, mais à l'intérieur des balises posées par le processus évaluatif (démarche, instruments etc.). Des différences de niveau de sévérité peuvent donc se manifester d'un examinateur à un autre, tant que ces différences ne sont pas arbitraires et capricieuses, et qu'elles sont justifiables aux yeux des parties prenantes concernées par l'évaluation. Cela suppose, entre autres, qu'un même examinateur ne voie pas son niveau de sévérité fluctuer de manière erratique au fil du temps. Si un même examinateur se retrouve à évaluer chaque semaine les performances langagières de candidats, à l'aide d'outils stables, on s'attend à ce que des performances similaires soient jugées de manière similaire, avec un niveau de sévérité relativement stable.

1. Traduction libre de *rater*, à la suite de Chénier (2018)

Cette préoccupation quant à la stabilité du niveau de sévérité des examinateurs est très présente dans le domaine de l'évaluation en L2 et elle se trouve à la fois dans les guides et dans les manuels d'organismes concernés (American Educational Research Association et al., 2014 ; Conseil de l'Europe, 2011 ; Council of Europe, 2009) et dans la littérature scientifique (Casanova & Demeuse, 2011 ; Eckes, 2012 ; Lamprianou et al., 2021 ; Wind & Engelhard Jr., 2016). D'un côté, les organismes responsables de l'évaluation en L2, garants de la validité de leurs processus évaluatifs, veulent que les examinateurs aient un niveau de sévérité stable et que ces niveaux soient similaires d'un examinateur à l'autre, de manière que le fait, pour un candidat, d'être évalué par une personne ou par une autre n'ait qu'un effet minime sur les résultats de l'évaluation. De l'autre côté, les études sur le sujet ont minimalement cherché à quantifier les écarts de niveaux de sévérité entre divers examinateurs appelés à évaluer les mêmes performances, de manière à décrire ces écarts, ce qui permet de voir jusqu'à quel point ces écarts posent un risque à la validité des évaluations. Si beaucoup de recherches ont décrit les écarts de niveau de sévérité en L2, elles se sont généralement cantonnées à les décrire de manière interindividuelle et synchronique, ou comportant un nombre restreint de temps de mesure, généralement moins de quatre (Chénier, 2018). Or, du point de vue de la validité des évaluations et de ses aspects éthiques, le problème est double : les organismes responsables des processus évaluatifs en L2 doivent s'assurer que les examinateurs travaillant pour une même épreuve ont des niveaux de sévérité similaires et ces organismes doivent veiller à ce que ces examinateurs aient, à travers le temps, un niveau de sévérité relativement stable et ne fluctuant pas de manière arbitraire. La question de la stabilité intra-individuelle du niveau de sévérité, appelée dérive temporelle du niveau de sévérité, a jusqu'ici été peu étudiée, presque toujours en anglais L2 ou L1, et plus souvent en production écrite qu'en production orale. À notre connaissance, seul Chénier (2018) a étudié les écarts intra- et interindividuels de niveau de sévérité en français L2 en production orale, et ce, d'une façon limitée étant donné l'envergure plus globale de cette recherche. La rareté des études sur ce sujet, malgré les aspects pratiques importants, justifie donc la question générale de recherche suivante : *quelle est l'amplitude des variations intra-individuelles et interindividuelles des niveaux de sévérité d'examinateurs en français L2* ?

Contexte théorique

La sévérité

Comme dans presque toutes les études sur la question, le niveau de sévérité d'un examinateur est ici défini de manière purement pragmatique, contextualisée et relative : un jugement évaluatif est considéré comme « sévère » ou « clément » si, en moyenne, il diffère du jugement d'autres examinateurs, tout en portant sur les mêmes performances (Myford & Wolfe, 2003). Cela suppose que la sévérité n'est pas une caractéristique intrinsèque à chacun mais bien qu'elle s'exprime contextuellement et différemment, un même examinateur pouvant être sévère pour un critère et clément pour un autre. Il faut donc un moyen d'estimer les différences de niveau de sévérité et ce moyen doit minimalement tenir compte du niveau d'habileté de la personne évaluée et du niveau de sévérité de l'examinateur. Le modèle de Rasch à multifacettes (MRM) a été développé précisément pour cela (Linacre, 1994). Il est devenu très utilisé dans le domaine de l'évaluation en L2 au cours des dernières années (Aryadoust et al., 2021) et presque toutes les études longitudinales du niveau de sévérité en L2 l'ont utilisé (Chénier, 2018 ; Congdon & McQueen, 2000 ; Davis, 2016 ; Kim, 2011 ; Lamprianou et al., 2021 ; Lim, 2009 ; Lumley & McNamara, 1995 ; Wolfe et al., 2007).

Le modèle de Rasch à multifacettes (MRM)

Le modèle de Rasch a d'abord été développé pour estimer la probabilité qu'une personne réponde correctement à un item, et ce, en fonction de deux paramètres : le niveau d'habileté du répondant et le niveau de difficulté de l'item. L'équation 1 présente ce modèle de base, sous sa forme logarithmique :

$$\ln\left[\frac{P_{nix=1}}{P_{nix=0}}\right] = \theta_n - \beta_i$$

où $P_{nix} = 1$ est la probabilité qu'un répondant n de niveau d'habileté θ ait la bonne réponse (1) à un item i d'un niveau de difficulté β (Bond & Fox, 2015). Les valeurs des paramètres sont en logits, soit le logarithme d'un rapport de cote tel qu'illustré par l'équation 1. Le modèle à multifacettes développe ce modèle de base en ajoutant tout autre paramètre pertinent, par exemple le niveau de sévérité des examinateurs. Ce modèle à multifacettes est normalement utilisé en L2 dans des contextes où la performance

est évaluée à l'aide d'une échelle descriptive à plusieurs échelons, éventuellement arrimée à un cadre de référence comme le *Cadre européen commun de référence* (CECR), où chaque échelon renvoie à un descripteur de performance. Les données sont donc polychotomiques et on peut modéliser ainsi la probabilité que le répondant n se voie octroyer l'échelon j plutôt que l'échelon $j - 1$ par l'examinateur e à l'item i, qui peut être un critère d'évaluation ou une tâche holistique, comme le montre l'équation 2 :

$$\ln\left[\frac{P_{neix=j}}{P_{neix=j-1}}\right] = \theta_n - \beta_i - \alpha_e - \tau_j$$

où α représente le niveau de sévérité de l'examinateur e et τ est un paramètre représentant les seuils séparant deux catégories adjacentes de l'échelle d'évaluation, entendu qu'une échelle à j catégories a $j - 1$ seuils au total. Précisons qu'il s'agit du modèle dit *rating scale* et que, pour celui-ci, l'échelle d'évaluation a un nombre fixe de catégories et tous les examinateurs sont réputés conceptualiser de la même manière l'ordre de ces catégories. Il existe une autre modélisation, le modèle à crédit partiel (*partial credit model*), où il peut y avoir une interaction entre les facettes du modèle. Cela permet, par exemple, d'estimer des niveaux de sévérité différents selon les échelons de l'échelle d'évaluation.

Le MRM a les mêmes conditions d'utilisation que le modèle de Rasch de base, soit l'unidimensionnalité, l'indépendance locale et la monotonicité de la relation entre le niveau d'habileté et la probabilité de se voir octroyer un échelon supérieur. Aryadoust et al. (2021) montrent toutefois que la majorité des études utilisant le MRM ne rapportent aucune procédure de vérification du respect de ces conditions d'utilisation. Tous les détails mathématiques et les aspects techniques sont expliqués dans Linacre (1994) et Eckes (2015).

Recension des écrits

Une revue de la littérature publiée en 2021 a recensé 215 études ayant utilisé le modèle de Rasch en L1 ou en L2, avec un nombre de textes publiés annuellement allant croissant depuis 2010 (Aryadoust et al., 2021). Chénier a, pour sa part, relevé 39 études en L1 ou en L2 ayant utilisé le MRM pour estimer le niveau de sévérité d'examinateurs. Le bilan de ces études est clair : les examinateurs travaillant pour les mêmes évaluations tendent à avoir des écarts de niveaux de sévérité interindividuels

importants (Chénier, 2018). Il faut garder à l'esprit que les mesures en logit de divers ensembles de données, et donc d'études différentes, ne peuvent être directement comparés, car les paramètres d'une analyse influencent l'étendue de l'échelle en logit. Il est tout de même intéressant de constater que l'écart interindividuel moyen entre les examinateurs les plus sévères et les moins sévères d'une même étude est de 2,60 logits, avec un écart-type de 1,40 logits, ce qui montre bien que des écarts de niveaux de sévérité sont communément observés (Chénier, 2018). Un écart aussi important représente généralement une différence substantielle dans la manière d'évaluer. Les études longitudinales, définies ici comme ayant au minimum trois temps de mesure distincts, sont plus rares, dix études ayant été identifiées. De ces dix, trois sont en L1 et sept en L2. Toutes sont en anglais, sauf l'étude de Chénier, qui porte sur une évaluation en français L2.

Congdon et McQueen (2000), Wolfe et al. (2007) et Leckie et Baird (2011) ont étudié des examinateurs d'évaluations de productions écrites en anglais L1. Congdon et McQueen (2000) avaient 16 examinateurs et six temps de mesure répartis sur huit jours. Leurs résultats détaillés montrent que l'écart interindividuel maximal observé à l'un des temps de mesure est de 3,30 logits, et l'écart intra-individuel maximal de 2,01 logits. L'étude de Wolfe et al. (2007), pour sa part, avait 101 examinateurs et huit temps de mesure étalés sur quatre jours. Leurs résultats ne présentent pas les résultats individuels détaillés mais des statistiques inférentielles développées par les chercheurs montrent que 33% des examinateurs avaient un écart intra-individuel de niveau de sévérité significativement différent. L'étude de Leckie et Baird (2011) comporte 689 examinateurs et cinq temps de mesure qui semblent répartis sur quelques jours. Ses résultats ne présentent aucun détail individuel mais les modèles multiniveaux log-linéaires utilisés s'ajustaient mieux aux données lorsque le paramètre de la sévérité des examinateurs variait en fonction du temps, ce qui montre qu'il y avait une certaine dérive temporelle du niveau de sévérité des examinateurs. Lim (2009 ; 2011) et Lamprianou et al. (2021) ont fait des études en production écrite en anglais L2. Lim (2009) a fait deux analyses : la première avec six examinateurs et sept temps de mesure sur 21 mois et la seconde avec sept examinateurs et cinq temps de mesure sur 15 mois. Les résultats de la première analyse montrent un écart interindividuel maximal de 2,62 logits et un écart intra-individuel maximal de 1,28 logits tandis que la seconde avait un écart interindividuel maximal de 0,93 logits et un écart intra-individuel maximal de 0,51 logits. L'étude de Lim (2011) avait 11 examinateurs suivis sur trois périodes, respectivement

de 12, 21 et 13 mois, mais ses résultats sont sous forme graphique, sans résultats individuels. Un examen des figures présentées montre toutefois que cinq des examinateurs ont vu leur niveau de sévérité dériver d'une manière appréciable. Lamprianou et al. (2021) ont suivi les examinateurs d'une évaluation d'admission à l'université durant 13 ans, à raison d'une passation par année. Les informations individuelles ne sont pas disponibles mais il semble y avoir eu une centaine d'examinateurs, dont 15 ayant participé à au moins six séances d'évaluation. Les résultats pour ces examinateurs ne sont présentés que graphiquement mais on peut voir que l'écart interindividuel maximal est d'environ 4 logits et l'écart intra-individuel maximal est d'environ 5 logits. Puisque les étendues des niveaux d'habileté en logit des candidats ne sont pas présentées, il est difficile de juger de l'importance de ces valeurs mais, considérant que leurs échelles d'évaluation ont de deux à cinq échelons, des écarts de 5 logits semblent très importants.

Les trois études suivantes ont eu lieu en production orale en anglais L2. Lumley et McNamara (1995) ont étudié 13 examinateurs d'une évaluation destinée aux professionnels de la santé. Il y a eu trois temps de mesure pendant 20 mois. Les écarts de niveau de sévérité interindividuels et intra-individuels les plus grands étaient respectivement de 2,43 et 1,76 logits. Kim (2011) a fait sa recherche avec neuf examinateurs, étudiés à trois reprises sur une période de trois mois. Les écarts interindividuels et intra-individuels maximaux étaient respectivement de 3,19 et 2,25 logits. De son côté, Davis (2016) a étudié l'évolution du niveau de sévérité de 20 examinateurs débutants sur une période de quatre semaines, avec quatre temps de mesure hebdomadaires. Les résultats montrent un écart interindividuel maximal de 1,60 logit et un écart intra-individuel maximal de 2,22 logits. Finalement, Chénier (2018), a étudié des examinateurs d'une évaluation de production orale en français L2. Il a suivi l'évolution des niveaux de sévérité de 10 examinateurs durant six périodes différentes, allant de 6 à 30 mois et ayant de 10 à 36 temps de mesure. Les écarts maximaux observés étaient de 1,90 logits pour l'intra-individuel et de 1,12 pour l'interindividuel.

La synthèse de la recension des écrits et les objectifs spécifiques de recherche

Au-delà des écarts en logits, non directement comparables d'une étude à l'autre, un constat s'impose : dans les études pour lesquelles les résultats individuels sont disponibles, on observe des écarts de niveau de sévérité

importants tant sur le plan intra- qu'interindividuel. Si nous prenons toutes les études pour lesquelles les résultats sont disponibles, la médiane des rapports intra- et interindividuels de niveaux de sévérité est de 0,78, ce qui signifie que, pour les examinateurs de ces études, l'écart maximal intra-individuel observé est égal à 78% de l'écart maximal interindividuel. Cela montre que le niveau de sévérité des examinateurs fluctue de manière importante intra-individuellement, et ce, malgré le fait que les études recensées aient chacune beaucoup d'examinateurs et peu de temps de mesure, à l'exception de l'étude de Chénier (2018). Qui plus est, rien dans ces études ne permet d'expliquer ces fluctuations.

Rappelons que ces fluctuations intra-individuelles importent, puisque les pratiques de formation initiale et continue des examinateurs supposent que le jugement évaluatif est à la fois malléable et durablement fixable. Il est malléable puisque l'on suppose que l'on peut, à l'aide de la formation ou de la modération sociale, amener l'examinateur à changer son jugement évaluatif afin de le rendre « acceptable ». Il est fixable puisque l'on souhaite que la formation ait un effet durable et que les changements du jugement évaluatif perdurent dans la pratique évaluative, au-delà des seules séances de formation. Il est donc important d'en savoir davantage sur l'amplitude des variations intra- et interindividuelles des niveaux de sévérité d'examinateurs, puisque peu de recherches ont étudié ce phénomène. Cette étude a donc pour objectifs spécifiques de recherche de :

1) Décrire les variations longitudinales intra- et interindividuelles des niveaux de sévérité des examinateurs ;
2) Comparer les écarts intra- et interindividuels de niveaux de sévérité des examinateurs.

Méthodologie

Les sources des données

Les données proviennent de six examinateurs (E1 à E6) ayant travaillé dans un centre homologué de passation du Test d'évaluation du français adapté au Québec (TEF/TEFaQ), un test de français L2 reconnu par les autorités québécoises et canadiennes à des fins d'immigration (CCI Paris-Île-de-France Education, 2021). Les données ont été collectées de septembre 2011 à avril 2014 dans ce centre situé à Montréal. Tous les examinateurs avaient de trois à dix années d'expérience comme enseignants de

français L2 au moment de commencer à travailler comme examinateurs. En septembre 2011, quatre de ceux-ci avaient de quelques mois à cinq années d'expérience comme examinateurs et les deux autres, E5 et E6, ont débuté durant la période de collecte. Ces six examinateurs ont évalué un total de 4083 candidats de niveaux de compétence variés, allant du débutant absolu au locuteur compétent (niveau pré A1 à C2 selon le CECR). Aucune information sociodémographique ou autre n'est disponible sur ces candidats.

Les instruments de collecte

Les données proviennent des notes accordées par les examinateurs aux épreuves d'expression orale du TEF/TEFAQ. L'épreuve d'expression orale prend la forme de deux jeux de rôle dans lesquels le candidat interagit avec deux examinateurs, chacun à tour de rôle. Les deux performances sont évaluées par chaque examinateur. Lors du premier jeu de rôle, d'une durée de cinq minutes, le candidat doit s'informer à propos d'un service ou d'un produit offert et, dans le second, de dix minutes, le candidat doit convaincre son interlocuteur d'adopter un comportement ou de faire une activité. À l'époque de la collecte des données, les performances étaient évaluées à l'aide d'une grille d'évaluation analytique comprenant 12 critères : six critères communicationnels – trois pour le premier jeu de rôle et trois pour le second – et six critères linguistiques, pour l'ensemble de la performance. Tous les critères sont associés à d'une échelle à 21 échelons arrimés aux six niveaux de compétence du CECR. Les deux examinateurs évaluent indépendamment les deux performances du candidat et, ensuite, ils se consultent pour en arriver à une notation consensuelle finale. Les données de cet article proviennent des évaluations indépendantes des examinateurs, c'est-à-dire des 12 notes attribuées à chaque candidat évalué.

Le déroulement et les considérations éthiques

Les données utilisées dans cet article sont des données secondaires originalement collectées à des fins de contrôle de la qualité par le centre de passation. Les autorités compétentes du centre ont transmis à l'auteur les données après les avoir anonymisées. Les autorités éthiques de l'université fréquentée par l'auteur à cette époque avaient jugé que la nature secondaire des données faisait en sorte qu'il n'avait pas à obtenir un certificat éthique pour son étude.

Les méthodes d'analyse

Les données disponibles ont d'abord été séparées en trois jeux distincts : un ensemble correspondant aux trois critères communicationnels utilisés pour noter le premier jeu de rôle (dorénavant « A »), un ensemble pour les trois critères communicationnels utilisés pour noter le second jeu de rôle (« B ») et un ensemble pour les six critères linguistiques utilisés pour noter l'ensemble de la performance du candidat (« L »). Les justifications détaillées se trouvent dans Casanova et Demeuse (2016) et Chénier (2018). Les données ont ensuite été examinées afin d'identifier des périodes distinctes durant lesquelles des paires d'examinateurs travaillaient conjointement. Cette vérification a révélé cinq périodes pour la prochaine phase analytique et le Tableau 1 montre les informations pertinentes sur ces périodes. La colonne *t* indique le nombre de temps de mesure.

Tableau 1
Informations sur les périodes et les examinateurs

Période	Durée	Fréquence	*t*	Examinateurs	*n* candidats
09-2011 à 07-2012	11 mois	1 par mois	11	E1 et E3	370
06-2012 à 03-2013	10 mois	2 par mois	19	E2 et E4	701
09-2012 à 09-2013	13 mois	3 par mois	38	E2 et E3	1380
12-2012 à 07-2013	8 mois	3 par mois	24	E3 et E5	899
11-2013 à 04-2014	6 mois	4 par mois	24	E5 et E6	733

Ensuite, chaque ensemble (A, B et L) a été modélisé avec le MRM à trois facettes : niveau d'habileté des candidats, niveau de sévérité des examinateurs et niveau de difficulté des critères d'évaluation. Cela a permis un examen rigoureux du respect des trois conditions d'utilisation, soit l'unidimensionnalité, la monotonicité de la relation entre le niveau d'habileté et la probabilité de se voir octroyer une note supérieure et l'indépendance locale. Ces vérifications n'ayant révélé aucune violation importante de ces conditions (voir Chénier, 2018, pour tous les détails), la prochaine étape a permis d'estimer les niveaux de sévérité de chaque examinateur à chacun des temps de mesure indiqués au Tableau 1. Pour ce faire, les résidus de l'étape précédente, soit la différence entre les valeurs attendues par le modèle et les valeurs réellement observées aux différents temps de mesure, ont été remodélisées avec un MRM à quatre facettes : niveau d'habileté des candidats, niveau de sévérité des examinateurs, niveau de difficulté des

critères d'évaluation et temps de mesure, cette dernière facette () étant une facette factice servant à estimer, pour chacun des temps de mesure, l'écart entre le niveau de sévérité global de l'examinateur et son niveau de sévérité au temps *t*. Cette facette est donc ancrée à une valeur de 0. L'équation 3 représente le modèle utilisé à cette étape :

$$\ln\left[\frac{P_{nei=j}}{P_{nei=j-1}}\right] = \hat{\theta}_n - \hat{b}_i - \alpha_{et} - \hat{\kappa}_t - \hat{\tau}_j$$

où est la difficulté du critère *i*, α_{et} est la sévérité d'un examinateur *e* au temps *t*, représente le temps *t* et la difficulté relative de chaque seuil de l'échelle d'évaluation. Les changements par rapport à l'équation 2 représentent le fait que l'estimation à cette étape se fait à partir des résidus des valeurs estimées à l'étape précédente, c'est-à-dire en fixant les valeurs des paramètres estimés lors de la première étape, d'où les accents circonflexes sur toutes les facettes sauf sur celle des examinateurs. Ces analyses sont possibles, car les données sont suffisamment liées par le travail en dyade des examinateurs, tous les examinateurs ayant travaillé ensemble à un moment ou un autre. Il s'agit donc d'un échantillonnage matriciel. Le lecteur intéressé est renvoyé à Chénier (2018), qui présente plus d'informations, et à Linacre (2021a, p. 610-611), qui explique en détails les aspects mathématiques. Il est à noter que, pour toutes les modélisations, le modèle MRM *rating scale* a été utilisé et l'estimation des paramètres a été faite par maximum de vraisemblance conjoint (Linacre, 1994), le tout avec le logiciel *Facets* version 3.83.6 (Linacre, 2021b).

Le résultat de cette dernière étape de modélisation est l'obtention, pour chaque temps de mesure, d'une valeur locale de sévérité, en logits, pour chaque examinateur. Cette valeur locale est ajoutée au niveau global de sévérité de chaque examinateur afin d'obtenir un niveau de sévérité en logit à chaque temps de mesure, c'est-à-dire des séries chronologiques de niveau de sévérité. Ces séries chronologiques ont ensuite été analysées à l'aide de graphiques chronologiques, de statistiques descriptives et d'un indice développé par Chénier (2018), le rapport intra-individuel/interindividuel des niveaux de sévérité (rapport intra/inter). Ce dernier se calcule de la sorte : pour une période donnée, on calcule l'écart maximal observé entre les différents niveaux de sévérité d'un même examinateur. Cela donne la valeur intra-individuelle. Pour la valeur interindividuelle, on calcule pour

chaque temps de mesure la différence entre les niveaux de sévérité des deux examinateurs et on prend la plus grande différence observée. On fait ensuite le rapport de ces deux valeurs, ce qui donne le rapport intra/inter. Une valeur inférieure à 1 montre que, pour cette période, l'écart interindividuel a été plus important que l'écart intra-individuel, alors qu'une valeur supérieure à 1 montre l'inverse, ce qui permet de comparer l'évolution temporelle intra- et interindividuelle du niveau de sévérité d'examinateurs travaillant ensemble. Ce rapport est sensible aux données extrêmes et doit être interprété avec prudence. Il est néanmoins utile puisque les données extrêmes représentent des écarts locaux importants du niveau de sévérité, écarts ayant un impact sur les résultats des candidats évalués lors de ces périodes. Finalement, les corrélations croisées entre les séries chronologiques d'examinateurs ont été estimées afin de voir s'il y avait des liens entre les niveaux de sévérité d'examinateurs travaillant de pair (Shin, 2017). Toutes les analyses statistiques de cette étape ont été faites avec le logiciel R, version 4.1.2 (R Core Team, 2021).

Résultats

Les résultats sont présentés pour chacune des cinq périodes, puis synthétisés. Remarquons d'abord que les valeurs en logit peuvent sembler très faibles, surtout par rapport aux valeurs recensées dans la littérature. Ceci s'explique par la grille d'évaluation utilisée par les examinateurs. Comme celle-ci compte 21 échelons, cela a un effet important de réduction sur l'étendue de l'échelle en logit obtenue pour les candidats et les examinateurs. Si une échelle à six échelons avait été utilisée, les valeurs observées auraient été environ trois fois plus grandes (Chénier, 2018). Deuxièmement, chaque paire de séries chronologiques a été centrée sur une moyenne globale de 0 afin d'en faciliter la visualisation : cela explique pourquoi les moyennes des séries de chaque paire sont l'opposée l'une de l'autre (p. ex. 0,04 et -0,04 logits dans la colonne *m*). Troisièmement, les droites (indigo et orange) dans les graphiques chronologiques représentent la tendance linéaire de chaque série. Finalement, à l'exception des rapports et des corrélations, tous les résultats sont en logits.

09-2011 à 07-2012

Les séries de cette période ont 11 temps de mesure mensuels. E1 et E3 ont évalué un total de 370 candidats, dont 215 conjointement (58%), pour des moyennes respectives de ~34 et ~20 candidats évalués par temps de mesure. La Figure 1 montre les trois paires (A, B et L) de séries chronologiques de E1 et E3 pour cette période et le Tableau 2, leurs statistiques descriptives. Notons que les ordonnées de toutes les figures du texte sont en logit. La courbe de E1 est en bleu foncé et celle de E3 en en rouge.

Puisqu'il n'y a ici que 11 temps de mesure, il faut être prudent dans l'interprétation des résultats. Nous observons toutefois que les séries sont stables, avec des tendances linéaires ayant une faible pente et des écarts limités entre les valeurs aux premiers et aux derniers temps de mesure. Les fluctuations sont limitées, avec des écarts-types assez faibles. Si les écarts sont globalement faibles, tant intra- qu'interindividuels, nous observons tout de même que E3 a des écarts intra-individuels presque aussi importants (« L ») et un peu plus importants (« B ») que les écarts interindividuels, comme en font foi ses rapports intra/inter près de ou supérieurs à 1. Pour finir, notons les corrélations croisées très élevées, tout en gardant à l'esprit que les coefficients de corrélations avec seulement 11 paires de données sont très instables.

06-2012 à 03-2013

Les séries de cette période ont 19 temps de mesure répartis sur 10 mois, à raison de deux temps de mesure par mois. E2 et E4 ont évalué 701 candidats, dont 123 conjointement (18%), en moyenne ~37 et ~6 par temps de mesure. La Figure 2 montre les séries chronologiques des niveaux de sévérité et le Tableau 3 leurs statistiques descriptives. La courbe de E2 est en bleu foncé et celle de E4 en rouge.

Il y a ici des différences importantes entre les séries des examinateurs. E4 a des niveaux de sévérité plus instables que E2, comme en font foi les étendues intra-individuelles et les rapports intra/inter. Il y a également quelques données extrêmes dans les séries de E4, aux temps 4, 13 et 19 de la série A et au temps 1 de la L. Ceci explique le fait que, pour les trois séries, E4 a un rapport intra/inter supérieur à 1, d'une valeur de presque 2 pour la série L. La dyade, pour sa part, a un comportement intéressant pour les séries A, où l'on remarque une tendance à la convergence des niveaux de sévérité, assez éloignés lors des cinq premiers temps et plus

Figure 1
Niveaux de sévérité des examinateurs E1 et E3 du 09-2011 au 07-2012

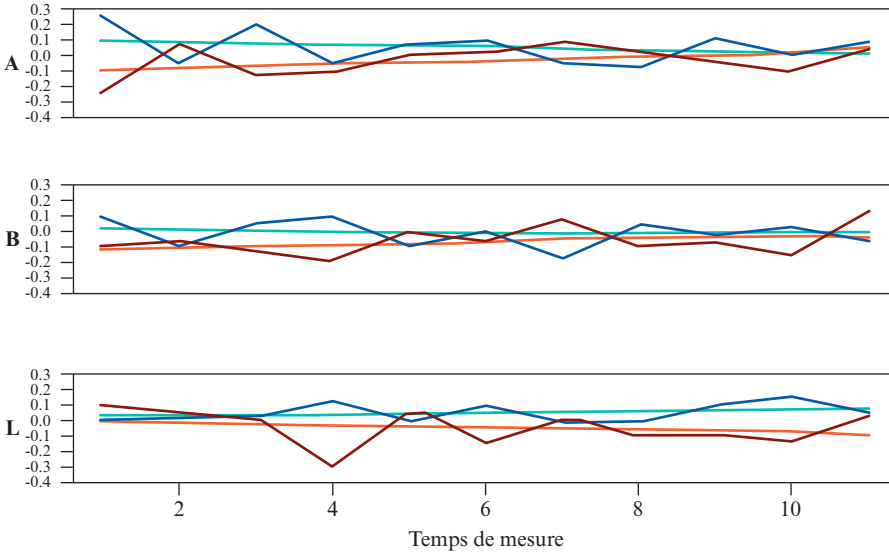


Tableau 2
Statistiques descriptives des séries du 09-2011 au 07-2012

	M	E-T	Étendue intra	Étendue inter	Intra/Inter	*r* croisée
E1 A	0,04	0,11	0,32		0,67	
E3 A	-0,04	0,10	0,32		0,67	
2 séries A				0,48		-0,63
E1 B	0,02	0,08	0,26		0,93	
E3 B	-0,02	0,10	0,32		1,14	
2 séries B				0,28		-0,78
E1 L	0,05	0,06	0,17		0,41	
E3 L	-0,05	0,11	0,38		0,93	
2 séries L				0,41		-0,74

près lors des cinq derniers, comme le montre l'écart entre les tendances linéaires, qui passe de 0,55 logit au temps 1 à 0,17 au temps 19. Terminons en relevant que des trois paires de séries, seule la série B est modérément négativement corrélée.

Figure 2
Niveaux de sévérité des examinateurs E2 et E4 du 06-2012 au 03-2013

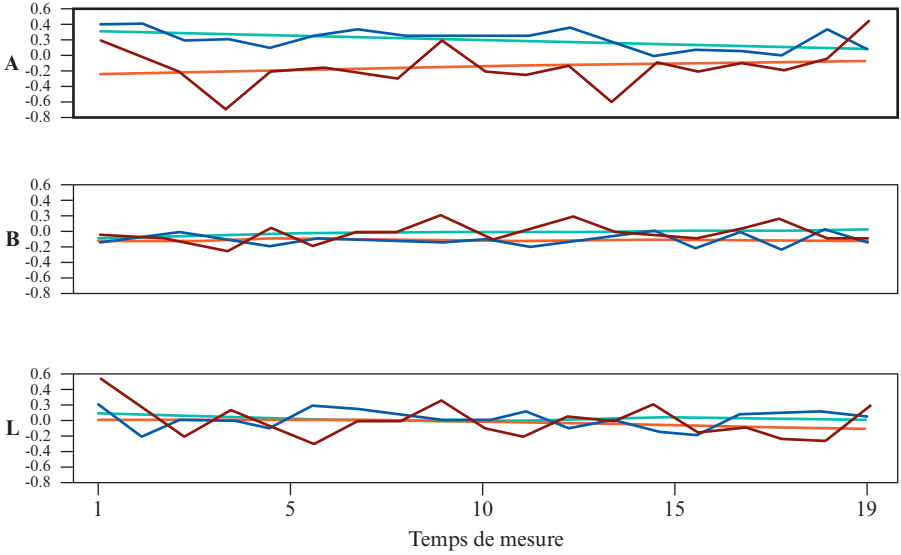


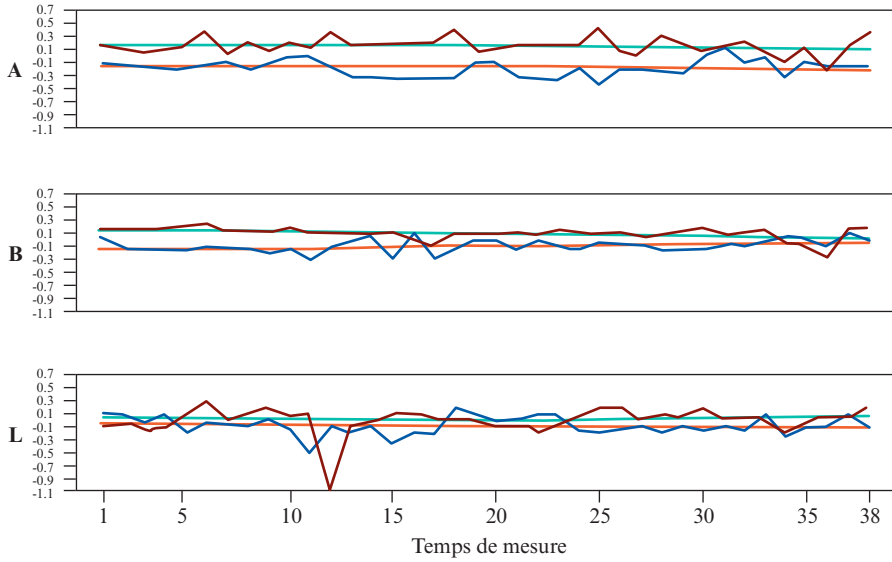
Tableau 3
Statistiques descriptives des séries du 06-2012 au 03-2013

	M	**E-T**	**Étendue intra**	**Étendue inter**	**Intra/Inter**	***r* croisée**
E2 A	0,18	0,13	0,44		0,48	
E4 A	-0,18	0,25	1,14		1,30	
2 séries A				0,91		-0,01
E2 B	-0,05	0,08	0,26		0,65	
E4 B	0,05	0,12	0,45		1,10	
2 séries B				0,40		-0,46
E2 L	0,02	0,12	0,43		0,93	
E4 L	-0,02	0,22	0,86		1,90	
2 séries L				0,46		-0,10

09-2012 à 09-2013

Les séries de cette période ont 38 temps de mesure, à raison de trois temps par mois durant 13 mois, et E2 et E3 ont évalué un total de 1380 candidats, dont 344 conjointement (~25%), en moyenne ~36 et ~9 par temps de mesure. La Figure 3 montre les trois paires de séries chronologiques de E2 et E3 pour cette période et le Tableau 4 leurs statistiques descriptives. La courbe de E2 est en bleu foncé et celle de E3 en rouge.

Figure 3
Niveaux de sévérité des examinateurs E2 et E3 du 09-2012 au 09-2013



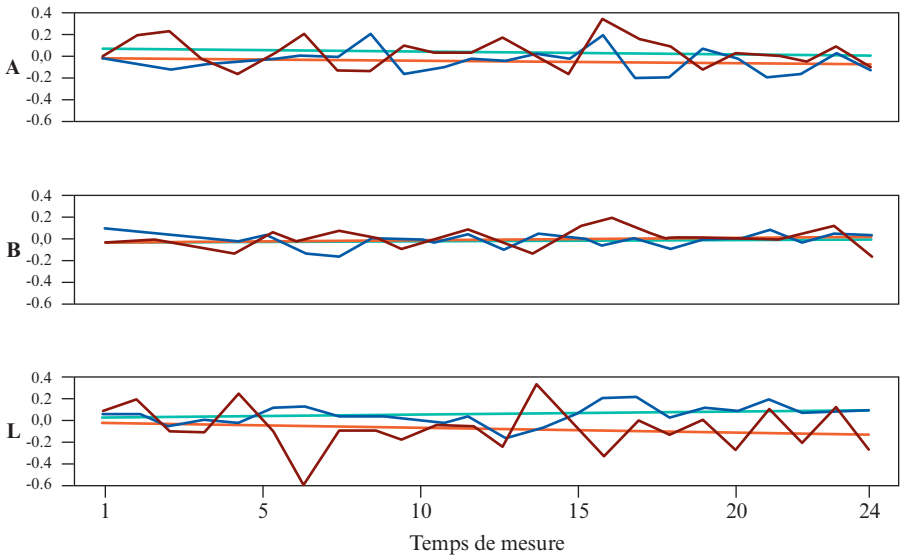
La valeur extrême de la série E3 L au temps 13 est possiblement attribuable au faible nombre de candidats évalués par E3 à ce moment-là, soit trois candidats. À l'exception de cette valeur extrême, les séries sont assez stables, comme le montrent les tendances linéaires de faible amplitude, sauf pour les séries B où l'écart interindividuel passe de 0,34 logit au temps 1 à 0,08 au temps 38. La relative stabilité des niveaux de sévérité s'observe également par le maintien de l'ordre de sévérité entre E2 et E3, E3 étant plus sévère 83% du temps à travers les trois séries. Finalement, relevons que les étendues intra-individuelles sont assez importantes, même si l'on exclut

la valeur de la série E3 L à cause de la donnée extrême, les autres valeurs allant de 0,65 à 1,22, ce qui montre des variations intra-individuelles substantielles pour E2 et E3 sur cette période. Rappelons néanmoins que cette période est d'une durée de 13 mois, ce qui est substantiel.

Tableau 4
Statistiques descriptives des séries du 09-2012 au 09-2013

	M	E-T	Étendue intra	Étendue inter	Intra/Inter	*r* croisée
E2 A	-0,18	0,14	0,60		0,67	
E3 A	0,18	0,14	0,77		0,86	
2 séries A				0,90		-0,31
E2 B	-0,10	0,11	0,45		0,98	
E3 B	0,10	0,11	0,56		1,22	
2 séries B				0,46		-0,09
E2 L	-0,05	0,15	0,75		0,65	
E3 L	0,05	0,24	1,52		1,30	
2 séries L				1,15		0,28

Figure 4
Niveaux de sévérité des examinateurs E3 et E5 du 12-2012 au 07-2013



12-2012 à 07-2013

Les séries de E3 et E5 ont 24 temps de mesure répartis sur huit mois, avec trois temps de mesure par mois. Durant cette période, E3 et E5 ont évalué 899 candidats, dont 144 conjointement (~16%), soit des moyennes par temps de mesure de ~37 et ~6 candidats. La Figure 4 montre les graphiques chronologiques des séries des niveaux de sévérité et le Tableau 5 contient leurs statistiques descriptives. La courbe de E3 est en bleu foncé et celle de E5 en rouge.

Tableau 5
Statistiques descriptives des séries du 12-2012 au 07-2013

	M	**E-T**	**Étendue intra**	**Étendue inter**	**Intra/Inter**	***r* croisée**
E3 A	-0,04	0,11	0,44		1,20	
E5 A	0,04	0,14	0,50		1,40	
2 séries A				0,37		-0,04
E3 B	-0,01	0,06	0,26		1,10	
E5 B	0,01	0,09	0,36		1,60	
2 séries B				0,23		-0,12
E3 L	0,06	0,09	0,40		0,56	
E5 L	-0,06	0,20	0,93		1,30	
2 séries L				0,72		-0,04

La valeur extrême observée au temps 7 de la série E5 L ne peut être expliquée par un faible nombre de candidats, E5 ayant évalué neuf candidats au temps 7. Les six séries de cette période sont stables, les tendances linéaires étant presque horizontales. Seules les séries L ont des tendances linéaires opposées, l'écart allant de 0,02 à 0,20 logit du premier au dernier temps de mesure. Les séries de E5 fluctuent systématiquement plus que celles de E3, comme en font foi les étendues intra-individuelles et les écarts-types plus élevés pour E5. Les rapports intra/inter sont particulièrement élevés pour cette période, tous, sauf un, étant supérieurs à 1. Il faut toutefois tenir compte du fait que ces rapports sont en partie attribuables au fait que les étendues interindividuelles sont assez faibles pour A et B, ce qui explique que leurs quatre rapports soient supérieurs à 1. Finalement, notons l'absence de corrélations entre les paires de séries de cette période.

11-2013 à 04-2014

Sur cette période, les séries de E5 et E6 ont 24 temps de mesure répartis sur six mois, à raison de quatre temps de mesure par mois. Les deux examinateurs ont évalué un total de 733 candidats, dont 203 conjointement (~28%), pour des moyennes respectives de ~31 et ~8 par temps de mesure. La Figure 5 montre les graphiques chronologiques des séries des niveaux de sévérité et le Tableau 6 contient leurs statistiques descriptives. La courbe de E5 est en bleu foncé et celle de E6 en rouge.

La série A de E6 a la particularité d'avoir la seule tendance linéaire locale franche, son niveau de sévérité allant constamment en diminuant du temps 1 au temps 9, pour ensuite remonter quasi systématiquement du temps 10 au temps 20. Il y a d'ailleurs convergence des tendances linéaires des niveaux de sévérité pour les séries A, l'écart initial étant de 0,25 logit, alors qu'il n'est plus que de 0,02 logit au dernier temps. Les tendances des deux autres paires sont quasi stables. Toutes les séries de cette période ont un patron en dents de scie, où les hausses et les baisses locales se succèdent, ce que reflètent les corrélations croisées observées, similaires pour les trois

Figure 5
Niveaux de sévérité des examinateurs E5 et E6 du 11-2013 au 04-2014

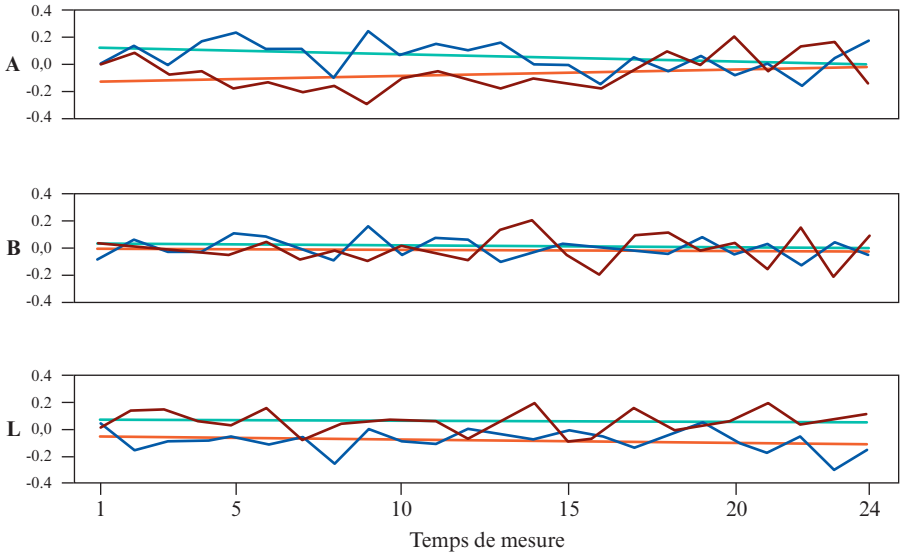


Tableau 6
Statistiques descriptives des séries du 11-2013 au 04-2014

	M	E-T	Étendue intra	Étendue inter	Intra/Inter	*r* croisée
E5 A	0,06	0,11	0,41		0,76	
E6 A	-0,06	0,13	0,52		0,96	
2 séries A				0,54		-0,46
E5 B	0,00	0,07	0,30		1,00	
E6 B	0,00	0,11	0,44		1,50	
2 séries B				0,29		-0,47
E5 L	-0,07	0,08	0,35		1,00	
E6 L	0,07	0,08	0,29		0,83	
2 séries L				0,35		-0,43

paires de séries et de valeurs négatives, montrant bien que lorsque le niveau de sévérité de E5 augmente, celui de E6 diminue, et vice-versa. Les rapports intra/inter sont assez élevés pour cette période, allant de 0,76 à 1,50. Notons toutefois le seul cas d'inversion à travers les cinq périodes modélisées dans cette étude : alors que E6 a un rapport intra/inter plus grand que E5 pour les séries A et B, c'est E5 qui a le plus grand pour les séries L. Finalement, alors que E5 est presque toujours plus sévère pour les séries A, c'est le contraire pour L, alors que les séries B se croisent constamment.

La synthèse des résultats

Un premier constat ressort d'un regard transversal sur les résultats de chacune des périodes : les niveaux de sévérité varient presque autant intra-individuellement qu'interindividuellement. Il y a pour chaque ensemble de données (A, B et L), un total de 10 rapports intra/inter et le Tableau 7 montre les valeurs minimales, médianes et maximales de ces 10 rapports, par ensemble de données.

Les médianes, proches de 1, montrent bien que, en moyenne, sur une période donnée, le niveau de sévérité d'un examinateur de cette étude varie autant intra- qu'interindividuellement. L'autre résultat important est que, à travers les trois séries de chacune des cinq périodes, c'est toujours le même examinateur qui a une étendue intra-individuelle plus grande que son collègue, à une exception près, soit la série L de la période du 11-2013 au 04-2014. Ainsi, dans l'ordre présenté précédemment, E3 a une plus grande étendue intra-individuelle que E1, E4 que E2, E3 que E2 et E5 que E3.

Tableau 7
Statistiques descriptives des rapports intralinter pour les cinq périodes étudiées

	Min	**Médiane**	**Max**
A	0,48	0,81	1,40
B	0,65	1,10	1,60
L	0,41	0,93	1,90
Toutes les séries	0,41	0,97	1,90

Conséquemment, encore à une exception près, pour chaque paire d'examinateurs, un examinateur a toujours une variance plus grande pour les séries chronologiques de ses niveaux de sévérité. Cela montre que certains examinateurs ont des niveaux de sévérité moins stables que ceux de leurs collègues.

Discussion

Les résultats de cette étude sont globalement en accord avec les résultats des études recensées. Alors que toutes ces études montrent que les niveaux de sévérité des examinateurs varient à la fois intra- et interindividuellement, l'amplitude de ces variations est variable. Il est toutefois intéressant de remarquer les grandes similitudes dans les valeurs des rapports intra/inter des étendues de niveaux de sévérité. Alors que la médiane, pour tous les rapports trouvés dans la littérature, est de 0,78, celle des rapports de cette recherche est de 0,97 (voir Tableau 7). Le résultat légèrement supérieur observé dans cette étude s'explique en partie par le devis utilisé : à l'exception de l'étude de Chénier (2018), les études recensées ont beaucoup d'examinateurs (*md* = 20) et peu de temps de mesure (*md* = 6,5). Or, le rapport intra/inter se calcule à l'aide des étendues intra- et interindividuelles. Plus il y a d'examinateurs dans une période, plus la probabilité augmente que l'étendue interindividuelle soit grande et, réciproquement, plus il y a de temps de mesure, plus la probabilité augmente que l'étendue intra-individuelle soit grande. Donc, le rapport intra/inter dépend en partie du devis retenu. Prenons deux cas extrêmes : il serait très étonnant qu'avec 6 000 examinateurs et deux temps de mesure, le rapport intra/inter ne soit pas minuscule. Inversement, avec deux examinateurs et 6 000 temps de mesure, le rapport serait probablement très élevé. Les cinq périodes de cette recherche ont, quant à elles, seulement deux examinateurs, mais de 11 à 38 temps de mesure. Il n'est donc pas étonnant que les rapports intra/inter observés soient plus élevés que ceux que l'on trouve dans la littérature. Ces

résultats constituent néanmoins un témoignage supplémentaire montrant que l'amplitude des variations intra-individuelles est substantielle lorsque ces variations sont étudiées sur une période suffisamment importante.

Les résultats sur les écarts-types et les étendues intra-individuelles des séries chronologiques des examinateurs tendent également à montrer que l'amplitude des variations intra-individuelles est non seulement substantielle, mais bien relativement stable. Pour chaque paire d'examinateurs étudiée, le même examinateur a toujours, pour ses séries chronologiques, des écarts-types et des écarts intra-individuels égaux ou supérieurs à ceux de son collègue, à une exception près. Cette constance suggère que les variations intra-individuelles des niveaux de sévérité sont liées à des manières d'évaluer, à des conceptions pour lesquelles les examinateurs diffèrent, constats que l'on trouve dans les résultats de plusieurs études sur les examinateurs en L2 (Barkaoui, 2010 ; Eckes, 2012 ; Park & Yan, 2019 ; Zhang, 2016). Finalement, les séries chronologiques des niveaux de sévérité des examinateurs de cette étude sont généralement stables, avec une tendance linéaire ayant une faible pente et une absence de cycles locaux importants. Cela semble en accord avec ce que l'on peut inférer des figures présentes dans les trois études recensées ayant le plus de temps de mesure (Lamprianou et al., 2021 ; Leckie & Baird, 2011 ; Lim, 2011).

Conclusion

Cette étude visait à décrire et à comparer les variations intra- et interindividuelles des séries chronologiques d'examinateurs en français L2. Les données, provenant de six examinateurs ayant évalué 4 083 candidats au TEFAQ/TEF, ont été analysées avec le MRM. Les résultats montrent que, en moyenne, les niveaux de sévérité des examinateurs varient presque autant intra- qu'interindividuellement. Cela a une importance pratique dont il faudrait tenir compte puisque les organismes responsables d'évaluations en L2 ne suivent pas nécessairement l'évolution longitudinale des niveaux de sévérité de leurs examinateurs. Plusieurs ne vérifient la qualité du jugement évaluatif des examinateurs qu'à leur embauche, puis périodiquement, mais à une fréquence qui peut être très faible. Si, sur une période de quelques mois, les niveaux de sévérité d'examinateurs fluctuent intra-individuellement autant que le laissent croire les résultats de cette étude, la qualité des évaluations peut être remise en question, surtout quand on sait combien les enjeux entourant ces évaluations peuvent être élevés : accès à des programmes d'études, immigration, accès à un ordre professionnel.

Cela dit, les résultats de cette étude doivent être interprétés avec une grande prudence. Il faut d'abord garder à l'esprit que les résultats du rapport intra/inter sont sensibles à la présence d'une seule valeur extrême, puisqu'un écart à un seul temps de mesure aura un impact important sur la valeur de cet indice. De plus, ce rapport est une statistique descriptive et, à ce titre, il résume une partie de l'information contenue dans la distribution des niveaux de sévérité, sans prétexte à l'exhaustivité. Une autre limite découle de la variante du modèle de mesure utilisé, le MRM *rating scale*. Une modélisation à crédit partiel montrerait d'éventuelles différences dans l'opérationnalisation de l'échelle d'évaluation et des critères d'évaluation par les examinateurs, ce qui changerait les estimations des niveaux de sévérité. Cette modélisation est toutefois impossible à utiliser avec notre plan d'échantillonnage longitudinal, car il aurait fallu pour chaque période au moins 10 notes par échelon de l'échelle d'évaluation, ce qui était impossible (Eckes, 2015). Finalement, au-delà des limites habituelles que l'on trouve aux études quantitatives idiographiques, une limite importante s'impose : les résultats de cette étude sont entièrement tributaires des modélisations temporelles retenues. Puisqu'il n'y a pas d'unités temporelles naturelles pour étudier l'évolution longitudinale des niveaux de sévérité – contrairement, par exemple, à la saisonnalité pour les phénomènes climatiques – il faut faire des choix arbitraires. Ainsi, les résultats de la période 11-2013 au 04-2014 dépendent de la fréquence retenue de quatre temps de mesure mensuels. Les mêmes données brutes, mais modélisées selon une autre périodicité, mèneraient potentiellement à d'autres résultats. Cela vaut évidemment pour toutes les études longitudinales recensées et c'est pourquoi, considérant la rareté des études longitudinales sur cette question, il faudrait davantage d'études, dans divers contextes évaluatifs, afin de voir jusqu'à quel point les résultats de cette recherche sont singuliers. Il serait particulièrement important que des études reprennent un plan d'échantillonnage similaire, mais avec un devis permettant l'utilisation du MRM à crédits partiels, ce qui permettrait de voir plus finement si les écarts locaux de sévérité sont attribuables à des changements du jugement évaluatif de l'examinateur ou à des différences de niveaux d'habileté chez les candidats.

Réception : 30 novembre 2021
Version finale : 02 mai 2022
Acceptation : 03 mai 2022

LISTE DE RÉFÉRENCES

American Educational Research Association, American Psychological Association et National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.

Aryadoust, V., Ying Ng, L., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment : Recommendations and guidelines for research. *Language Testing*, *38*(1), 6-40. https ://doi.org/10.1177%2F0265532220927487

Barkaoui, K. (2010). Do ESL Essay Raters' Evaluation Criteria Change With Experience ? A Mixed-Method, Cross-Sectional Study. *TESOL Quarterly*, *44*(1), 31-57. https ://doi.org/10.5054/tq.2010.214047

Bond, T. G., & Fox, C. M. (2015). Applying the Rasch Model : Fundamental Measurement in the Human Sciences (3e éd.). Routledge.

Cardinet, J. (1987). *L'objectivité de l'évaluation*. Recherches, Institut romand de recherches et de documentations pédagogiques.

Casanova, D., & Demeuse, M. (2011). Analyse des différentes facettes influant sur la fidélité de l'épreuve d'expression écrite d'un test de français langue étrangère. *Mesure et évaluation en éducation, 34*(1), 25-53. https ://doi.org/10.7202/1024862ar

Casanova, D., & Demeuse, M. (2016). Évaluateurs évalués : évaluation diagnostique des compétences en évaluation des correcteurs d'une épreuve d'expression écrite à forts enjeux. *Mesure et évaluation en éducation*, *39*(3), 59-96. https ://doi.org/10.7202/1040137ar

CCI Paris-Île-de-France Éducation. (2021). *Test d'évaluation du français*. https ://www.lefrancaisdesaffaires.fr/tests-diplomes/test-evaluation-francais-tef/

Chénier, C. (2018). Étude longitudinale du niveau de sévérité d'examinateurs d'un test d'expression orale en français langue étrangère [Thèse de doctorat non publiée]. Université du Québec à Montréal.

Congdon, P. J., & McQueen, J. (2000). The Stability of Rater Severity in Large-Scale Assessment Programs. *Journal of Educational Measurement*, *37*(2), 163-178. https ://www.jstor.org/stable/1435283

Conseil de l'Europe (2011). Manuel pour l'élaboration et la passation de tests et d'examens de langue. Association of language testers in Europe.

Council of Europe (2009). Relating Language Examinations to the Common European Framework of Reference for Languages : Learning, Teaching, Assessment (CEFR). Council of Europe Publishing.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, *33*(1), 117-135. https ://doi.org/10.1177%2F0265532215582282

Eckes, T. (2012). Operational Rater Types in Writing Assessment : Linking Rater Cognition to Rater Behavior. *Language Assessment Quarterly*, *9*(3), 270-292. https ://doi.org/10.1080/15434303.2011.649381

Eckes, T. (2015). Introduction to Many-Facet Rasch Measurement (2e éd.). Peter Lang.

Edgeworth, F. Y. (1888). The Statistics of Examinations. *Journal of the Royal Statistical Society*, *51*(3), 599-635. https ://www.jstor.org/stable/2339898

Edgeworth, F. Y. (1890). The Element of Chance in Competitive Examinations. *Journal of the Royal Statistical Society*, *53*(4), 644-663. https://www.jstor.org/stable/2979446

Gerard, F.-M. (2002). L'indispensable subjectivité de l'évaluation. *Antipodes*, 156, 26-34.

Hadji, C. (1992). *L'évaluation des actions éducatives*. Presses Universitaires de France.

Kim, H. J. (2011). Investigating raters' development of rating ability on a second language speaking assessment [Thèse de doctorat non publiée]. Teachers College, Columbia University.

Lamprianou, I., Tsagari, D., & Kyriakou, N. (2021). The longitudinal stability of rating characteristics in an EFL examination : Methodological and substantive considerations. *Language Testing*, *38*(2), 273-301. https://doi.org/10.1177%2F0265532220940960

Leckie, G., & Baird, J.-A. (2011). Rater Effects on Essay Scoring : A Multilevel Analysis of Severity Drift, Central Tendency, and Rater Experience. *Journal of Educational Measurement*, *48*(4), 399-418. https://doi.org/10.1111/j.1745-3984.2011.00152.x

Lim, G. (2009). Prompt and rater effects in second language writing and performance assessment. [Thèse doctorale non publiée]. Michigan State University.

Lim, G. (2011). The development and maintenance of rating quality in performance writing assessment : A longitudinal study of new and experienced raters. *Language testing*, *28*(4), 543-560. https://doi.org/10.1177%2F0265532211406422

Linacre, J. M. (1994). *Many-Facet Rasch Measurement* (2e éd.). Mesa.

Linacre, J. M. (2021a). Facets computer program for many-facet Rasch measurement Program Manual. Winsteps.com.

Linacre, J. M. (2021b). Facets computer program for many-facet Rasch measurement, version 3.83.6. Winsteps.com.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias : implications for training. *Language Testing*, *12*(1), 54-71. https://doi.org/10.1177%2F026553229501200104

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement : Part 1. *Journal of Applied Measurement*, *4*(4), 386-422.

Park, H., & Yan, X. (2019). An investigation into rater performance with a holistic scale and a binary, analytic scale on an ESL writing placement test. *Papers in Language Testing and Assessment*, *8*(2), 34-64.

R Core Team (2021). *R : A language and environment for statistical computing* (version 4.1.2). R Foundation for Statistical Computing.

Shehadeh, A. (2012). Task-Based Language Assessment : Components, Development, and Implementation. Dans C. Coombe, P. Davidson, B. O'Sullivan et S. Stoynoff (dir.), *The Cambridge Guide to Second Language Assessment* (p. 156-163), Cambridge University Press.

Shin, Y. (2017). *Time Series Analysis in the Social Sciences*. University of California Press.

Spolsky, B. (2000). Language Testing in *The Modern Language Journal*. *The Modern Language Journal*, *84*(4), 536-552. https://doi.org/10.1111/0026-7902.00086

Uyaniker, P. (2017). Language Assessment : Now and Then. *Avrasya Dil E-itimi ve Araştırmaları Dergisi*, *1*(1), 1- 20.

Wind, S. A., & Engelhard Jr., G. (2016). Exploring Rating Quality in Rater-Mediated Assessments Using Mokken Scale Analysis. *Educational and Psychological Measurement*, *76*(4), 685-706. https://doi.org/10.1177/0013164415604704

Wolfe, E. W., Myford, C. M., Engelhard Jr., G., & Manalo, J. R. (2007). *Monitoring ReaderPerformance and DRIFT in the AP® English Literature and Composition Examination Using Benchmark Essays*. (Rapport de recherche no 2007-2). College Board.

Zhang, J. (2016). Same text different processing? Exploring how raters' cognitive and meta-cognitive strategies influence rating accuracy in essay scoring. *Assessing Writing*, *27*, 37-53. https://doi.org/10.1016/j.asw.2015.11.001