Mesure et évaluation en éducation

Utilité des échelles descriptives et différences individuelles dans l'autoévaluation de l'écrit

Dany Laveault and Carol Miles

Volume 31, Number 1, 2008

URI: https://id.erudit.org/iderudit/1025011ar
DOI: https://doi.org/10.7202/1025011ar

See table of contents

Publisher(s)

ADMEE-Canada - Université Laval

ISSN

0823-3993 (print)
2368-2000 (digital)

Explore this journal

Cite this article

Laveault, D. & Miles, C. (2008). Utilité des échelles descriptives et différences individuelles dans l'autoévaluation de l'écrit. *Mesure et évaluation en éducation*, *31*(1), 1–29. https://doi.org/10.7202/1025011ar

Article abstract

The purpose of this research was to study individual differences among students using rating scales ("rubrics") to assess a complex performance. More precisely, we tried to determine which characteristics of a student's judgment in terms of accuracy, severity and confidence have an impact on writing achievement. Accuracy was found to be the best predictor of writing achievement. Several important differences in the students' ability to use rating scales were found among students with learning difficulties, gifted and talented students. Despite the fact that girls usually had better marks in writing, no differences were found between boys and girls as far as accuracy, severity and confidence were concerned. Multiple regressions on francophone and anglophone students' results indicated that rating scales contributed in the same way to writing achievement in both groups. Degree of confidence in one's assessment of a peer's writing was significantly lower among francophones. No significant improvement in judgment accuracy was found from grade 5 to grade 8. The research concludes by stressing the importance of socio-affective aspects of a student's judgment in the use of rating scales.

Utilité des échelles descriptives et différences individuelles dans l'autoévaluation de l'écrit

Dany Laveault
Université d'Ottawa

Carol Miles
Carleton University

Mots clés : Évaluation par des pairs, autoévaluation, échelle descriptive, performance complexe, copie type, différences individuelles, écriture

Cette recherche porte sur les différences individuelles entre élèves lorsque ceux-ci utilisent des échelles descriptives pour évaluer une production complexe. Plus précisément, elle a cherché à déterminer quelles caractéristiques du jugement de l'élève, pour ce qui est de la précision, de la sévérité et de la confiance, affectent leur rendement en écriture. La précision du jugement s'est avérée le meilleur prédicteur du rendement en écriture. Des différences importantes quant à l'habileté avec laquelle les élèves des groupes talentueux, doués ou en difficulté d'apprentissage utilisent les échelles descriptives ont également été observées. Même si les filles obtiennent de meilleures notes que les garçons à l'écrit, les différences entre les deux groupes quant à la précision, à la sévérité et à la confiance du jugement ne sont pas significatives. La comparaison des groupes francophone et anglophone au moyen des régressions multiples séparées indique que la compréhension des échelles descriptives contribue de la même manière au rendement en écriture dans chacun des groupes. Le degré de confiance, toutefois, s'est révélé significativement plus bas chez les francophones. Il n'y pas eu d'amélioration significative dans la capacité à évaluer des textes écrits entre la 5e et la 8e année. La recherche conclut en soulignant l'importance des dimensions socioaffectives du jugement de l'élève dans l'utilisation d'échelles descriptives.

Note des auteurs – Toute correspondance peut être adressée comme suit : Dany Laveault, Faculté d'éducation, Université d'Ottawa, Pavillon Lamoureux, bureau 474, 145, rue Jean-Jacques Lussier, Ottawa, Ontario, K1N 6N5, ou par courriel aux adresses suivantes : [dlaveaul@uottawa.ca] [carol_miles@carleton.ca].

Key words: Peer assessment, self-assessment, rating scales, complex performance, exemplar, individual differences, writing

The purpose of this research was to study individual differences among students using rating scales ("rubrics") to assess a complex performance. More precisely, we tried to determine which characteristics of a student's judgment in terms of accuracy, severity and confidence have an impact on writing achievement. Accuracy was found to be the best predictor of writing achievement. Several important differences in the students' ability to use rating scales were found among students with learning difficulties, gifted and talented students. Despite the fact that girls usually had better marks in writing, no differences were found between boys and girls as far as accuracy, severity and confidence were concerned. Multiple regressions on francophone and anglophone students' results indicated that rating scales contributed in the same way to writing achievement in both groups. Degree of confidence in one's assessment of a peer's writing was significantly lower among francophones. No significant improvement in judgment accuracy was found from grade 5 to grade 8. The research concludes by stressing the importance of socio-affective aspects of a student's judgment in the use of rating scales.

Palavras-chave: Avaliação pelos pares, auto-avaliação, escala descritiva, desempenho complexo, teste-modelo, diferenças individuais, escrita

Esta investigação incide nas diferenças individuais entre alunos quando estes utilizam escalas descritivas para avaliar uma produção complexa. Mais precisamente, procurou determinar quais as categorias do juízo do aluno, em termos de precisão, de severidade e de confiança, que afectam o seu rendimento na escrita. A precisão do juízo revelou-se o melhor indicador de rendimento, na escrita. Observaram-se, também, diferenças significativas, relativas à capacidade dos alunos muito bons, bons ou com dificuldades de aprendizagem para utilizar escalas descritivas. Apesar de as raparigas, na escrita, alcançarem melhores classificações do que os rapazes, as diferenças entre os dois grupos, relativas à precisão, à severidade e à confiança do juízo não são significativas. A comparação entre os grupos de alunos francófonos e angloxaxónicos, por meio de regressões múltiplas separadas, revela que a compreensão de escalas descritivas contribui, em ambos os grupos, para o rendimento na escrita. Contudo, o grau de confiança revelou-se significativamente mais baixo junto dos alunos francófonos. Não houve melhorias significativas na capacidade de avaliar os textos escritos entre o 5º e o 8º anos. A investigação conclui pela importância das dimensões socio-afectivas do juízo do aluno na utilização de escalas descritivas.

L'autoévaluation et l'évaluation par des pairs ont pris une importance accrue dans la formation à l'élémentaire et au secondaire (Ministère de l'Éducation de l'Ontario, 1997 ; Ministère de l'Éducation du Québec, 2001 ; 2003). Dans le contexte de nouveaux programmes dont les attentes sont formulées en termes de compétences, l'autoévaluation et l'évaluation par les pairs interviennent à toutes les étapes de l'apprentissage, depuis la définition des buts et des critères, en passant par l'observation d'une performance (*monitoring*), jusqu'au jugement final sur la production ou sur le processus suivi.

Malgré que ces pratiques d'évaluation contrôlées par l'élève soient de plus en plus encouragées, plusieurs recherches sur leur validité et leur utilité restent à faire. Plus précisément, il s'agit de déterminer dans quelles conditions leur usage est préférable à d'autres pratiques d'évaluation ou encore de préciser dans quels contextes elles fournissent de meilleurs résultats pour l'apprentissage. Ce type de recherche n'est pas simple, et ce, pour plusieurs raisons :

1. L'autoévaluation regroupe une famille de pratiques qui peut comprendre l'autocorrection, l'autoobservation et l'autonotation, pour ne nommer que celles-là (Allal, 1999, 2007 ; Laveault, 2007). Ces formes de pratique diffèrent entre elles par le type et le degré de contrôle exercé par l'étudiant. La différenciation de ces formes permet d'établir des nuances pertinentes, rendant d'autant plus ardue et complexe l'étude de l'impact de l'autoévaluation sur l'apprentissage.
2. À la diversité des modalités s'ajoute la multiplicité des contextes d'utilisation. En effet, l'efficacité des pratiques d'évaluation s'inscrit à l'intérieur de modalités d'enseignement spécifiques. Celles-ci varient selon les époques, les lieux et le type de population. Par exemple, comment traiter sur le même pied l'efficacité d'une forme d'évaluation par des pairs dans le contexte d'un apprentissage par objectifs à l'élémentaire avec celle d'une pratique similaire dans le contexte d'un apprentissage par compétences avec des adultes ?

L'ensemble de ces raisons nous conduit à analyser de manière critique les résultats de la recherche sur l'autoévaluation quant à sa généralisation à des contextes, des modalités et des populations variées. Nous sommes conduits à nous interroger sur les conditions qui doivent se retrouver pour assurer l'efficacité de ces pratiques en termes d'apprentissage. Pour y arriver, il faut mieux comprendre comment la personne qui s'autoévalue ou qui évalue l'un de ses pairs formule son appréciation, quels facteurs ont un impact sur son jugement et comment ce jugement affecte son rendement académique.

Plusieurs recherches ont porté sur la façon dont l'enseignant développe et utilise son jugement professionnel en matière d'évaluation des élèves (Bressoux & Pansu, 2003). L'élève est aussi intéressé par ce type de recherches, car il est important de mieux comprendre comment il forme lui-même son jugement, apprend à s'autoévaluer et comment cet apprentissage lui permet de progresser. Ce type d'apprentissage est décrit dans les écrits scientifiques en anglais sous l'expression *assessment as learning* (Stiggins, Arter, Chappuis & Chappuis, 2004). C'est précisément ce que cette recherche ambitionne d'accomplir : déterminer comment les différents aspects du jugement de l'élève ont un impact sur sa capacité à évaluer par lui-même et comment cette capacité est reliée à la réussite scolaire en écriture, l'une des productions complexes les plus fréquemment évaluées.

État de la question

Parmi les facteurs les plus souvent mentionnés, en rapport avec l'efficacité des pratiques d'autoévaluation et d'évaluation par des pairs, figurent le degré de contrôle exercé par l'élève et le degré de compréhension des critères d'évaluation. Ces facteurs ne sont pas totalement indépendants lorsque, dans certaines situations, les élèves sont appelés à définir eux-mêmes les critères qu'ils auront à appliquer.

L'appréciation de l'efficacité de l'autoévaluation et de l'évaluation par des pairs est souvent liée au degré d'accord entre les évaluations faites par l'enseignant, l'élève lui-même et ses pairs. Miller (2003) a pu démontrer que la corrélation entre l'autoévaluation et l'évaluation par des pairs était meilleure lorsque les critères étaient plus spécifiques. De plus, le recours à des critères plus spécifiques a pour effet de réduire la moyenne des résultats et d'accroître la sensibilité de l'évaluation aux différences individuelles. On peut se demander si cette réduction de la moyenne n'est pas due à un accroissement de la sévérité ou de la rigueur rendue possible par des critères plus précis et mieux compris.

Il existe plusieurs méthodes pour rendre les critères d'évaluation plus spécifiques, moins ambigus, donc plus univoques pour l'élève :

- recourir à des échelles descriptives fournissant des précisions pour chaque niveau de performance ;
- illustrer les niveaux de performance des échelles descriptives au moyen de productions typiques ou « copies types » de ces niveaux ;
- engager directement les élèves dans la définition et le choix des critères de performance qui serviront à les évaluer.

La compréhension des critères d'évaluation ressort fréquemment comme une condition d'efficacité de l'évaluation par des pairs ou de l'autoévaluation. Bergee (1997, p. 610) suggère que la capacité à s'autoévaluer peut s'améliorer si des discussions sont prévues pour aborder les écarts relevés entre les appréciations du tuteur et celles des élèves. Fallows et Chandramohan (2001, p. 243) considèrent qu'une formation à l'évaluation – centrale dans leur approche – renforce la sensibilité des étudiants au processus d'évaluation et les aide à devenir plus critiques envers leur travail et celui de leurs pairs. Beaman (1998, p. 54) a mis au point un jeu, *The Egg Game*, pour amener les étudiants à réfléchir à ce qu'implique l'évaluation par des pairs et à soulever, avant qu'ils ne surviennent, les problèmes associés à cette forme d'évaluation tels que la collusion, l'équité et la validité. Bref, la formation des élèves en tant que juges de leurs résultats exerce une influence déterminante sur leur apprentissage du processus même de l'évaluation.

Orsmond, Merry et Reiling (2002) ont démontré que l'utilisation de critères de correction conjointement avec des copies types a pour effet d'accroître le degré d'accord entre le tuteur et l'étudiant. Ce résultat diffère d'une recherche précédente d'Orsmond, Merry et Reiling (2000) alors que la discussion des critères, sans utilisation de copies types, n'avait eu aucun effet sur le degré d'accord tuteur-étudiant. Selon Orsmond et collaborateurs (2002, p. 318), la participation à la construction des critères et la discussion de ces critères au moyen de copies types a permis aux étudiants de mieux comprendre ce qui était requis et de mieux discriminer entre les niveaux de performance. Toutefois, même avec des copies types, Orsmond et collaborateurs (2002, p. 319) ne rapportent aucun effet de la discussion des critères sur le degré d'accord entre les étudiants eux-mêmes. Ils concluent que, même si les copies types peuvent aider les étudiants à mieux percevoir ce que le tuteur attend d'eux, elles n'augmentent pas nécessairement leur capacité à produire l'amélioration attendue.

L'équipe d'Orsmond (2000, p. 34) répartit les étudiants en quatre groupes quant aux difficultés qu'ils éprouvent à s'exprimer à travers un processus d'évaluation.

1. Un premier groupe s'évalue avec difficulté, sans que cela puisse être imputable à une méconnaissance des critères ou de la matière.
2. Un second groupe s'évalue avec difficulté parce que les étudiants connaissent mal les critères et la matière elle-même.
3. Un troisième groupe s'évalue avec difficulté, parce que les étudiants méconnaissent les critères, et ce, même si la matière est bien comprise.

4. Un quatrième groupe s'évalue avec difficulté, car même si les étudiants connaissent bien les critères, la matière n'est pas bien comprise.

Les résultats des recherches d'Orsmond et collaborateurs (1997, 2000, 2002) montrent que, même si l'étudiant veut être plus critique et réfléchir davantage aux questions d'évaluation, il n'est pas toujours en mesure de le faire par lui-même. Pour ce faire, il a besoin de soutien : les copies types sont un moyen parmi d'autres pour l'aider à approfondir les critères d'évaluation. L'efficacité relative des copies types n'est pas déterminée de façon certaine, car celles-ci peuvent être utilisées avec ou sans discussion avec un tuteur, ou encore être employées dans des contextes où les critères ont été définis par l'étudiant ou encore par l'enseignant. Il semble enfin que l'amélioration du degré d'accord entre un tuteur et un étudiant puisse favoriser une amélioration des apprentissages, mais qu'elle ne soit pas suffisante en elle-même.

La plupart des recherches portant sur l'autoévaluation ou sur l'évaluation par des pairs ont donc jugé de l'efficacité de ces pratiques en fonction de la concordance des jugements exprimés. Cette fidélité a été évaluée selon plusieurs méthodes : corrélation entre autoévaluation et évaluation par le tuteur, concordance entre les évaluations faites par des pairs et, enfin, concordance entre l'évaluation faite par un tuteur et celle des élèves de la classe. Par exemple, Liu, Lin et Yuan (2002, p. 401) rapportent une plus grande concordance entre l'évaluation du tuteur et celle des pairs qu'entre celle du tuteur et l'autoévaluation. L'autoévaluation peut aussi manquer de réalisme (Bishop, 1994, p. 5).

L'impact de l'autoévaluation ou de l'évaluation par des pairs sur l'apprentissage est bien illustré par l'étude de Church (1997). Elle démontre l'existence d'un lien direct entre la performance et la congruence entre résultats de l'autoévaluation et ceux de multiples sources de renseignements. Dans ce cas-ci, plus l'autoévaluation est considérée comme réaliste en rapport avec les autres sources (clients, superviseurs, etc.), meilleurs sont les progrès. L'étude de Church (1997, p. 1013) appuie l'existence d'un lien direct entre la performance et la prise de conscience de ses forces et faiblesses par un employé. Elle confirme également l'importance de faire valider l'autoévaluation par des *feed-backs* de plusieurs sources.

Pour Boud (1986, in Somervell, 1993, p. 228), l'évaluation par les pairs est considérée comme une partie intégrante de l'autoévaluation et elle contribue à l'enrichir. Miller (2003, p. 390) abonde dans le même sens et utilise le concept de « triangulation » pour illustrer la convergence des sources. À son

avis, cependant, la triangulation fait ressortir la présence de perspectives multiples qui n'ont pas à s'accorder entre elles. La complémentarité des points de vue peut être tout aussi importante que la concordance. Selon Miller (2003), il importe que les outils d'évaluation soient sensibles à différents niveaux de performance. Des outils grossiers peuvent donner lieu à des concordances élevées entre les évaluateurs, occultant ainsi des différences fines mais pertinentes dans les évaluations.

Les résultats d'Orsmond et collaborateurs (2002) indiquent que des autoévaluations fidèles (concordance élevée avec le tuteur) ne sont pas suffisantes pour améliorer les apprentissages. La question demeure donc ouverte : dans quelles conditions la capacité à évaluer ses travaux ou ceux des autres permettra à l'élève de progresser dans ses apprentissages ? Notre étude se propose de déterminer quelles sont les caractéristiques des élèves qui ont développé des habiletés à évaluer et comment ces habiletés ont un impact sur leur apprentissage. Nous avons choisi de faire porter la recherche sur le domaine particulier de l'évaluation de l'écrit parce qu'il s'agit là d'une performance complexe, fréquemment évaluée, qui nécessite la prise en compte simultanée de plusieurs critères. De plus, cette étude se situe dans le contexte particulier de la réforme du programme de français entreprise en Ontario depuis 1997, qui a apporté plusieurs changements à la méthode d'évaluation utilisée par les enseignants et les élèves.

Les enseignants et les élèves de l'Ontario utilisent maintenant des échelles descriptives (*rubrics*) en quatre points fondées sur les attentes du programme ontarien. Au centre de cette réforme, on retrouve des copies types (*exemplars*). Celles-ci répondent à deux besoins principaux :

1. accroître la cohérence des enseignants dans l'évaluation des travaux d'élèves ;
2. améliorer l'apprentissage de l'élève.

Dans le curriculum de l'Ontario (Ministère de l'Éducation de l'Ontario, 1999, p. 7), il est clairement dit : « Le rendement des élèves augmente lorsqu'elles et ils comprennent bien ce que l'on attend d'eux ainsi que les critères au moyen desquels sont évalués leurs travaux. » Quoiqu'une telle affirmation soit appuyée par plusieurs recherches empiriques, elle peut difficilement être généralisée à l'ensemble des élèves. En effet, il existe des différences individuelles entre élèves quant à leur compréhension des échelles descriptives et quant à leur capacité à les utiliser adéquatement.

Une étude détaillée des rapports provinciaux de l'Office pour la qualité et la responsabilité en éducation (OQRE) de l'Ontario, pour la période de 1999 à 2003, révèle en effet des écarts importants dans le pourcentage des élèves qui atteignent les standards provinciaux (les niveaux 3 et 4) et le pourcentage de ceux qui répondent par « oui » à l'affirmation « Je suis bon en écriture » (Laveault & Miles, 2006). Des différences importantes ont pu être observées entre élèves de troisième et de sixième années, entre garçons et filles et entre francophones et anglophones. Les francophones tendent à se sous-estimer davantage que les anglophones, et les filles plus que les garçons. Enfin, les élèves de sixième année se sous-estiment davantage que ceux de troisième.

De telles données impliquent qu'il y a d'importantes différences quant au jugement que différentes catégories d'élèves font de leur performance à l'écrit. Ces écarts peuvent avoir des conséquences imprévisibles. Par exemple, un élève qui se surestime pourra démontrer une confiance excessive et ne pas réviser sa copie. Il se peut aussi que certains élèves soient tout simplement plus « sévères » que d'autres et se fixent des cibles plus élevées à atteindre. De tels élèves peuvent sous-estimer leur performance, peu importe leur véritable rendement. Il est donc capital d'approfondir comment de telles différences individuelles affectent le jugement de l'élève, qu'il s'agisse d'autoévaluation de sa propre production ou de celle de ses pairs.

Objectifs et questions de recherche

Notre recherche tentera d'examiner les différences interindividuelles quant au jugement que les élèves exercent lorsqu'ils utilisent des échelles descriptives et comment les différentes caractéristiques de ce jugement influencent leur rendement en écriture. En plus des indicateurs habituels de fidélité tels que la concordance des évaluations entre pairs d'un même groupe-classe, nous porterons notre attention sur trois aspects importants du jugement de l'élève : sa précision, sa sévérité et son degré de certitude ou de confiance. Enfin, nous tenterons de valider les observations en tenant compte de plusieurs variables contextuelles : le niveau scolaire de l'élève, son sexe, son groupe linguistique et s'il éprouve ou non des difficultés d'apprentissage.

Les questions de recherche sont les suivantes.

1. Dans quelle mesure les différentes caractéristiques du jugement (précision, sévérité, confiance) de l'élève sont-elles reliées à la réussite à l'écrit ?
2. Existe-t-il des différences individuelles en rapport avec le sexe, la langue ou les difficultés d'apprentissage de l'élève quant à sa capacité à évaluer des textes écrits ?

3. Existe-t-il une progression dans l'habileté à utiliser adéquatement les échelles descriptives d'évaluation de textes écrits – telle que révélée par la précision de ces évaluations – en fonction du niveau scolaire de la cinquième à la huitième année ?

Méthode

Participants

Au total, 770 élèves d'un conseil scolaire francophone (342) et de deux conseils scolaires anglophones (428) de la région d'Ottawa-Carleton ont participé à cette recherche. Comme il s'agit d'un échantillonnage pratique (*convenience sampling*), la distribution des élèves n'est pas nécessairement proportionnelle à la distribution des caractéristiques de sexe, de niveau scolaire et de langue de la population d'origine.

Instruments

Huit exercices d'évaluation de textes écrits ont été mis au point afin de déterminer le degré de précision du jugement des élèves lorsqu'ils évaluent des textes rédigés par des pairs. Ces textes ont été tirés des copies types du programme d'étude de l'Ontario (Ministère de l'Éducation de l'Ontario, 1999) et représentaient des cas typiques de chaque niveau de rendement pour chaque niveau scolaire. Ceux-ci étaient constitués de brefs paragraphes rédigés par des élèves du même niveau scolaire que le participant. Les exemples de textes écrits étaient dactylographiés tels quels et incluaient donc les erreurs d'orthographe et de grammaire de même que les erreurs structurelles caractéristiques des niveaux de performance qu'ils étaient censés représenter. Quatre exercices (un pour chaque niveau scolaire de la cinquième à la huitième année) ont été construits pour vérifier l'habileté des élèves francophones à utiliser les échelles descriptives (*rubrics*) du programme d'écriture de l'Ontario et quatre autres pour les élèves anglophones. Les textes français et anglais originaux étaient différents d'une langue à l'autre, autrement dit, il ne s'agissait pas de traductions.

Chaque exercice d'un niveau scolaire donné comprenait six textes correspondant aux quatre niveaux de rendement des échelles descriptives utilisées dans les programmes d'étude. Il y avait un texte de niveau 1, deux textes de niveau 2, deux textes de niveau 3 et un texte de niveau 4, pour un total de six textes différents à évaluer. Pour chacun des textes, l'élève devait indiquer le niveau de rendement ainsi que le degré de confiance qu'il avait en son appréciation. Chaque élève pouvait avoir recours à une copie de l'échelle descriptive d'évaluation des productions écrites du programme d'étude de l'Ontario.

Procédure

Les exercices d'évaluation ont été réalisés dans la salle de classe et ont nécessité environ 15 minutes. Les élèves ont eu pour consigne de considérer chacun des six textes comme s'il avait été composé par un élève de son niveau scolaire. L'échelle descriptive d'appréciation a été distribuée, expliquée et discutée en classe. Par la suite, les élèves ont eu pour tâche d'assigner un niveau de rendement à chacun des six textes écrits et d'indiquer leur degré de confiance dans ce jugement. Les exercices étaient administrés soit par un membre de l'équipe de chercheurs, soit par l'enseignant durant le temps de classe ; dans les deux cas cependant, l'enseignant était présent en classe pour toute la durée de l'exercice. Enfin, les enseignants ont également fourni le résultat le plus récent en écriture de chaque élève participant selon l'échelle de notation de leur conseil respectif. Afin de permettre la comparaison des notes, les notes à l'écrit ont été transposées sur une échelle unique de notation qui a préservé les propriétés ordinales des notes de chaque conseil scolaire.

Résultats

Afin de mesurer différents aspects du jugement des élèves, trois scores ont été construits à partir des réponses faites aux exercices. Ces scores visaient à mesurer la précision, la sévérité et le degré de confiance des élèves. Ils ont été mis au point et calculés selon les procédures suivantes.

1. Un score de différence D (précision). Il est calculé au moyen de la différence entre l'appréciation faite par l'élève du niveau de rendement du texte et le niveau de rendement tel qu'attribué par le programme d'études de l'Ontario. Ce score D est calculé en faisant la somme du carré des différences pour chacun des six textes. Cette méthode de calcul accorde une plus grande pondération aux plus grandes erreurs d'écart entre l'appréciation de l'élève et le niveau véritable. Plus le score D est élevé, moins l'élève comprend ou utilise correctement les échelles descriptives.
2. Un score de sévérité S. Ce score est la somme des appréciations des élèves pour les six textes. Le score maximum de 24 consisterait à accorder un niveau de rendement 4 à chacun des six textes. Le minimum possible serait 6. Pour rendre le score S directement proportionnel au concept mesuré, la somme des cotes attribuées a été soustraite de la valeur maximale de 24 pour obtenir une valeur s'étendant de 0 (aucune sévérité) à 18 (sévérité maximale). La valeur attendue de S est la somme des niveaux de rendement de chacun des textes (voir section Instruments), soit $24 – (1 + 2 + 2 + 3 + 3 + 4) = 24 – 15 = 9$.

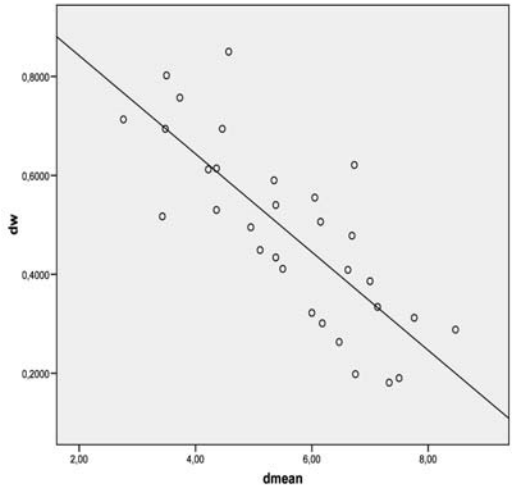
3. Un score de confiance C. Ce score est constitué de la somme des valeurs de confiance accordées à chacune des six appréciations. Une transformation similaire à celle du score S a également été utilisée pour ce score. Celui-ci s'étend également de 0 (aucune confiance sur l'ensemble des six appréciations) à 18 (confiance totale).

Facteurs affectant la qualité des appréciations dans chaque classe

Avant de procéder à l'analyse des résultats individuels, il est important de déterminer si les élèves d'une même classe interprètent uniformément les attentes de chaque niveau de rendement. Nous avons voulu déterminer quel était le degré d'accord entre élèves d'une même classe lorsque ceux-ci sont considérés comme juges de chacun des textes. Nous avons également cherché à déterminer quelles variables construites étaient associées à cet accord entre les élèves.

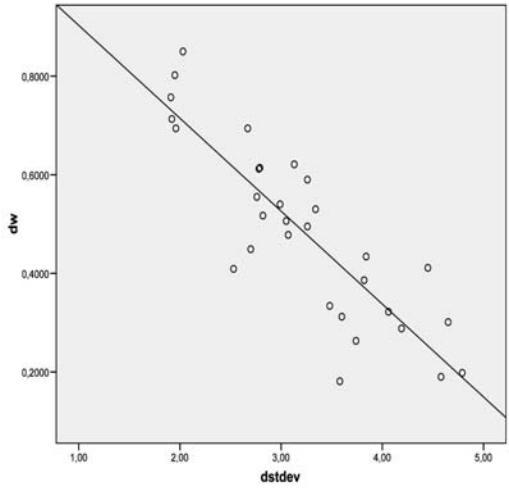
Pour ce faire, nous avons effectué des analyses de régression sur les valeurs du coefficient d'accord de chaque classe en fonction des valeurs moyennes et des écarts types des valeurs de différence D, de sévérité S et de confiance C. Les figures 1 et 2 illustrent les fonctions de régression permettant de prédire le degré d'accord entre les élèves, tel que mesuré par le W de Kendall (Siegel & Castellan, 1988), à partir des valeurs des moyennes et des écarts types des valeurs de D pour chaque groupe-classe. Les groupes de moins de 12 élèves ont été supprimés de cette analyse à cause de la faiblesse de l'échantillon ainsi qu'un groupe extrême pour lequel l'écart type des valeurs de D était considérablement plus élevé que celui des autres groupes.

La figure 1 indique que l'accord entre les élèves d'une même classe (DW) tend à diminuer lorsque la moyenne des valeurs de D (DMEAN) augmente, indiquant que plus les élèves d'une même classe se trompent dans l'appréciation du niveau d'un texte, moins ils manifestent un degré d'accord élevé entre eux dans leur estimation. ($r = -0{,}79$, $n = 31$, $p < 0{,}001$). La figure 2 confirme également que le degré d'accord des élèves est plus élevé lorsque leurs erreurs d'appréciation, représentées par le score de différence D, présentent une variance réduite ($r = -0{,}86$, $n = 31$, $p < 0{,}001$). De ces deux figures, il est possible de conclure que les élèves qui appartiennent à un groupe où la valeur moyenne de D est faible possèdent une meilleure compréhension du niveau de rendement au moyen des échelles descriptives, tendent à partager la même compréhension des exigences d'un bon texte écrit (faible écart type et faible moyenne des scores de différence) et utilisent les critères de la même manière (W de Kendall élevé ou même ordonnancement des textes en niveaux de rendement).



Légende : dw : degré de concordance entre élèves d'un même classe (W de Kendall)
dmean : moyenne du degré de précision (D)

Figure 1. ***Concordance entre élèves en fonction des valeurs moyennes de D pour chaque groupe classe***



Légende : dw : degré de concordance entre élèves d'un même classe (W de Kendall)
dstdev : écart type du degré de précision (D)

Figure 2. ***Concordance entre élèves en fonction des écarts types de D pour chaque groupe classe***

Par ailleurs, il existe deux autres corrélations significatives entre variables construites, même si elles sont de moindre importance. Une corrélation indique que plus les élèves d'une même classe sont en accord (W de Kendall élevé),

plus le degré de confiance qu'ils expriment en leurs appréciations est élevé ($r = 0{,}39$, $n = 31$, $p < 0{,}05$). Une autre corrélation, entre la précision des appréciations mesurée par le score de différence (D) moyen et le degré de sévérité moyen de chaque classe, indique que ce sont les classes les plus sévères qui réalisent les estimations les plus précises ($r = -0{,}47$, $n = 31$, $p < 0{,}01$).

Différences entre les groupes

Les élèves identifiés comme éprouvant des difficultés d'apprentissage, comme doués ou talentueux n'ont pas été inclus dans la première série d'analyses de variance. Il nous est apparu important de traiter les élèves de ces groupes séparément, d'abord à cause de leur petit nombre, ensuite, afin de ne pas confondre leurs résultats avec ceux des autres regroupements tels que le sexe, la langue et le niveau scolaire. La répartition des élèves varie considérablement en fonction des trois facteurs précédents. C'est ainsi que beaucoup plus de garçons sont identifiés comme ayant des difficultés d'apprentissage et que beaucoup plus de filles sont considérées comme «douées» ou «talentueuses». La majorité des élèves ne fait pas l'objet d'une identification particulière et sera désignée « groupe de référence ».

Élèves du groupe de référence

Le tableau 1 présente l'analyse de variance en plan factoriel complètement aléatoire pour la variable de différence D. Un seul effet principal est significatif, celui du niveau scolaire. Il confirme que la valeur de différence D dans l'appréciation du niveau de rendement de chaque texte varie en fonction du niveau de scolaire des élèves. De plus, il existe une interaction significative entre le niveau scolaire et le groupe linguistique de l'élève, indiquant que les moyennes des valeurs D par niveau scolaire ne sont pas les mêmes selon le groupe linguistique. La figure 3 illustre cette interaction. On peut y constater que les tendances des groupes francophone et anglophone sont à l'opposé l'une de l'autre. Les résultats du groupe anglophone indiquent une tendance à s'améliorer en précision de la cinquième à la huitième année (moyennes de D décroissantes) alors que celles du groupe francophone indiquent une tendance à se détériorer (moyennes de D croissantes, sauf en sixième). Une analyse des effets simples (tableau 1) indique que les différences entre les deux groupes linguistiques sont significatives pour la sixième et la huitième année. Par ailleurs, il n'y a aucun effet principal dû au sexe, ni aucune interaction significative entre le sexe des élèves et les autres facteurs. Ceci tend à confirmer que les filles sont aussi précises que les garçons dans leurs estimations du niveau de rendement, et ce, peu importe le niveau scolaire ou le groupe linguistique.

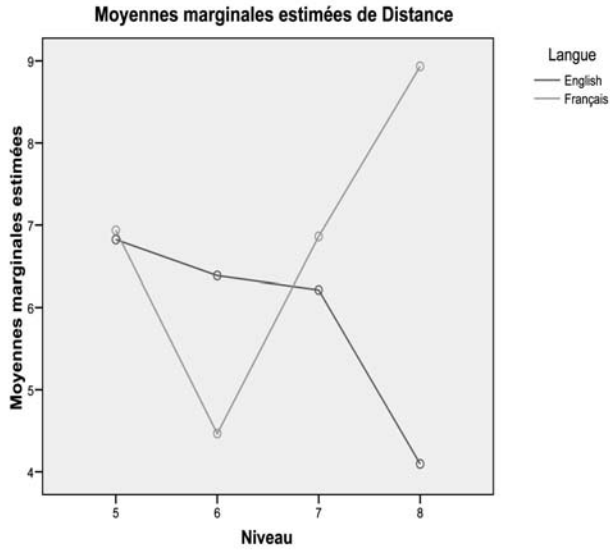


Figure 3. ***Variable D : interaction significative entre niveaux scolaires et groupes linguistiques***

Tableau 1
ANOVA (plan factoriel 4 × 2 × 2) pour la variable D

Source de variation	*SC*	*dl*	*CM*	*F*	*Sign.*
Niveau scolaire (NS)	116,16	3	38,72	2,641	*
LA en 5e	4,08	1	4,08	0,28	
LA en 6e	66,27	1	66,27	4,52	*
LA en 7e	10,46	1	10,46	0,71	
LA en 8e	228,81	1	228,81	15,61	***
Genre (GE)	3,80	1	3,80	0,26	
Langue (LA)	51,52	1	51,52	3,51	
NS × GE	56,76	3	18,92	1,29	
NS × LA	302,60	3	100,87	6,88	***
GE × LA	25,06	1	25,06	1,71	
NS × GE × LA	105,08	3	35,03	2,39	
Erreur	6 993,10	477	14,67		
Total	26 799,00	493			
Total corrigé	7 793,48	492			

Légende : Test de Levene d'homogénéité des variances, $F(15,477) = 1,51$; $p > 0,05$
* Effet significatif à 0,05
** Effet significatif à 0,01
*** Effet significatif à 0,001

Tableau 2
ANOVA (plan factoriel 4 × 2 × 2) pour la variable S

Source de variation	*SC*	*dl*	*CM*	*F*	*Sign.*
Niveau scolaire (NS)	80,76	3	26,92	5,26	***
LA en 5e	1,09	1	1,09	0,18	
LA en 6e	244,06	1	244,06	41,23	***
LA en 7e	5,10	1	5,10	0,86	
LA en 8e	15,49	1	15,49	2,62	
Genre (GE)	3,46	1	3,46	0,68	
Langue (LA)	11,35	1	11,35	2,22	
NS × GE	16,27	3	5,42	1,06	
NS × LA	206,04	3	68,68	13,41	***
GE × LA	2,25	1	2,25	0,44	
NS × GE × LA	16,14	3	5,38	1,05	
Erreur	2 442,52	477	5,12		
Total	22 342,00	493			
Total corrigé	2 773,57	492			

Légende : Test de Levene d'homogénéité des variances, $F(15,477) = 3,22$; $p < 0,001$
* Effet significatif à 0,05
** Effet significatif à 0,01
*** Effet significatif à 0,001

Le tableau 2 présente les résultats d'une analyse de variance similaire, réalisée cette fois sur les résultats de la variable sévérité (S). Tout comme pour le cas de la variable D, les différences entre niveaux scolaires constituent le seul effet principal significatif de même que l'interaction entre le niveau scolaire et le groupe linguistique. La figure 4 illustre cette interaction. Un test des effets simples confirme que c'est uniquement en sixième année que la différence entre les deux groupes est significative, le groupe des élèves anglophones étant plus sévère que le groupe des élèves francophones, ces derniers ayant tendance à surestimer les niveaux de rendement des copies évaluées. Aucune différence liée au sexe des participants n'est significative, ni aucun autre effet d'interaction.

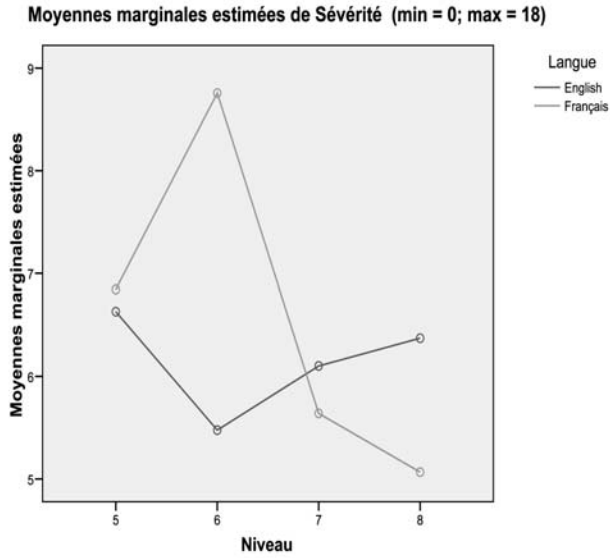


Figure 4. ***Variable S : interaction significative entre niveaux scolaires et groupes linguistiques***

Tableau 3
ANOVA (plan factoriel 4 × 2 × 2) pour la variable C

Source de variation	*SC*	*dl*	*CM*	*F*	*Sign.*
Niveau scolaire (NS)	113,98	3	37,994	6,419	***
Genre (GE)	0,30	1	0,30	0,05	
Langue (LA)	28,17	1	28,17	4,76	***
NS × GE	7,78	3	2,59	0,44	
NS × LA	24,44	3	8,15	1,38	
GE × LA	1,79	1	1,79	0,30	
NS × GE × LA	3,95	3	1,32	0,22	
Erreur	2 687,25	454			
Total	77 848,00	470			
Total corrigé	2 877,58	469			

Légende : Test de Levene d'homogénéité des variances, $F(15,454) = 1,51$; $p > 0,05$
* Effet significatif à 0,05
** Effet significatif à 0,01
*** Effet significatif à 0,001

Le tableau 3 présente les résultats de l'analyse de variance pour le niveau de confiance total exprimé dans les appréciations. Deux effets principaux se sont avérés significatifs :

1. L'effet principal du niveau scolaire. Un test de comparaisons multiples indique que cet effet est dû uniquement aux groupes de septième année qui manifestent une confiance moins grande en leurs appréciations.
2. L'effet principal du groupe linguistique. Les élèves anglophones manifestent une plus grande confiance que les élèves francophones. Ceci est confirmé par l'absence d'une interaction significative entre le niveau scolaire et le groupe linguistique. La figure 5 reproduit graphiquement ces résultats.

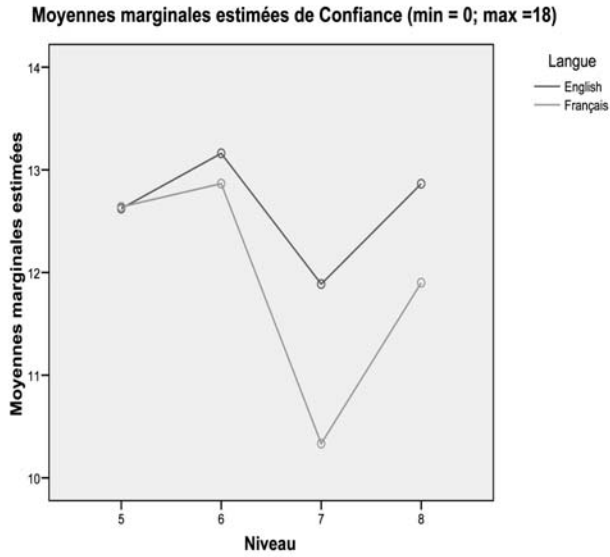


Figure 5. ***Variable C : interaction non significative entre niveaux scolaires et groupes linguistiques***

Élèves considérés comme « exceptionnels »

D'autres analyses de variance ont été effectuées sur chacune des trois variables de la recherche afin de comparer les élèves identifiés comme exceptionnels avec ceux du groupe de référence. Les élèves exceptionnels se répartissent en trois groupes : les élèves en difficulté d'apprentissage, les élèves considérés comme talentueux et ceux considérés comme doués.

Ces analyses ne prennent pas en compte le groupe linguistique, le sexe et le niveau scolaire car le nombre d'élèves par cellule est trop petit. De plus, la répartition des élèves selon ces facteurs est disproportionnée. Par exemple, les filles sont beaucoup plus présentes dans le groupe des élèves doués et les garçons, plus présents dans le groupe des élèves en difficulté d'apprentissage.

Deux des trois analyses de variance ont révélé des différences significatives entre les quatre groupes d'élèves. Aucune différence significative n'a été observée entre ces groupes d'élèves quant à la sévérité ($F(3,621) = 0,73$, $p > 0,05$). Par contre, ces catégories d'élèves ont obtenu des moyennes significativement différentes sur la variable D portant sur la précision ($F(3,621) = 7,64$, $p < 0,001$) et sur la variable C portant sur le degré de confiance ($F(3,595) = 3,71$, $p < 0,05$).

Deux tests F de Ryan-Einot-Gabriel-Welsch (REGW : Howell, 1998) de comparaisons multiples des moyennes *a posteriori* ont été effectués pour comparer les différents groupes d'élèves pour les deux variables (D et C) où un rapport F significatif avait été obtenu. Le tableau 4 rapporte les résultats du test REGW pour la variable D. Il permet de regrouper les groupes d'élèves en trois sous-ensembles. Le premier est constitué uniquement des élèves doués et se caractérise par les appréciations les plus précises (moyennes de D très faibles). Le second sous-ensemble regroupe les élèves talentueux avec les élèves du groupe de référence. Ce sous-ensemble fournit des appréciations moins précises que le sous-ensemble regroupant les élèves doués. Par contre, elles sont plus précises que celles des élèves du groupe « difficulté d'apprentissage » qui forment le troisième sous-ensemble.

Tableau 4
Test de comparaisons multiples REGW pour la variable D

F de Ryan-Einot-Gabriel-Welsch

		Sous-ensemble		
Catégorie	*N*	*1*	*2*	*3*
doué	18	3,56		
pas difficulté apprentissage	493		6,21	
talentueux	86		6,50	
difficulté d'apprentissage	28			9,11
Signification		1,000	0,781	1,000

Les moyennes des groupes formant des sous-ensembles homogènes sont affichées dans la même colonne.
Le terme d'erreur est la moyenne des carrés (erreur) = 15,881.
Alpha = 0,05.

Le tableau 5 rapporte les résultats du test REGW pour la variable C. Deux sous-ensembles ont été créés. Un premier sous-ensemble regroupe les élèves qui ont exprimé le moins de confiance : il s'agit des élèves doués, talentueux et en difficulté d'apprentissage. Le deuxième sous-ensemble regroupe les élèves les plus confiants : il s'agit des élèves qui ne présentent pas de difficultés d'apprentissage et de ceux qui en présentent. Le groupe des élèves en difficulté d'apprentissage se retrouve dans les deux sous-ensembles, car il occupe une position mitoyenne entre les groupes et ne se distingue pas suffisamment ni du groupe des élèves doués et talentueux, ni du groupe de référence.

Tableau 5
Test de comparaisons multiples REGW pour la variable C.

F de Ryan-Einot-Gabriel-Welsch

		Sous-ensemble	
Catégorie	*N*	*1*	*2*
doué	17	11,65	
talentueux	84	11,77	
difficulté d'apprentissage	28	12,07	12,07
pas difficulté apprentissage	470		12,63
Signification		0,819	0,435

Les moyennes des groupes formant des sous-ensembles homogènes sont affichées dans la même colonne.
Le terme d'erreur est la moyenne des carrés (erreur) = 6,168.
Alpha = 0,05.

Analyses de régression multiple

Pour chaque groupe linguistique, une régression multiple a été effectuée pour prédire la note en écriture à partir des trois variables construites D, S et C tout en contrôlant pour les effets du sexe. Une méthode de régression hiérarchique a été employée pour choisir l'ordre d'entrée des variables indépendantes construites, une fois contrôlé l'effet du sexe dans une première étape. Deux analyses ont été effectuées parce que la variable dépendante, la note à l'écrit, a été évaluée au moyen de systèmes différents de notation dans les conseils scolaires anglophones et dans le conseil scolaire francophone.

Tableau 6
Matrice des corrélations pour les groupes anglophone et francophone

	Corrélations(a)				
	Note en écriture	*Sexe*	*D*	*S*	*C*
Note en écriture	–	– 0,20	– 0,18	0,02	0,00
Sexe	– 0,27	–	0,11	0,00	0,02
D	– 0,31	0,05	–	– 0,34	– 0,16
S	0,08	0,05	– 0,49	–	– 0,16
C	– 0,07	0,015	– 0,06	– 0,08	–

Matrice diagonale supérieure = Anglophones (n = 352)
Matrice diagonale inférieure = Francophones (n = 141)
Les corrélations significativement différentes de 0 sont soulignées ($p < 0,05$)

Le tableau 6 présente la matrice des corrélations pour les groupes francophone (matrice diagonale inférieure) et anglophone (matrice diagonale supérieure). Les corrélations significatives à $p < 0,01$ sont soulignées. La structure des deux matrices est similaire sur plusieurs points. Dans les deux groupes, les variables indépendantes les plus corrélées avec la note à l'écrit sont les variables D et le sexe. Plus les élèves ont réalisé des appréciations précises lors des exercices d'évaluation, plus leurs notes à l'écrit sont élevées. De plus, la corrélation significative entre le sexe et la note à l'écrit indique que les filles ont tendance à se mériter de meilleures notes. Par ailleurs, la corrélation la plus élevée entre variables indépendantes se produit entre le degré de sévérité et le degré de précision. En effet, tant chez les élèves francophones qu'anglophones, plus les évaluations sont sévères, plus les estimations sont exactes. Enfin, il existe dans le groupe anglophone une corrélation significativement différente de 0 entre le degré de confiance et la sévérité, indiquant que plus l'élève est sévère, moins il est confiant dans sa réponse. Une telle corrélation n'est pas significativement différente de 0 dans le groupe francophone. De plus, un test z sur la différence entre corrélations calculées chez les francophones (– 0,08, N = 141) et chez les anglophones (– 0,16, N = 352) révèle qu'il n'y a pas de différence significative entre les deux groupes ($z = 0,805$, $p > 0,05$). L'absence de différence significative ne permet donc pas de conclure à l'existence d'une corrélation différente chez les deux groupes entre confiance et sévérité.

Tableau 7

Analyse de régression multiple pour le groupe anglophone

Récapitulatif du modèle

					Changement dans les statistiques				
Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation	Variation de R-deux	Variation de F	ddl 1	ddl 2	Modification de F signification
1	0,203[a]	0,041	0,038	1,868	0,041	15,007	1	350	0,000
2	0,259[b]	0,067	0,062	1,845	0,026	9,681	1	349	0,002

a. Valeurs prédites : (constantes), Sexe
b. Valeurs prédites : (constantes), Sexe, Distance
c. Langue = Anglais

Coefficients[a,b]

		Coefficients non standardisés		Coefficients standardisés		
Modèle		B	Erreur standard	Bêta	t	Signification
1	(constante)	4,403	0,139		31,713	0,000
	Sexe	–0,772	0,199	–0,203	–3,874	0,000
2	(constante)	4,818	0,191		25,199	0,000
	Sexe	–0,707	0,198	–0,186	–3,571	0,000
	Distance	–0,076	0,024	–0,162	–3,111	0,002

a. Variable dépendante : Note en écriture
b. Langue = Anglais

Tableau 8

Analyse de régression multiple pour le groupe francophone

Récapitulatif du modèle

					Changement dans les statistiques				
Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation	Variation de R-deux	Variation de F	ddl 1	ddl 2	Modification de F signification
1	0,272[a]	0,074	0,068	1,564	0,074	11,140	1	139	0,001
2	0,399[b]	0,159	0,147	1,496	0,085	13,957	1	138	0,000

a. Valeurs prédites : (constantes), Sexe
b. Valeurs prédites : (constantes), Sexe, Distance
c. Langue = Français

Coefficients[a,b]

		Coefficients non standardisés		Coefficients standardisés			Statistiques de colinéarité	
Modèle		B	Erreur standard	Bêta	t	Signification	Tolérance	VIF
1	(constante)	5,024	0,170		29,610	0,000		
	Sexe	–0,899	0,269	–0,272	–3,338	0,001	1,000	1,000
2	(Constante)	5,735	0,250		22,916	0,000		
	Sexe	–0,849	0,258	–0,257	–3,291	0,001	0,997	1,003
	Distance	–0,114	0,031	–0,292	–3,736	0,000	0,997	1,003

a. Variable dépendante : Note en écriture
b. Langue = Français

Les tableaux 7 et 8 présentent les résultats des régressions multiples pour chaque groupe linguistique. Les R^2 multiples sont statistiquement significatifs, mais très faibles (anglophones, $R^2 = 0{,}26$, $F(2{,}35) = 12{,}53$, $p < 0{,}001$; francophones, $R = 0{,}40$, $F(2, 138) = 13{,}07$, $p < 0{,}001$). La variable sexe, utilisée ici comme covariable, a permis de contrôler l'effet des différences entre garçons et filles en écriture. Lorsque cette variable est contrôlée, la seule variable indépendante qui contribue à accroître significativement la prédiction des résultats à l'écrit est la variable D, chez les deux groupes linguistiques. Ces résultats confirment les observations précédentes faites à partir des matrices de corrélations des deux groupes linguistiques.

Discussion et conclusion

La question à la base de cette recherche consistait à déterminer s'il existe une relation entre la capacité à interpréter et à utiliser des échelles descriptives pour l'évaluation des textes produits par des pairs et la note en écriture. Cette recherche a démontré que le score D de différence entre l'appréciation du niveau de rendement d'une production écrite et son niveau réel est le meilleur prédicteur du rendement en écriture d'un élève parmi les trois variables étudiées. Le score de sévérité S et le score de confiance C n'ajoutent rien à la prédiction du résultat en écriture à ce qui est déjà expliqué par la variable D. Ceci étant dit, les résultats ont également démontré que le score D diminue lorsque le degré de sévérité de l'élève augmente. Ce résultat s'explique par la tendance générale des élèves à errer dans le sens de la surestimation. En fait, il serait plus juste de dire que le score D diminue lorsque la sévérité se rapproche d'une valeur optimale. Sévérité et précision sont par conséquent liées, les élèves trop ou pas assez sévères ayant tendance à errer davantage dans leur appréciation.

Si la relation entre la capacité à évaluer avec précision et le résultat à l'écrit est significative, l'effet expérimental tel que révélé par les corrélations (simple et multiple) est plutôt faible. De fait, il se peut que les relations entre ces variables soient plus fortes que ce que les résultats obtenus laissent entendre. Il est en effet difficile de croire que les exercices d'évaluation représentent des tâches équivalentes pour chaque groupe linguistique et pour chaque niveau scolaire. L'interaction entre groupe linguistique et niveau scolaire observée entre les variables D et S le confirme. Les différences probables entre exercices en français et en anglais peuvent avoir contribué à diminuer la valeur de l'effet expérimental et ont rendu plus difficile une réponse précise à la question de recherche #3 sur la progression du jugement des élèves de la cinquième à la huitième année.

On peut également s'interroger sur l'existence d'une relation de cause à effet entre la précision et la réussite à l'écrit. La précision dépend-elle de la réussite à l'écrit ou est-ce la réussite à l'écrit qui dépend de la précision ? Les analyses effectuées ne permettent pas de répondre directement à cette question, mais laissent toutefois entrevoir une réponse possible.

Il semble probable que la relation entre la précision et la réussite à l'écrit se caractérise par une implication réciproque. Sans une compréhension adéquate des textes à évaluer, comment en effet fournir une appréciation précise ? Par ailleurs, comment fournir une appréciation précise sans un certain approfondissement des critères d'évaluation ? Ceci nous ramène à la distinction faite par Orsmond et son équipe (2000) entre catégories d'élèves : certains s'évaluent avec difficulté parce qu'ils connaissent mal les critères et d'autres parce qu'ils connaissent mal la matière, parfois les deux. L'analyse des corrélations entre le degré d'accord et la précision des jugements d'évaluation a tendance à confirmer que les classes où les élèves connaissent et utilisent bien les critères ont tendance également à ordonner les textes de la même façon, ce qui indiquerait une compréhension plus homogène des critères. Les classes où les élèves manquent de précision sont celles pour lesquelles il y a de grandes différences dans le classement des copies. Les résultats confirment que l'enseignant peut, en améliorant la précision des évaluations des élèves, favoriser le développement de certains consensus dans l'application des critères par les élèves d'une même classe et, de là, contribuer à la réussite à l'écrit. Ceci rejoint les observations de Stiggins et al. (2004) sur l'importance d'apprendre à évaluer (« assessment as learning »). Enfin, le lien entre la précision des évaluations et des variables telles que la confiance et la sévérité laisse entrevoir plusieurs implications quant à la façon dont l'enseignant peut s'acquitter de sa responsabilité de planifier l'évaluation en salle de classe.

Ceci nous amène à répondre à la deuxième question de recherche sur l'existence de différences individuelles en rapport avec le sexe, la langue ou les difficultés d'apprentissage. Là encore, une certaine prudence s'impose. La puissance des tests de signification effectués n'est pas toujours très grande, ce qui accroît le risque de déclarer non significatifs des effets qui existent en réalité. Ceci étant dit, aucune des analyses de variance effectuées sur les trois variables ne rapporte d'effets significatifs en rapport avec le sexe, ni aucune interaction significative du sexe avec le niveau scolaire ou le groupe linguistique. Il est possible de conclure que, même si les filles obtiennent de meilleures notes que les garçons à l'écrit, les différences entre les deux quant à la précision, la sévérité et la confiance ne sont pas assez grandes pour être considérées comme significatives, et ce, peu importe la langue ou le niveau scolaire.

La comparaison des groupes francophone et anglophone au moyen des régressions multiples séparées appuie l'hypothèse selon laquelle la compréhension des échelles descriptives contribue de la même manière à l'apprentissage de l'écriture dans chacun des groupes. Le degré de confiance, toutefois, s'est révélé significativement plus bas chez les francophones et ceux-ci, quoiqu'aussi précis que leurs pairs anglophones, sont moins confiants. De faibles niveaux de confiance chez les francophones peuvent être attribués à des facteurs de perceptions personnelles associés au vécu de groupe minoritaire des francophones de l'Ontario. En cela, nos résultats confirment ce qu'il a été possible d'observer année après année dans les résultats des tests provinciaux de l'Ontario (Laveault & Miles, 2006). Lorsque questionnés sur leur habileté à l'écrit, les francophones se jugent beaucoup moins bons qu'ils ne le sont en réalité.

L'étude a également permis de confirmer l'existence de différences importantes quant à l'habileté avec laquelle les élèves des groupes talentueux, doués ou en difficulté d'apprentissage utilisent les échelles descriptives. Les élèves identifiés comme doués sont les plus précis dans leur estimation et les élèves en difficulté d'apprentissage sont les moins précis. Par contre, rien ne distingue les quatre groupes quant à la sévérité. Quant au degré de confiance exprimé, ce sont les élèves doués qui expriment le moins de confiance, ce en quoi ils ne se distinguent pas des élèves talentueux ou en difficulté d'apprentissage. Il semble ici que la confiance exprimée par les élèves doués sous-tend une réalité différente de celle des élèves en difficulté d'apprentissage. On peut émettre l'hypothèse que les élèves doués sont plus conscients de la difficulté des exercices, ce qui se refléterait par un effet de modération sur la confiance. Chez les élèves en difficulté d'apprentissage, le manque de confiance serait davantage lié à l'expérience personnelle de difficultés persistantes. Ces deux cas typiques soulèvent toute la délicate question de l'importance d'évaluer l'élève sans nuire globalement à son degré de confiance.

Quant à la réponse à la troisième question de recherche, les résultats ne permettent pas de réponse précise. Il y a bien des effets significatifs du niveau scolaire pour les trois variables de la recherche, mais l'interprétation des résultats ne permet pas de dégager une véritable tendance qui se serait traduite dans notre recherche par une diminution progressive de la valeur de D (précision plus grande) de la cinquième à la huitième année. À cause d'interactions significatives entre l'appartenance au groupe linguistique et le niveau scolaire pour les variables de précision et de sévérité, il est difficile de déterminer quelle part de ces interactions est le résultat d'une variation dans le degré de difficulté des exercices pour chaque niveau scolaire en français et en anglais et

quelle part est le résultat d'une véritable amélioration dans l'habileté à évaluer entre niveaux scolaires. Par contre, les résultats sont beaucoup plus facilement interprétables pour la variable confiance puisqu'une telle interaction significative ne s'y retrouve pas. Ceci peut être dû au fait que la confiance exprimée est indépendante de la difficulté des exercices ou, du moins, qu'elle est influencée par celle-ci dans une moindre mesure que la précision ou la sévérité. Ces résultats révèlent une plus grande confiance chez les étudiants anglophones et un effet principal dû au niveau scolaire identique pour les deux groupes linguistiques, qui se traduit par une diminution significative en septième année. Cette diminution de la confiance rejoint les résultats des recherches sur ce groupe scolaire, à la charnière du passage du primaire au secondaire.

Portée et limites de la recherche

Les résultats indiquent que certaines sources de variance ont pu être confondues, limitant ainsi l'estimation de l'effet expérimental et la puissance des tests statistiques. Ceci résulte de la combinaison de plusieurs facteurs, tels que, par exemple, des textes nécessairement différents pour francophones et anglophones ainsi que des différences inévitables entre écoles et conseils scolaires dans les échelles de notes. Le recours à un examen uniforme en écriture permettrait sans doute d'obtenir des effets expérimentaux plus élevés. Ceci étant dit, le recours à des méthodes statistiques a permis de contrôler *a posteriori* certains de ces effets confondus en utilisant le sexe comme covariable dans les analyses de régression. Les analyses de régression ont été conduites séparément pour les élèves francophones et anglophones dans un but similaire.

En réalisant la même étude sur deux groupes linguistiques et avec des exercices différents, quoique jugés équivalents, il a été possible de valider les observations de façon croisée. Même si cette condition de la recherche a pu contribuer à réduire la validité interne du devis et à sous-estimer l'effet expérimental, elle permet par contre d'accroître la généralisation des résultats observés à des exercices et à des groupes différents. La portée des conclusions et la validité externe en sont accrues. Il est donc possible d'affirmer que l'association établie entre la précision de l'évaluation et la réussite en écriture est généralisable aux deux groupes linguistiques et à tous les niveaux scolaires de la cinquième à la huitième année de notre échantillon.

Recommandations

Ces résultats permettent de faire un certain nombre de recommandations quant à l'utilisation d'échelles descriptives en salle de classe pour l'évaluation par des pairs ou encore pour l'autoévaluation. Premièrement, il semble qu'un entraînement à une utilisation compétente de ces échelles est possible et même souhaitable. Les élèves qui savent mieux que les autres ce que l'on attend d'eux et ce qui différencie un bon texte d'un moins bon performent mieux. On ne peut tenir pour acquis que les élèves savent nécessairement utiliser une échelle sans qu'il y ait eu un certain apprentissage préalable. Ceci correspond à l'apprentissage de l'évaluation (*assessment as learning*) décrit par Stiggins et al. (2004). Il est donc nécessaire d'encourager les enseignants à prendre du temps pour expliquer ces échelles descriptives aux élèves et faire la démonstration de leur pertinence.

Par ailleurs, l'observation voulant qu'il existe un lien entre le degré de sévérité et la précision de l'évaluation de productions écrites par des pairs devrait sensibiliser les enseignants à la dynamique particulière qui peut exister entre certains élèves. Si des élèves bons en écriture peuvent fournir une évaluation plus précise des productions écrites de leurs pairs, celle-ci risque également d'être plus sévère. Les enseignants devraient prendre en considération non seulement l'impact cognitif, mais aussi affectif et émotionnel de l'évaluation par des pairs. En effet, comme l'a révélé la recherche de McCurdy et Shapiro (1992), l'ensemble des élèves de leur étude préférait recevoir un renforcement ou des encouragements de la part de l'enseignant plutôt que de la part des pairs. Ces auteurs en ont conclu que le rôle de l'enseignant demeure indispensable pour motiver les élèves.

L'importance de prendre en compte les aspects affectif et émotionnel de l'évaluation ressort également des résultats obtenus quant au degré de confiance. Les francophones en milieu minoritaire et les élèves en difficulté ont manifesté des degrés de confiance plus faibles. Dans le cas des francophones, même si ceux-ci se sont révélés ni plus ni moins précis dans leurs estimations, leur degré de confiance est apparu significativement plus bas que celui des anglophones, et ce, pour trois des quatre niveaux scolaires étudiés. Il est naturellement facile d'attribuer cette différence à l'expérience minoritaire des élèves francophones de l'Ontario. Il serait cependant plus utile de pouvoir relier ce manque de confiance à des aspects précis de leur condition minoritaire à l'origine de ce résultat.

Les nouveaux programmes d'études, que ce soit au Québec, en Ontario ou à l'extérieur du Canada, mettent l'accent sur la réalisation de tâches de plus en plus complexes. La planification et l'observation continue de toutes les étapes de réalisation de ces tâches complexes nécessitent une capacité à évaluer son travail qui n'est pas forcément acquise ou présente chez tous les élèves. Cette recherche a permis de faire ressortir les impacts cognitifs, mais aussi émotionnels de l'évaluation. Par exemple, il peut s'avérer utile de comprendre pourquoi certaines catégories d'élèves manifestent moins de confiance en leur évaluation. Ce type d'observation, ainsi que celles qu'il nous a été possible de faire ressortir sur la sévérité, confirment la pertinence de tenir compte non seulement des dimensions cognitives, mais également socioaffectives de l'utilisation d'échelles descriptives, que ce soit pour l'évaluation de la production par des pairs ou pour l'autoévaluation.

RÉFÉRENCES

Allal, L. (1999). Impliquer l'apprenant dans le processus d'évaluation : promesses et pièges de l'autoévaluation. In C. Depover & B. Noël (éds), *L'évaluation des compétences et des processus cognitifs*. Bruxelles : DeBoeck-Université.

Allal, L, (2007). Régulations des apprentissages : orientations conceptuelles pour la recherche et la pratique en éducation. In L. Allal & L. Mottier Lopez (éds), *Régulation des apprentissages en situation scolaire et en formation* (pp. 7-23). Bruxelles : De Boeck.

Beaman, R. (1998). The unquiet… even loud andragogy. Alternative assessments for adult learners. *Innovative Higher Education, 23*(1), 47-59.

Bergee, M.J. (1997). Relationships among faculty, peer and self-evaluations of applied performances. *Journal of Research in Music Education, 45*(4), 601-612.

Bishop, N. (1994). Grading actors. A student evaluation based on effort and improvement. *Teaching Theatre, 5*(2), 3-7.

Bressoux, P., & Pansu, P. (2003). *Quand les enseignants jugent leurs élèves.* France : PUF.

Church, A.H. (1997). Do you see what I see ? An exploration of congruence in ratings from multiple perspectives. *Journal of Applied Social Psychology, 27*(11), 983-1020.

Fallows, S., & Chandramohan, B. (2001). Multiple approaches to assessment : reflections on use of tutor, peer and self-assessment. *Teaching in Higher Education, 6*(2)*,* 229-245.

Howell, D.C. (1998). *Méthodes statistiques en sciences humaines.* Bruxelles : ITP/De Boeck.

Laveault, D. (2007). De la régulation au réglage : étude des dispositifs d'évaluation favorisant l'autorégulation des apprentissages. In L. Allal & L. Mottier Lopez (éds), *Régulation des apprentissages en situation scolaire et en formation* (pp. 207-234). Bruxelles : De Boeck.

Laveault, D., & Miles, C. (2006). *Educational indicators of two Language Arts curricula conceptual equivalence*. Bruxelles : International Test Commission.

Liu, E.Z., Lin, S.S.J., & Yuan, S.M. (2002). Alternatives to instructor assessment : A case study of comparing self and peer assessment with instructor assessment under a networked innovative assessment procedures. *International Journal of Instructional Media, 29*(4), 393-404.

McCurdy, B.L., & Shapiro, E.S. (1992). A comparison of teacher-, peer- and self-monitoring with curriculum-based measurement in reading among students with learning difficulties. *The Journal of Special Education, 26*(2), 162-180.

Miller, P.J. (2003). The effect of scoring criteria specificity on peer and self-assessment. *Assessment & Evaluation in Higher Education, 28*(4), 383-394.

Ministère de l'Éducation de l'Ontario (1997). *Le curriculum de l'Ontario de la 1re à la 8e année – Français*. Toronto : Auteur.

Ministère de l'Éducation de l'Ontario (1999). *Le curriculum de l'Ontario : Copies types de la 1re à la 8e année – Écriture*. Toronto : Auteur.

Ministère de l'Éducation du Québec (2001). *Programme de formation de l'école québécoise. Éducation préscolaire et enseignement primaire*. Québec : Auteur.

Ministère de l'Éducation du Québec (2003). *Politique d'évaluation des apprentissages*. Québec : Auteur.

Orsmond, P., Merry, S., & Reiling, K. (1997). A study in self-assessment : tutor and students' perceptions of performance criteria. *Assessment & Evaluation in Higher Education, 22*(4), 357-369.

Orsmond, P., Merry, S., & Reiling, K. (2000). The use of student derived marking criteria in peer and self-assessment. *Assessment & Evaluation in Higher Education, 25*(1), 23-38.

Orsmond, P., Merry, S., & Reiling, K. (2002). The use of exemplars and formative feedback when using student derived marking criteria in peer and self-assessment. *Assessment & Evaluation in Higher Education, 27*(4), 309-323.

Siegel, S., & Castellan, N.J. (1988). *Nonparametric Statistics for the Behavioral Sciences* (2e édition). New York : McGraw-Hill.

Somervell, H. (1993). Issues in assessment, enterprise and higher education : the case for self-, peer and collaborative assessment. *Assessment & Evaluation in Higher Education, 18*(3), 221-233.

Stiggins, R.J., Arter, J.A., Chappuis, J., & Chappuis, S. (2004). *Classroom Assessment for Student Learning. Doing it right – Using it Well.* Portland (Oregon) : Assessment Training Institute.