Pearson

GSL Research Series

Extension de la Global Scale of English (GSE) à la Global Scale of Languages (GSL)

Partie 1 : Aligner les descripteurs du français sur la GSL

Septembre 2023

Ying Zheng, University of Southampton
Catherine Doyle, Pearson
David Booth, Pearson
Mike Mayor, Pearson

Global Scale of Languages
Fast-track your progress

Sommaire

Résumé exécutif

La Global Scale of English (GSE) offre un moyen plus détaillé de décrire et d'évaluer les progrès et les performances des apprenants de la langue anglaise. Pearson a mené une étude approfondie (**Pearson**) sur l'utilisation des descripteurs de la GSE comme échelle de référence pour étendre l'ensemble des expressions d'aptitude des apprenants du Cadre européen commun de référence pour les langues (CECR) de 2001, afin de répondre aux besoins d'un plus grand nombre d'apprenants.

Cette étude a comparé l'ordre de classement des descripteurs de la GSE traduits en français afin de déterminer si les valeurs existantes de la GSE sont applicables aux apprenants adultes du français langue étrangère (FLE). 320 descripteurs ont été traduits en français. Un panel de 20 évaluateurs qualifiés issus d'un vivier d'enseignants du FLE a été invité à comparer 25 jugements comparatifs par descripteur, ce qui a permis d'obtenir 16 000 points de données. Plusieurs analyses, notamment les statistiques adéquates des évaluateurs et des items, ont été effectuées afin d'évaluer la difficulté des descripteurs actuels pour l'anglais dans les logits de Rasch et les comparer aux estimations du jugement comparatif entre les versions anglaise et française pour quatre compétences linguistiques. Des équations de transformation ont été dérivées de ces comparaisons afin d'aligner les résultats des descripteurs pour le français sur la GSE actuelle, ce qui a conduit à la création d'une nouvelle Global Scale of French.

Pour de plus amples informations sur la Global Scale of Languages, veuillez consulter la page **pearson.com/languages**.

1. Présentation de la GSE et des descripteurs de la GSE

La GSE est une échelle de compétence normalisée de 10 à 90 pour l'anglais, alignée de manière psychométrique sur le Cadre européen commun de référence pour les langues (CECR, Conseil de l'Europe, 2001). Un ensemble de descripteurs de la GSE a été développé pour décrire les compétences de l'apprenant à chaque point de l'échelle, en intégrant et en étendant l'ensemble des descripteurs du CECR. Ces descripteurs ont été évalués par des enseignants d'anglais langue étrangère (EFL) et calibrés par rapport à la Global Scale of English (de Jong, Mayor et Hayes, 2016). Contrairement au CECR et à d'autres échelles qui décrivent les compétences acquises à l'aide de scores généraux, la Global Scale of English identifie la capacité d'un apprenant à chaque point de l'échelle dans les domaines de l'expression orale, de l'écoute, de la lecture et de l'écriture, afin de fournir une description plus détaillée de la maîtrise croissante de la langue. Le travail d'élaboration des descripteurs de la GSE s'appuie sur l'étude menée par Brian North et le Conseil de l'Europe lors de la création du CECR (North, 2000) tout en l'étendant. Les descripteurs de la GSE ont été élaborés par Pearson English depuis plusieurs années en collaboration avec plus de 6 000 enseignants, auteurs et spécialistes de l'enseignement de l'anglais du monde entier.

2. Objet de l'étude

Le but de cette étude est de comparer l'ordre de classement des descripteurs de la GSE qui ont été traduits en français, afin de voir si les valeurs actuelles de la GSE sont applicables aux apprenants adultes du français langue étrangère, c'est-à-dire s'ils peuvent être placés sur la même échelle. L'hypothèse de travail est la suivante : étant donné que la GSE est basée sur le CECR, lui-même indépendant des langues, on s'attend à ce que l'ordre général soit fortement corrélé à la fois pour la GSE et pour le CECR. Ce projet vise donc à vérifier cette hypothèse en utilisant l'approche du jugement comparatif.

3. Méthodologie

3.1 Le jugement comparatif et ses applications

Le jugement comparatif implique les jugements holistiques par un groupe de juges indépendants de paires de travaux rédigés par des élèves, afin de déterminer celui dont la construction globale spécifiée est la meilleure. Il en résulte une matrice de décision binaire du « gagnant » et du « perdant » pour chaque paire de travaux, qui est ensuite ajustée au modèle Bradley-Terry (Bradley et Terry, 1952) afin de produire des valeurs de paramètres (scores) et des erreurs types pour chaque travail d'élève. La valeur des paramètres permet d'établir un ordre de classement des travaux de ces élèves, du « meilleur » au « pire », pouvant être utilisé à des fins d'évaluation telles que la notation.

En plus de son utilisation par les jurys d'examen britanniques pour examiner la comparabilité des comités d'examen (par exemple, Fearnley, 2000 ; Gray, 2000), la comparabilité des normes dans le temps et le maintien des normes (par exemple, Chambers et Cunningham, 2022), le jugement comparatif a également été appliqué à divers contextes éducatifs. Il s'agit notamment de l'évaluation par les pairs de rapports de projet de conception créative d'étudiants de premier cycle (Mentzer *et al.*, 2021), de tests écrits sur la compréhension conceptuelle d'un cours de mathématiques (Jones et Alcock, 2014), de l'évaluation par les enseignants de statistiques sommatives et d'évaluations d'anglais (Marshall *et al.*, 2020), d'essais (Steedle et Ferrara, 2016) et de textes argumentatifs (Lesterhuis *et al.*, 2022). Pearson a utilisé le jugement comparatif pour aligner les descripteurs de la Global Scale of English (GSE) pour les jeunes apprenants sur les compétences de la Chinese Scale of English (CSE) en comparant la difficulté des descripteurs dans chaque norme (Pearson, 2020).

Le fondement psychologique du jugement comparatif est que les êtres humains sont capables de comparer un objet à un autre, mais ne sont pas assez fiables lorsqu'il s'agit d'évaluer des objets isolément (Gill et Bramley, 2013 ; Thurstone, 1927). Les approches analytiques traditionnelles impliquent que les enseignants notent le travail de chaque élève individuellement, de manière absolue, à l'aide de grilles d'évaluation, ce qui peut conduire à différentes interprétations et applications des descripteurs de ces grilles, ainsi qu'à la possibilité de s'inspirer de leur perception du travail d'autres élèves. En revanche, le jugement comparatif minimise cette influence comparative à partir des grilles d'évaluation détaillées et spécifiques (Pollitt, 2004). Il exploite directement l'aspect comparatif de l'évaluation, en se passant des grilles et de la notation. Les articles bibliographiques précédents ont montré la manière dont le jugement comparatif répond à des normes strictes de validité, de fiabilité et d'efficacité.

3.2 Conception de l'étude

L'outil de jugement comparatif *NoMoreMarking* (Wheadon, 2019) a été utilisé pour réaliser cette étude. Le nombre de fois qu'un objet donné est jugé par rapport à un autre est un élément important dans une étude de jugement comparatif. Verhavert *et al.* (2019) recommandent d'avoir 10 à 30 comparaisons par objet pour assurer une fiabilité acceptable. Conformément à cette recommandation, 25 comparaisons par descripteur ont été collectées pour garantir une conception robuste.

Dans cette étude, nous avons sélectionné 320 descripteurs de la GSE pour adultes, ce qui représente 30 % du nombre total disponible. En termes de taille et de sélection de l'échantillon, 20 % est généralement le chevauchement minimum nécessaire pour aligner les échelles (Kolen et Brennan, 2004). L'échantillon est stratifié pour être représentatif à la fois du nombre de descripteurs dans chacune des quatre compétences et pour chaque niveau du CECR (cf. tableau 1 ci-dessous).

Tableau 1 : Répartition des descripteurs

CECR/GSE	Écoute	Lecture	Expression orale	Écriture	TOTAL	% de la base de données
Inférieur à A1 (10-21)	3	3	10	4	20	*34 %*
A1 (22-29)	5	5	14	8	32	*27 %*
A2 (30-35)	6	6	17	10	39	*30 %*
A2+ (36-42)	6	6	16	10	38	*27 %*
B1 (43-50)	7	7	18	11	43	*35 %*
B1+ (51-58)	7	7	18	11	43	*33 %*
B2 (59-66)	7	7	19	11	44	*28 %*
B2+ (67-75)	5	5	14	8	32	*27 %*
C1 (76-84)	3	3	10	4	20	*28 %*
C2 (85-90)	1	1	4	3	9	*47 %*
TOTAL	**50**	**50**	**140**	**80**	**320**	***30 %***
% de la base de données	*26 %*	*35 %*	*28 %*	*33 %*	*30 %*	

Il a également été tenu compte de la diversité et de l'étendue des fonctions linguistiques, ainsi que de la nécessité d'éviter de sélectionner des descripteurs très similaires. Certains des descripteurs originaux du CECR (qui sont aussi inclus dans la GSE) ont également été sélectionnés pour fournir des liens statistiques vers le modèle CECR/North.

3.2.1 Traductions des descripteurs : de l'anglais au français

Les 320 descripteurs de la GSE ont été traduits en français par une agence de traduction. L'agence a reçu les traductions officielles du CECR du Conseil de l'Europe de l'anglais vers le français pour référence. Après les traductions initiales, deux membres de l'équipe de publication de Pearson en France ont été invités à vérifier le travail sur place. Des problèmes mineurs dans la traduction ont été identifiés et rectifiés avant que la version française finale des descripteurs ne soit utilisée.

Pour s'assurer que les versions française et anglaise soient évaluées dans le même cadre de référence, la version française et la version anglaise du même descripteur ont été placées dans le même pool de notation (divisé par compétences), de sorte que les évaluateurs ont examiné soit deux versions françaises, soit deux versions anglaises, soit une version de chaque.

3.2.2 Sélection des évaluateurs

Les évaluateurs ont été recrutés au sein d'un vivier d'enseignants du français langue étrangère qui étaient ou avaient été employés comme examinateurs pour le français dans le cadre du General Certificate of Secondary Education (GCSE) et/ou du A-Level (diplômes d'études secondaires/du lycée au Royaume-Uni, équivalent du Baccalauréat français) par le jury d'examen Pearson Edexcel. 138 personnes ont exprimé leur intérêt à participer à la recherche et fourni leurs parcours d'enseignement. Sur la base de leur expérience dans l'enseignement aux apprenants adultes, ainsi que de leur connaissance du CECR, 20 évaluateurs ont été sélectionnés pour le projet. Il a également été tenu compte de la nécessité de créer un groupe d'évaluateurs aussi diversifié que possible en termes de genre, de nationalité et d'expérience. En outre, l'ensemble des évaluateurs avaient de l'expérience dans l'enseignement d'au moins une autre langue en plus du français (voir l'annexe pour les données démographiques des évaluateurs).

Les évaluateurs ont ensuite reçu des instructions écrites sur la tâche à effectuer et la plateforme utilisée avant d'être invités à porter un jugement comparatif basé sur la question suivante : « Lequel de ces descripteurs décrit la compétence la plus difficile pour un apprenant en langues ? »

3.2.3 Description de l'ensemble de données

Tableau 2 : Nombre de descripteurs et de comparaisons pour chaque compétence

Compétence	Descripteurs de l'anglais	Descripteurs du français	Nombre total de jugements
Écoute	50	50	2 500
Lecture	50	50	2 500
Expression orale	140	140	7 000
Écriture	80	80	4 000
TOTAL	**320**	**320**	**16 000**

4. Résultats

Dans le cadre d'un jugement comparatif, la fiabilité de l'échelle de séparation est utilisée comme indicateur de fiabilité, en l'occurrence la fiabilité de l'ordre de classement des descripteurs produits par l'activité de jugement comparatif. La fiabilité de l'échelle de séparation est rapportée sur une échelle de 0 à 1, les valeurs supérieures à 0,90 indiquant une échelle de jugement comparatif très fiable. Le tableau 3 ci-dessous présente la fiabilité de l'échelle de séparation pour les quatre compétences.

Tableau 3 : Fiabilité de la séparation des échelles

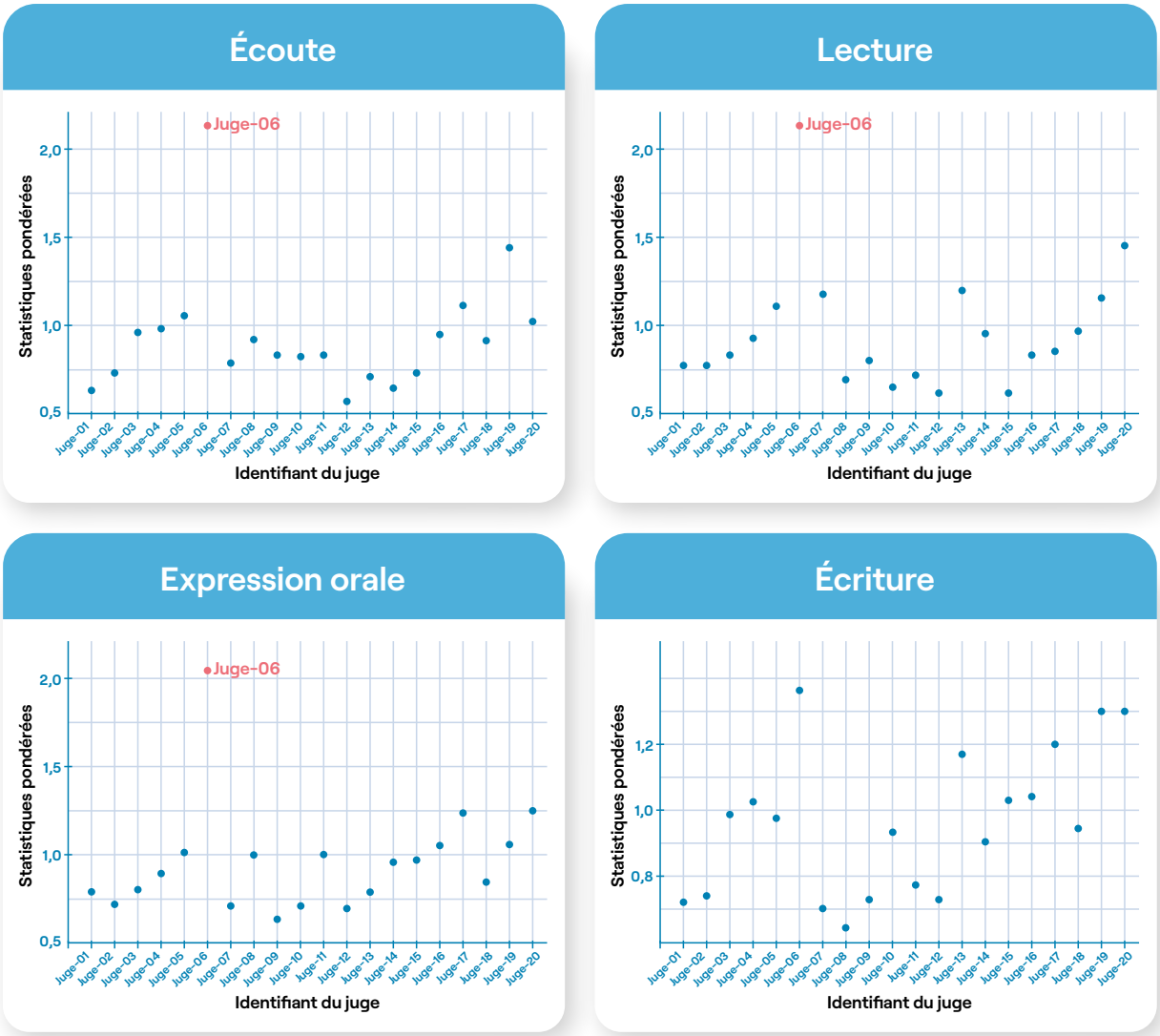
Écoute	Lecture	Expression orale	Écriture
0,939	0,938	0,938	0,943

4.1 Statistiques pondérées des juges

Des statistiques adéquates ont été calculées à la fois pour les évaluateurs et les items (c.-à-d. les descripteurs) utilisés dans cet exercice de jugement comparatif. Les évaluateurs dont les statistiques adéquates étaient supérieures de plus de deux écarts-types des statistiques pondérées moyennes ont été exclus, car cela indique qu'ils ont pu juger de manière incohérente ou qu'ils ne sont pas alignés sur le consensus des autres évaluateurs. 19 évaluateurs sur 20 (95 %) ont obtenu des statistiques pondérées acceptables pour les quatre exercices. Un évaluateur a fait preuve d'inadéquation dans l'écoute, la lecture et l'expression orale (peut-être en raison d'une mauvaise interprétation de la tâche). Bien que les statistiques pondérées de cet évaluateur pour l'écriture ne se situent pas en dehors de la fourchette acceptable, les statistiques d'inadéquation de l'évaluation pour cette tâche étaient encore les plus élevées parmi l'ensemble des évaluateurs.

Par conséquent, les données d'évaluation pour les quatre tâches ont été retirées des analyses ultérieures. En supprimant les données de cet évaluateur, la corrélation globale entre les valeurs existantes de la GSE et les scores de jugement comparatif s'est améliorée en passant de 0,79 à 0,93. Veuillez vous reporter à la figure 1 pour une représentation visuelle des statistiques pondérées des évaluateurs.

Figure 1 : Statistiques pondérées des juges (quatre compétences)



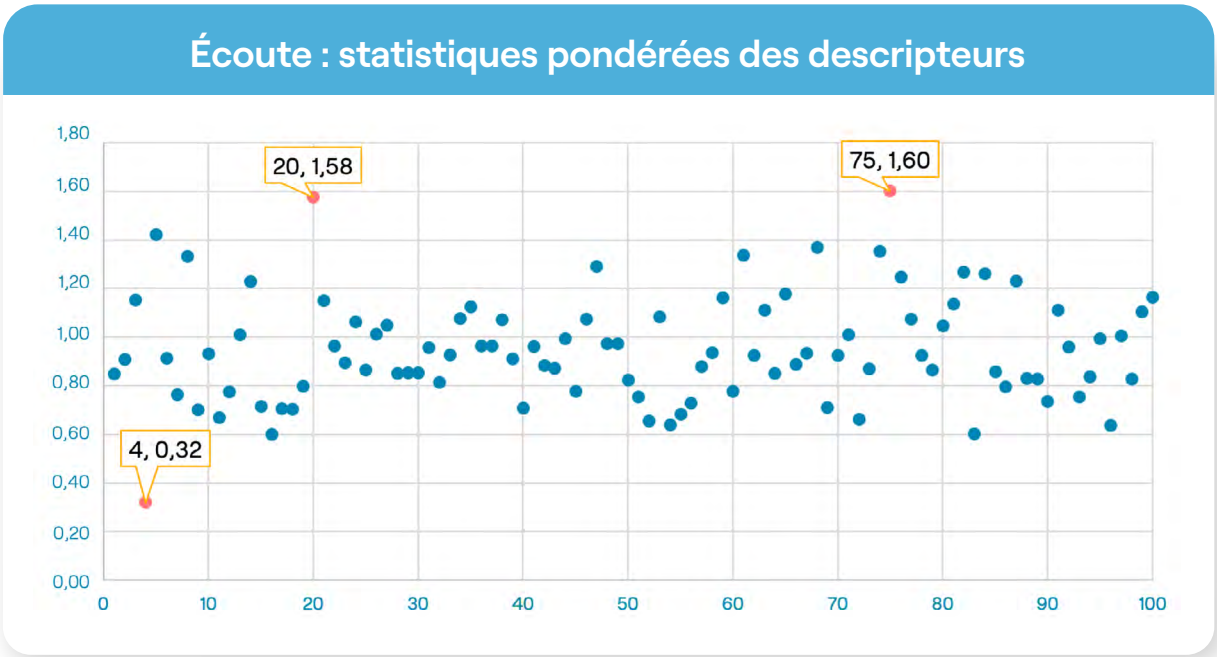
4.2 Statistiques pondérées des descripteurs

Les sections suivantes présentent les statistiques pondérées des descripteurs pour les quatre compétences. Les figures 2, 4, 6 et 8 montrent les nuages de points pour chaque compétence, l'axe Y indiquant les statistiques pondérées des items et l'axe X indiquant le nombre d'items. Il y a 100 items d'écoute, 100 items de lecture, 160 items d'écriture et 260 items d'expression orale.

4.2.1 Écoute

La corrélation globale entre les valeurs existantes de la GSE des descripteurs pour l'écoute (en anglais et en français) et les valeurs de jugement comparatif est de **0,931**.

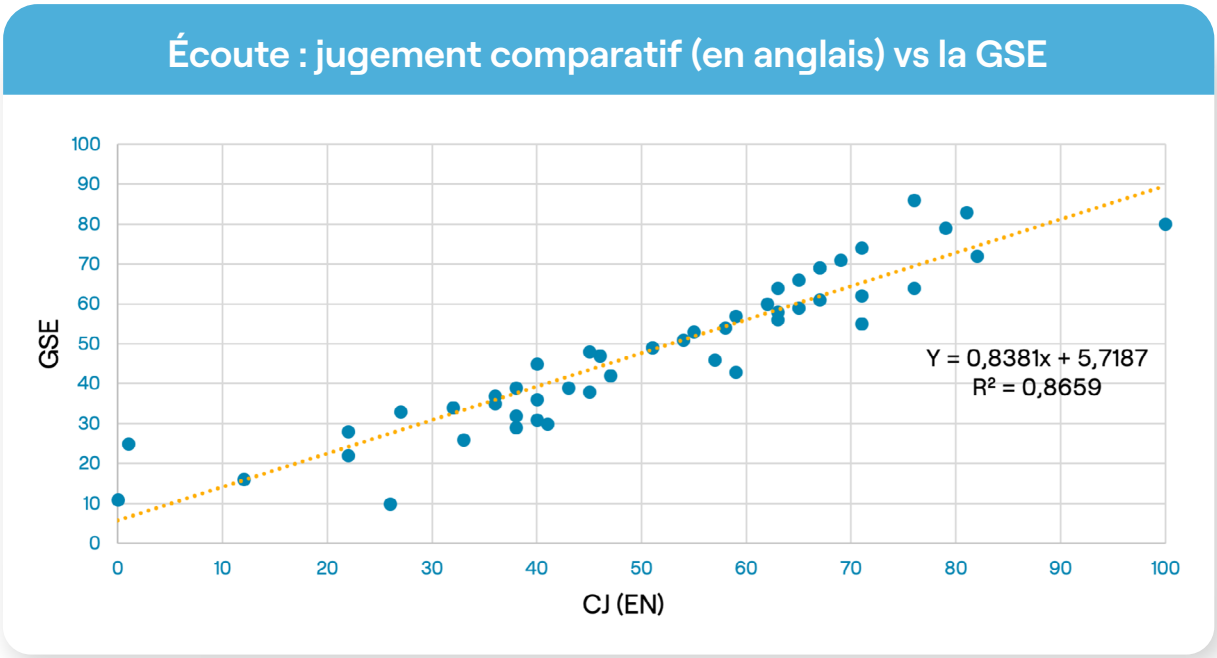
Figure 2 : Écoute – Statistiques pondérées des descripteurs

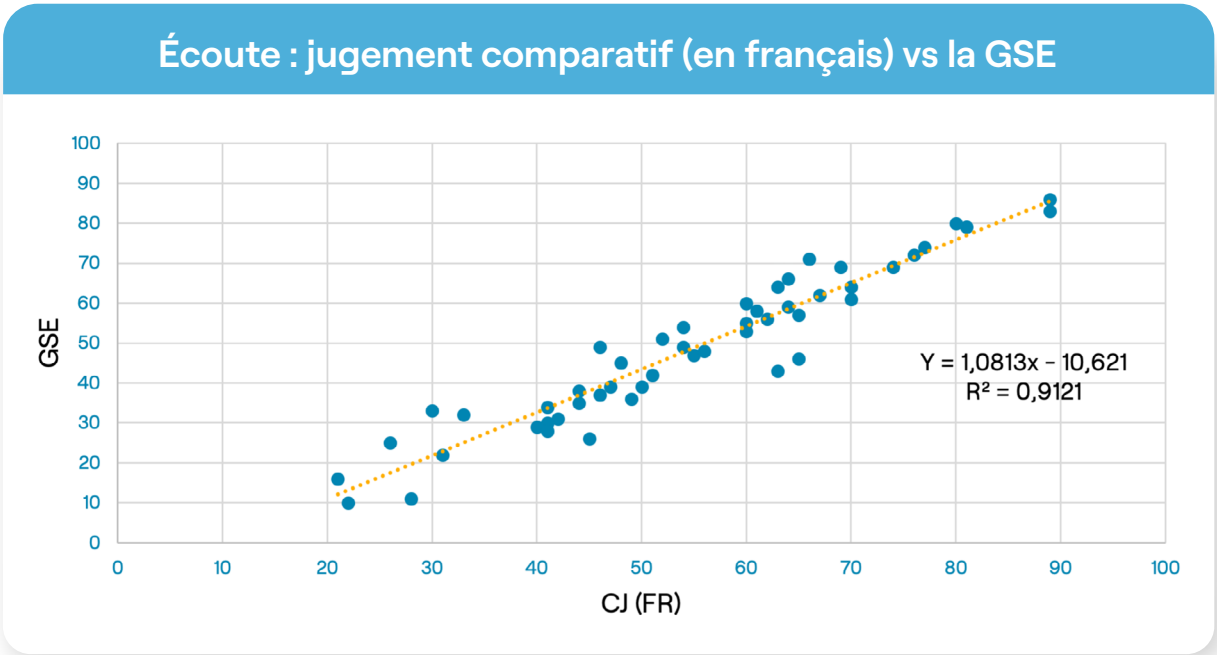


Les trois descripteurs mis en évidence dans la figure 2 ont été signalés pour des contrôles qualitatifs supplémentaires. Comme ils se situaient dans la fourchette acceptable des statistiques pondérées, ils ont été maintenus dans l'étude. *[Remarque : dans la figure ci-dessus, les points rouges indiquent l'identifiant de l'item suivi de sa statistique pondérée.]*

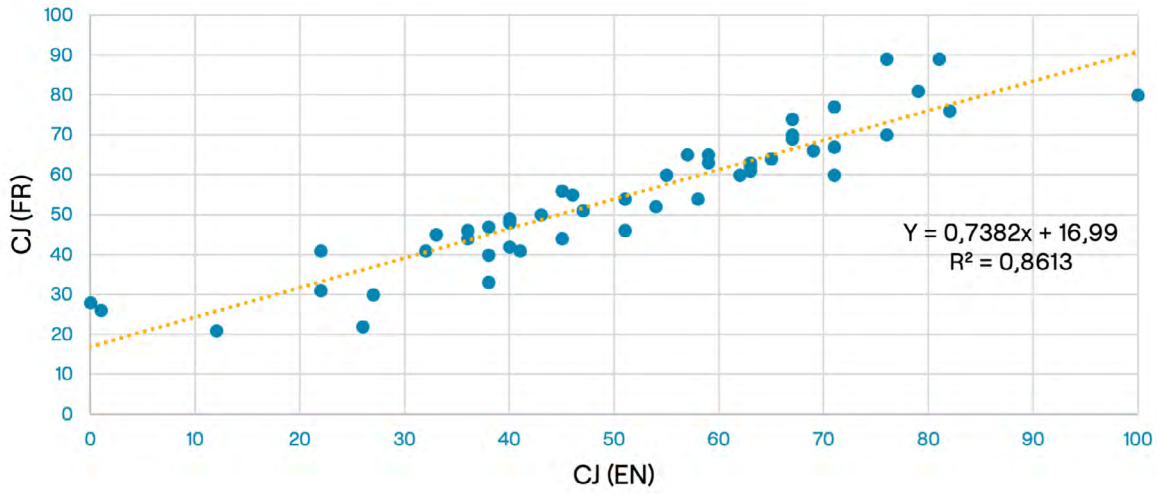
Les trois graphiques (cf. figure 3) ci-dessous montrent les comparaisons des valeurs existantes de la GSE pour l'écoute avec les estimations de jugement comparatif des descripteurs en anglais, et avec les estimations de jugement comparatif des descripteurs en français, ainsi qu'avec les deux estimations de jugement comparatif des mêmes descripteurs dans les deux langues. Les scores de jugement comparatif échelonnés fournis par *NoMoreMarking* ont été utilisés pour la comparaison. Les résultats indiquent que ces comparaisons sont très proches les unes des autres.

Figure 3 : Écoute – Comparaison entre la difficulté des descripteurs actuels et les estimations du jugement comparatif





Écoute : jugement comparatif (en anglais) vs jugement comparatif (en français)

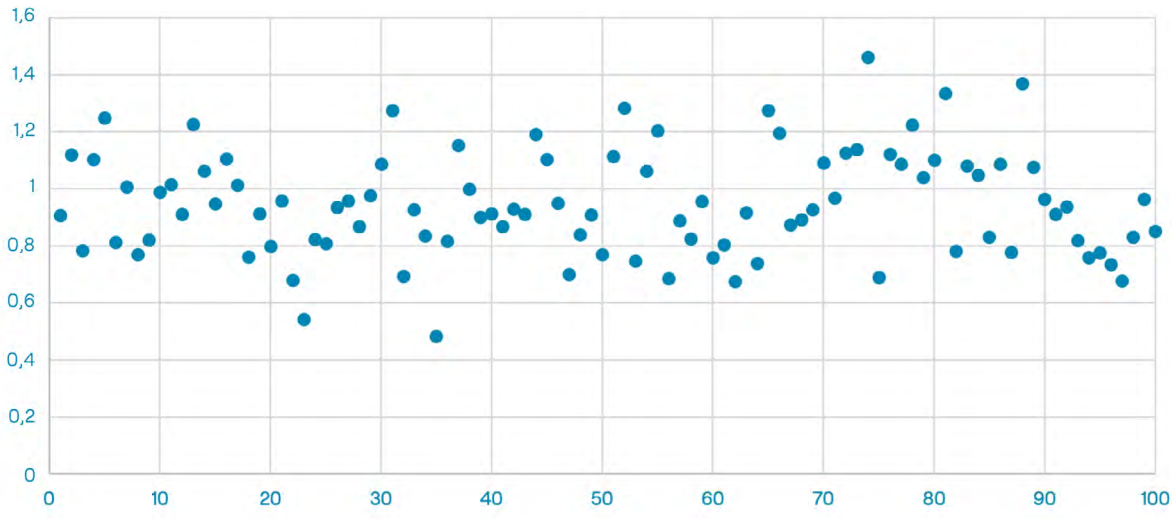


4.2.2 Lecture

La corrélation globale entre les valeurs existantes de la GSE des descripteurs pour la lecture (en anglais et en français) et les valeurs de jugement comparatif est de **0,923**. Aucun descripteur n'a été signalé pour un examen plus approfondi (cf. figure 4).

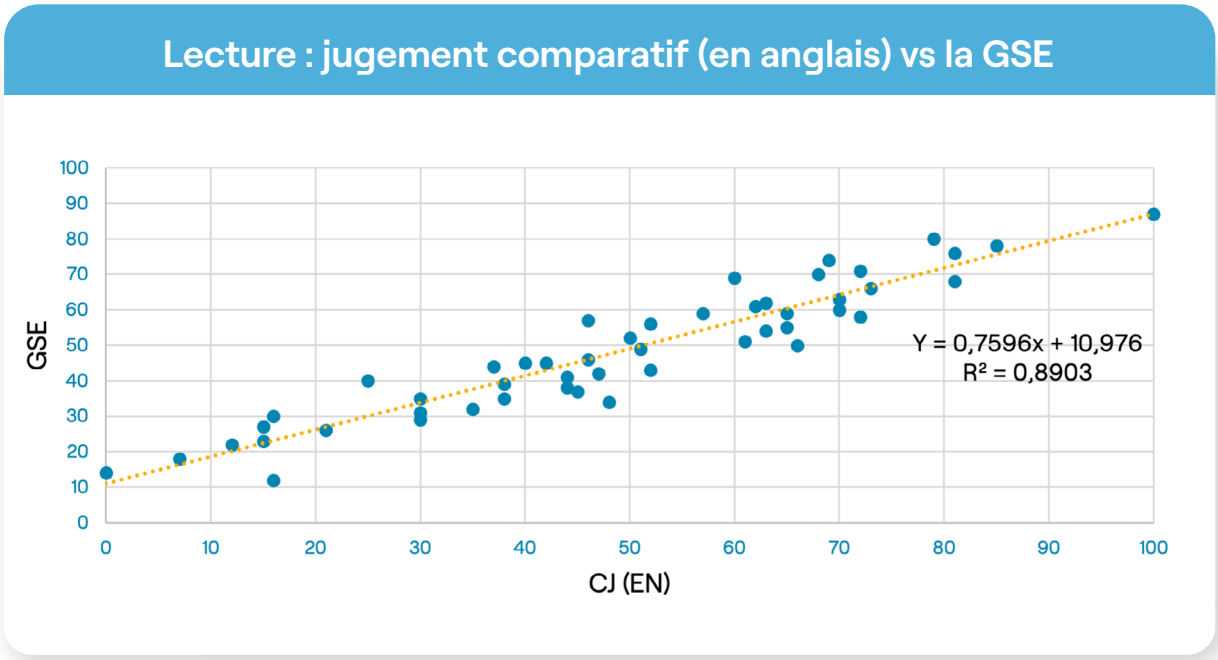
Figure 4 : Lecture – Statistiques pondérées des descripteurs

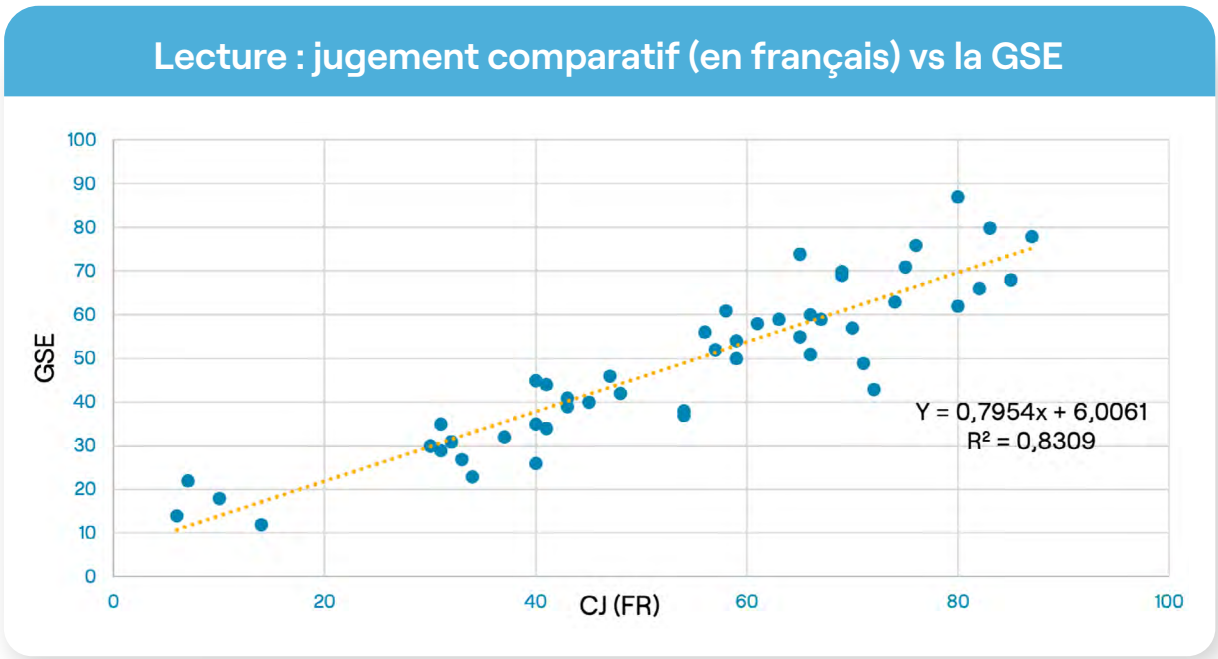
Lecture : statistiques pondérées des descripteurs



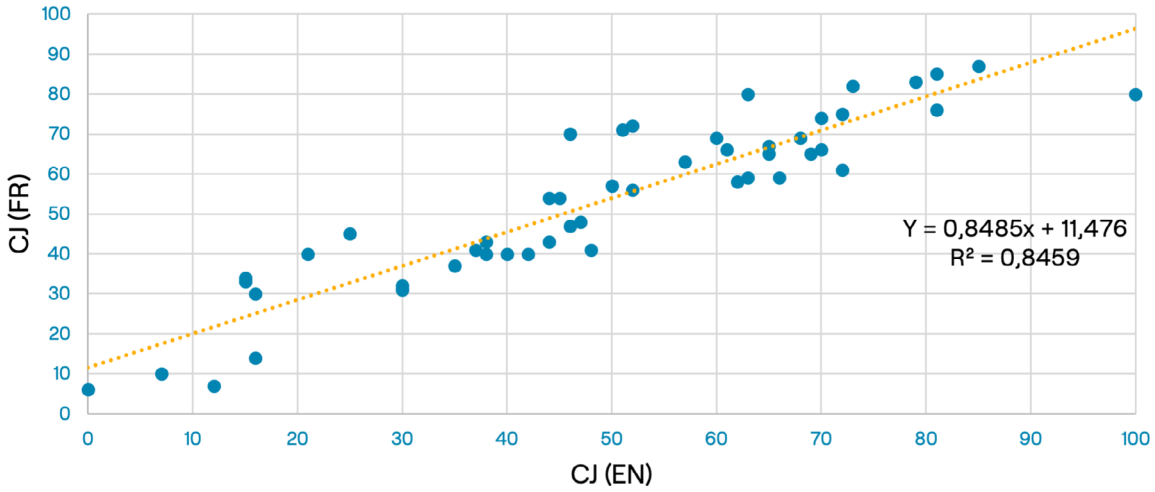
Les trois graphiques ci-dessous (cf. figure 5) montrent les comparaisons des valeurs existantes de la GSE pour la lecture avec les estimations de jugement comparatif des descripteurs en anglais, et avec les estimations de jugement comparatif des descripteurs en français, ainsi qu'avec les deux estimations de jugement comparatif des mêmes descripteurs dans les deux langues. Les résultats indiquent que ces comparaisons sont très proches les unes des autres.

Figure 5 : Lecture – Comparaison entre la difficulté des descripteurs actuels et les estimations du jugement comparatif





Lecture : jugement comparatif (en anglais) vs jugement comparatif (en français)

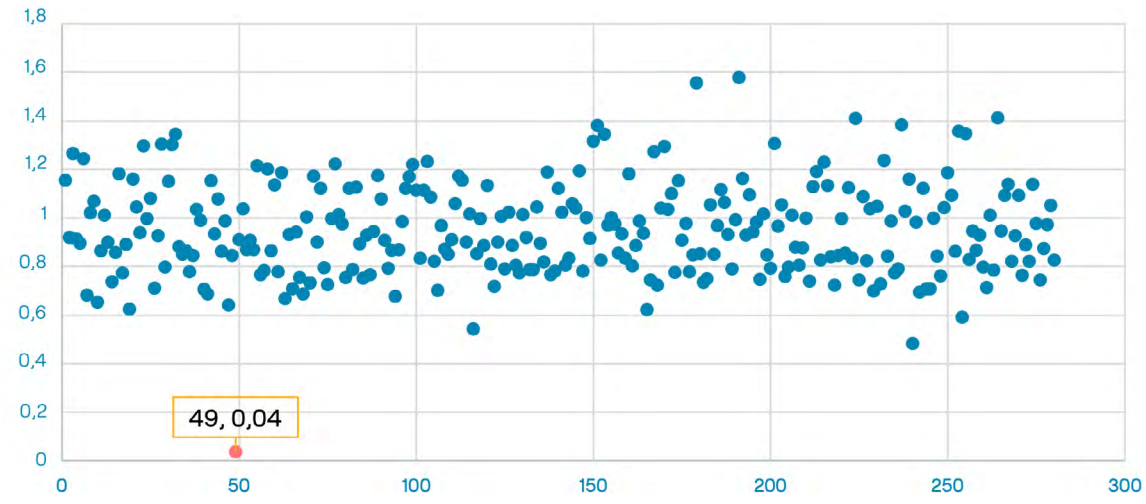


4.2.3 Expression orale

La corrélation globale entre les valeurs existantes de la GSE des descripteurs pour l'expression orale (en anglais et en français) et les valeurs de jugement comparatif est de **0,917**. Un descripteur a été signalé pour un examen plus approfondi (cf. figure 6). *[Remarque : les deux chiffres au-dessus du point rouge indiquent l'identifiant de l'item suivi de sa statistique pondérée.]*

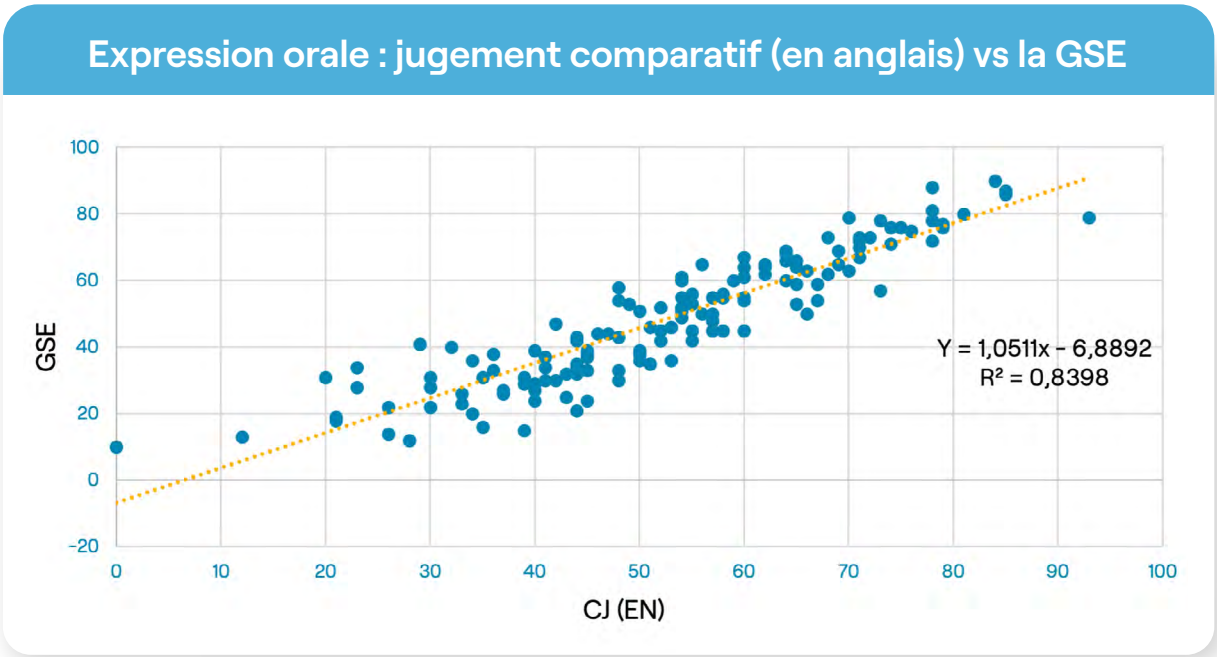
Figure 6 : Expression orale – Statistiques pondérées des descripteurs

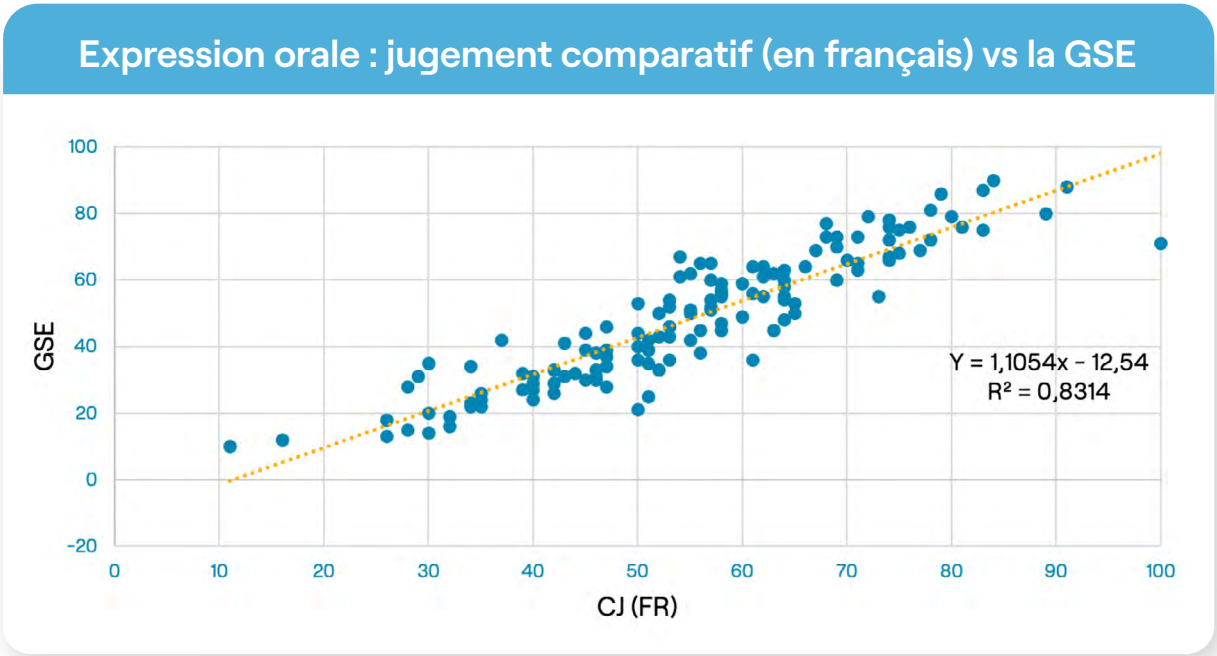
Expression orale : statistiques pondérées des descripteurs

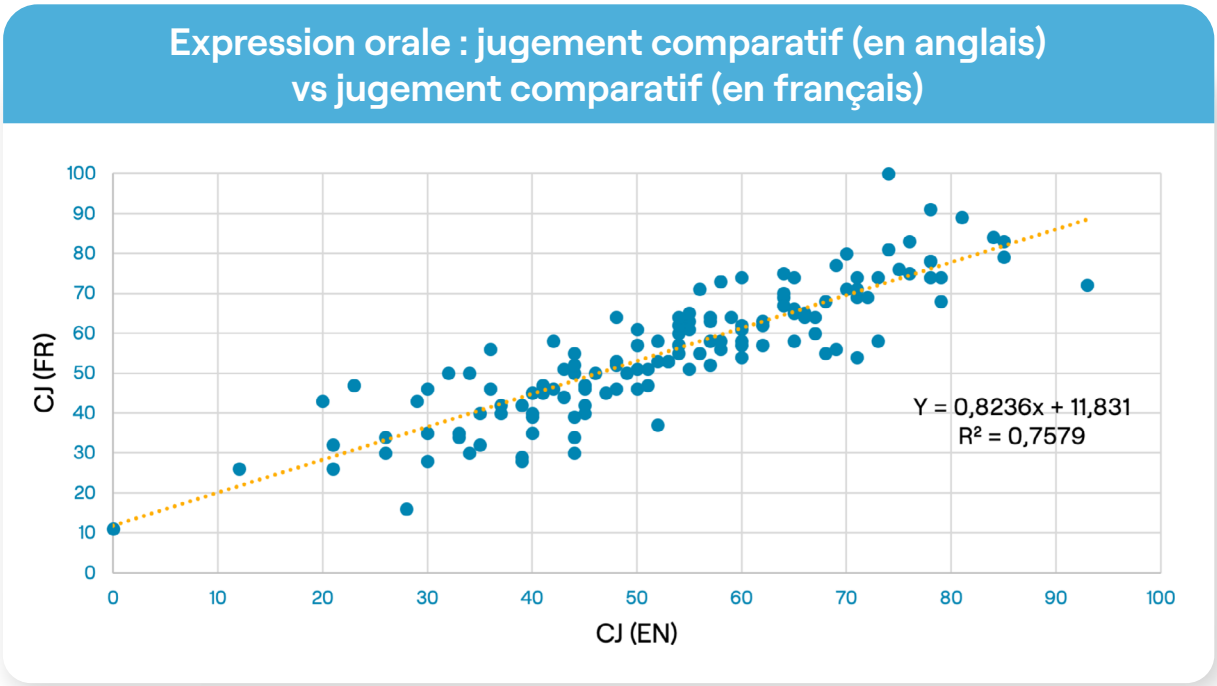


Les trois graphiques ci-dessous (cf. figure 7) montrent les comparaisons des valeurs existantes de la GSE pour l'expression orale avec les estimations de jugement comparatif des descripteurs en anglais, et avec les estimations de jugement comparatif des descripteurs en français, ainsi qu'avec les deux estimations de jugement comparatif des mêmes descripteurs dans les deux langues. Les résultats indiquent que ces comparaisons sont très proches les unes des autres.

Figure 7 : Expression orale – Comparaison entre la difficulté des descripteurs actuels et les estimations du jugement comparatif



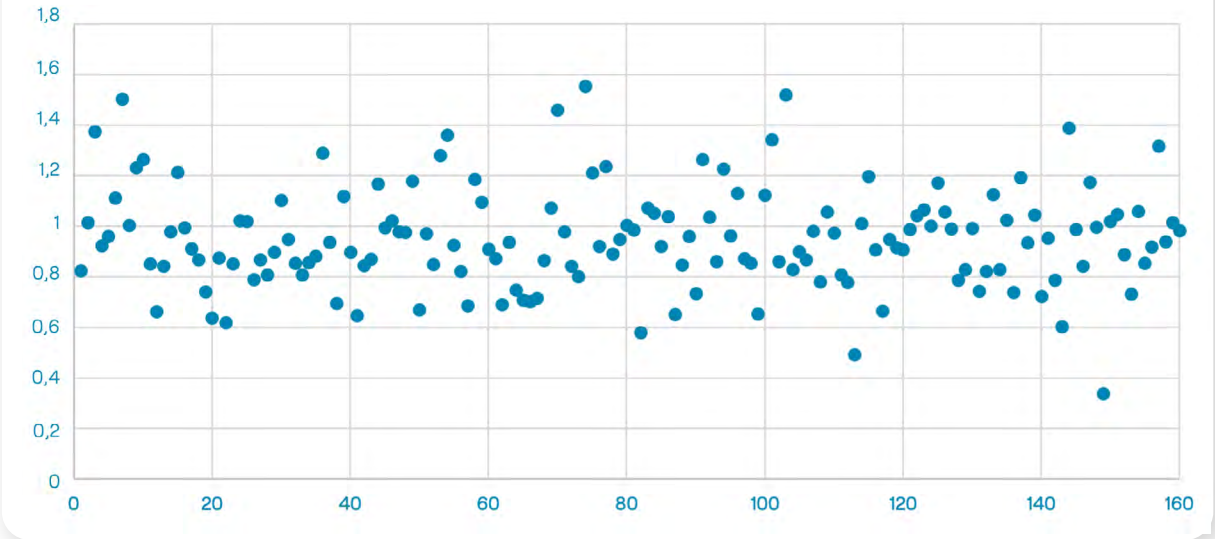




4.2.4 Écriture

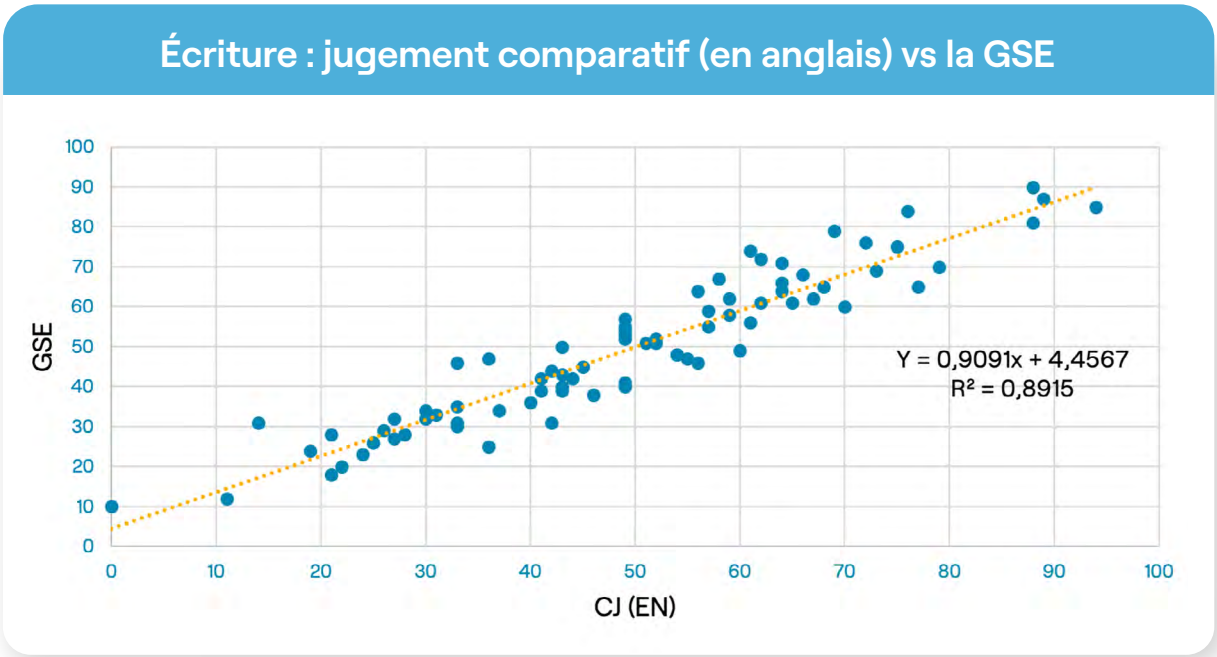
La corrélation globale entre les valeurs existantes de la GSE des descripteurs pour l'écriture (en anglais et en français) et les valeurs de jugement comparatif est de **0,925**. Aucun descripteur n'a été signalé pour un examen plus approfondi (cf. figure 8).

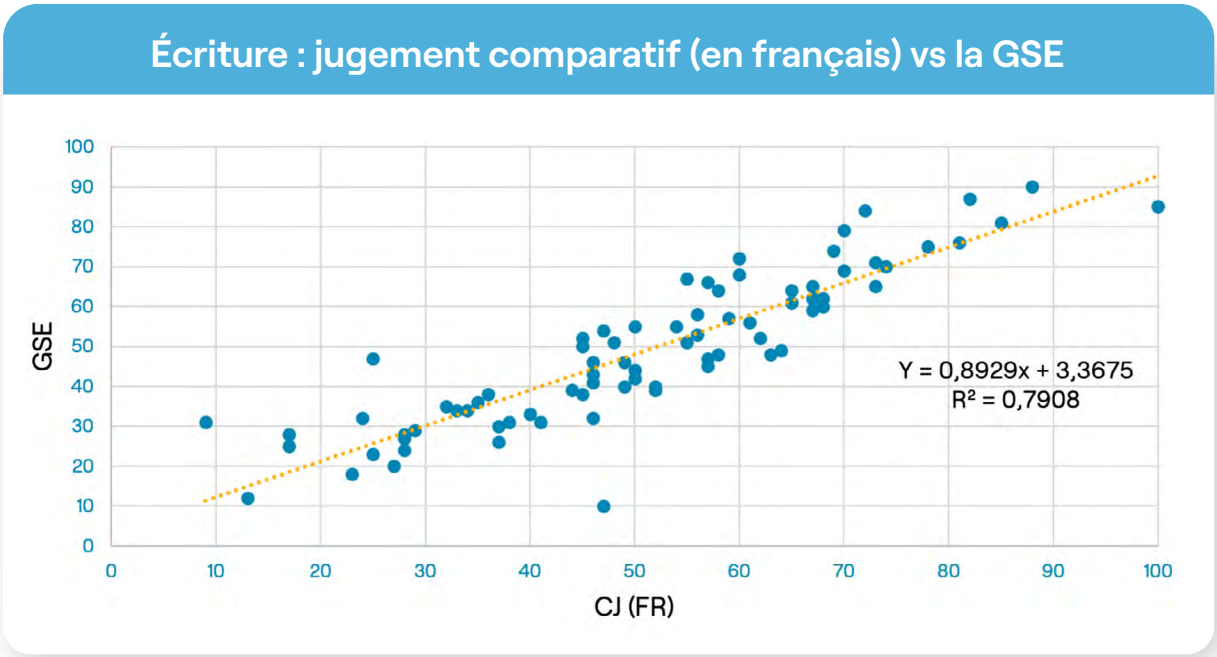
Figure 8 : Écriture – Statistiques pondérées des descripteurs



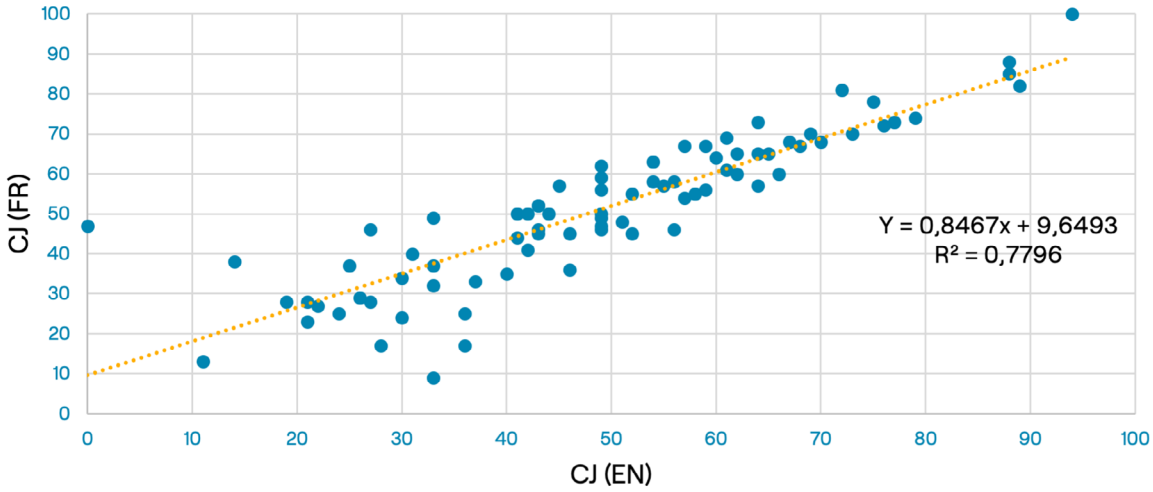
Les trois graphiques ci-dessous (cf. figure 9) montrent les comparaisons des valeurs existantes de la GSE pour l'écriture avec les estimations de jugement comparatif des descripteurs en anglais et avec les estimations de jugement comparatif des descripteurs en français, ainsi qu'avec les deux estimations de jugement comparatif des mêmes descripteurs dans les deux langues. Les résultats indiquent que ces comparaisons sont très proches les unes des autres.

Figure 9 : Écriture – Comparaison entre la difficulté des descripteurs actuels et les estimations du jugement comparatif





Écriture : jugement comparatif (en anglais) vs jugement comparatif (en français)



4.3 Corrélations entre les estimations de la GSE et du jugement comparatif

Les corrélations entre les valeurs existantes des descripteurs de la GSE et les valeurs produites dans l'étude de jugement comparatif sont présentées dans le tableau ci-dessous. Des comparaisons ont été faites entre les valeurs de jugement comparatif des descripteurs de la GSE existants et celles des traductions en français. À une exception près (pour l'écriture), toutes les corrélations sont supérieures à 0,9 dans les comparaisons.

Tableau 4 : Corrélations entre les valeurs des descripteurs de la GSE existantes et les valeurs produites dans l'étude de jugement comparatif par langue et par compétence

	Valeurs de la GSE existantes vs valeurs du jugement comparatif (anglais)	Valeurs de la GSE existantes vs valeurs du jugement comparatif (français)
Écoute	0,931	0,955
Lecture	0,944	0,912
Expression orale	0,916	0,912
Écriture	0,944	0,889

4.4 Équation de transformation pour chaque compétence

Sur la base des résultats satisfaisants obtenus jusqu'à présent, des équations de transformation ont été générées pour chaque compétence linguistique. Le tableau 5 ci-dessous présente les équations de transformation des estimations du jugement comparatif en valeurs de jugement comparatif de la GSE.

À des fins de démonstration, deux équations sont présentées pour chaque compétence, la première pour transformer les valeurs de jugement comparatif en anglais en valeurs de la GSE et la seconde pour transformer les valeurs de jugement comparatif en français en valeurs de la GSL. En pratique, seule la seconde équation est nécessaire pour aligner les descripteurs en français sur la Global Scale 10-90 de Pearson.

Tableau 5 : Équation de transformation des estimations du jugement comparatif en valeurs de la GSE

X = Score du jugement comparatif échelonné ; Y = GSE

Écoute : jugement comparatif (en anglais) vs la GSE
y = 0,8381x + 5,7187
Écoute : jugement comparatif (en français) vs la GSL
y = 1,0813x - 10,621

Lecture : jugement comparatif (en anglais) vs la GSE
y = 0,7596x + 10,976
Lecture : jugement comparatif (en français) vs la GSL
y = 0,7954x + 6,0061

Expression orale : jugement comparatif (en anglais) vs la GSE
y = 1,0511x - 6,8892
Expression orale : jugement comparatif (en français) vs la GSL
y = 1,1054x - 12,54

Écriture : jugement comparatif (en anglais) vs la GSE
y = 0,9091x + 4,4567
Écriture : jugement comparatif (en français) vs la GSL
y = 0,8929x + 3,3675

5. Discussion et conclusions

Les résultats de cette étude de jugement comparatif montrent une forte corrélation entre le niveau de compétence des mêmes descripteurs en anglais et en français. À titre de validation supplémentaire, des équations de transformation spécifiques sont fournies pour chaque compétence, ce qui conduit finalement à établir la concordance entre les estimations de difficulté des descripteurs en français de la GSE et du CECR respectivement.

Le CECR lui-même est un cadre indépendant des langues qui « peut être adapté et utilisé dans de multiples contextes et appliqué à toutes les langues » (**Conseil de l'Europe**). Depuis son développement en 2001, le CECR a été traduit en 40 langues *(ibid.)*. Le travail de Pearson pour étendre le CECR et créer la GSE a été conçu à l'origine dans un contexte d'anglais comme langue étrangère (ALE), mais il a été estimé que cette extension pourrait également être pertinente et utile pour les enseignants et les apprenants d'autres langues. L'étude basée sur le jugement comparatif décrite dans le présent article fournit des preuves qui étayent le point de vue selon lequel les actes linguistiques communicatifs et fonctionnels exprimés dans les expressions d'aptitude des apprenants en anglais et en français ont une valeur comparable en termes de compétence, c'est-à-dire qu'ils peuvent tous deux être placés sur la même échelle.

Compte tenu des similitudes entre le français et d'autres langues romanes telles que l'espagnol et l'italien, nous croyons que ces résultats peuvent être extrapolés et appliqués dans ces contextes, de la même manière que le CECR est applicable à ces langues européennes. D'autres études sont en cours pour des langues non romanes, qui viendront compléter le corpus de preuves de validation de la Global Scale of Languages.

Glossaire

CECR : Cadre européen commun de référence pour les langues

CJ : jugement comparatif

CSE : China Scale of English

GSE : Global Scale of English

GSL : Global Scale of Languages

Références

Bradley R. A. et Terry M. E. (1952), *Rank analysis of incomplete block designs. I. The method of paired comparisons,* Biometrika, 39, pp. 324-345.

Chambers L. et Cunningham E. (2022), *Exploring the Validity of Comparative Judgement: Do Judges Attend to Construct-Irrelevant Features?,* Frontiers in Education (7).

Conseil de l'Europe (2001), *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer,* Cambridge, Cambridge University Press.

De Jong J., Mayor M. et Hayes C. (2016), *Developing Global Scale of English Learning Objectives aligned to the Common European Framework,* disponible à l'adresse : https://www.pearson.com/languages/why-pearson/the-global-scale-of-english/resources.html.

Fearnley A. (2000), « A comparability study in GCSE mathematics. A study based on the summer 1998 examination », dans *Assessment and Qualifications Alliance* (Northern Examinations and Assessment Board), Manchester, Joint Forum for the GCSE and GCE.

Gill T. et Bramley T. (2013), « How accurate are examiners' holistic judgements of script quality? », *Assessment in Education: Principles, Policy & Practice,* 20 (3), pp. 308-324.

Gray E. (2000), *A comparability study in GCSE science 1998. A study based on the 1998 summer examination,* organisée par Oxford, Cambridge et RSA Examinations (Midland Examining Group) au nom du forum commun pour le GCSE et le GCE.

Jones I. et Alcock L. (2014), « Peer assessment without assessment criteria », *Studies in Higher Education,* 39(10), pp. 1774-1787.

Kolen M. J. et Brennan R. L. (2004), *Test equating, scaling, and linking: Methods and practices,* 2e éd., New York, Springer.

Lesterhuis M., Verhavert S., Coertjens L., Donche V. et De Maeyer S. (2017), « Comparative judgement as a promising alternative to score competences », dans *Innovative practices for higher education assessment and measurement,* IGI Global, pp. 119-138.

Marshall N., Shaw K., Hunter J. et Jones I. (2020), « Assessment by comparative judgement: An application to secondary statistics and English in New Zealand », *New Zealand Journal of Educational Studies,* 55, pp. 49-71.

Mentzer N., Lee W. et Bartholomew S. R. (2021), « Examining the Validity of Adaptive Comparative Judgment for Peer Evaluation in a Design Thinking Course », dans *Frontiers in Education,* Frontiers, p. 492.

North B. (2000), *The development of a common framework scale of language proficiency,* New York, Peter Lang.

Pollitt A. (2004), *Let's stop marking exams,* International Association for Educational Assessment Conference, Philadelphia, PA.

Steedle J. T. et Ferrara S. (2016), « Evaluating comparative judgment as an approach to essay scoring », *Applied Measurement in Education,* 29 (3), pp. 211-223.

Pearson technical report (2020), « Aligning Global Scale of English-Young Learner to the CSE », disponible à l'adresse : https://m.i21st.cn/elt/15934.html.

Thurstone L. L. (1927), « A law of comparative judgment », *Psychological Review,* 34 (4), p. 273.

Verhavert S., Bouwer R., Donche V. et Maeyer S. D. (2019), « A meta-analysis on the reliability of comparative judgement », *Assessment in Education: Principles, Policy & Practice,* 26(5), pp. 541-562.

Wheadon C. (2019), *No More Marking* [Computer Software], extrait de https://www.nomoremarking.com/.

Annexe : Données démographiques des évaluateurs

Nationalité	Nombre
Britannique	8
Français	1
Italien	1
Mexicain	2
Espagnol	8
TOTAL	**20**

Genre	Nombre
Homme	7
Préfère ne pas répondre	1
Femme	12
TOTAL	**20**

Années d'enseignement du français	Nombre
> 10 ans	16
5 à 10 ans	4
TOTAL	**20**

Connaissance du CECR	Nombre
Connaissance détaillée	5
Compréhension générale	8
Au courant	7
TOTAL	**20**

Autres langues enseignées	Nombre*
Français	**16**
Anglais	**6**
Allemand	**2**
Italien	**2**
Latin	**2**
Grec ancien	**1**
Japonais	**1**
Catalan	**1**

* Tous ont enseigné au moins une autre langue

Groupes d'âge des apprenants (français)	Nombre*
Adultes (18 ans et plus)	**20**
Secondaire/lycée (15 à 19 ans)	**18**
Collège (12 à 15 ans)	**18**
Primaire grande section (9 à 12 ans)	**12**
Primaire petite section (6 à 9 ans)	**3**
Maternelle (3 à 5 ans)	**2**

Global Scale of Languages

Fast-track your progress

Pearson