Les sciences perdent-elles du sens en traduction ? Les effets de la langue sur les évaluations des matières du groupe Sciences du Programme du diplôme du Baccalauréat International

Résumé

Joshua McGrane, Yasmine El Masri, Heather Kayton, Robert Woore, & Kit Double

Oxford University Centre for Educational Assessment (OUCEA)

Département d'éducation, Université d'Oxford

Résumé

Contexte

Les programmes du Baccalauréat International (IB) sont proposés dans 153 pays du monde entier. Bien que l'anglais soit la langue d'enseignement de la majorité des écoles du monde de l'IB, l'espagnol et le français sont utilisés par de nombreux établissements pour enseigner les programmes de l'IB. Les évaluations du Programme du diplôme de l'IB ont donc lieu en anglais, en espagnol et en français, et certaines sont aussi proposées dans plus de 75 langues. La complexité du processus de traduction des évaluations et l'engagement de l'IB à proposer des évaluations plurilingues de qualité équivalente dans toutes les langues soulèvent des questions essentielles. Ces questions portent sur la comparabilité des évaluations en anglais, en espagnol et en français sur le plan linguistique et, par extension, sur le plan des exigences cognitives. Elles portent aussi, de manière plus générale, sur leur comparabilité en matière de difficulté empirique des éléments. Ces questions sont importantes parce qu'elles ont un rapport direct avec le degré de comparabilité des notes de l'IB. L'équivalence des notes entre les versions dans les différentes langues présente des enjeux considérables pour l'équité en matière d'accès à l'enseignement supérieur, particulièrement pour un diplôme international comme celui de l'IB. Cette étude, qui constitue une première recherche sur le sujet, examine la comparabilité des versions anglaise, espagnole et française des évaluations de 2019 des matières du groupe Sciences du Programme du diplôme.

Portée et objectifs

Les objectifs généraux suivants ont guidé l'approche et les méthodes utilisées pour cette étude ainsi que les conclusions qui en ont été tirées.

- Examiner les tendances et les évolutions dans les différences observées entre les résultats obtenus par les élèves en réponse à la version source (anglais) et aux versions cibles (espagnol et français) des questions des examens de 2019 du groupe Sciences du Programme du diplôme pour évaluer dans quelle mesure la difficulté des questions variait d'une langue à l'autre.
- Étudier la mesure dans laquelle ces différences observées sont dues à la traduction des questions en espagnol et en français, qui a induit des changements en matière de linguistique et d'exigences cognitives.
- Développer un modèle pour expliquer les différences observées concernant un ensemble d'effets de traduction, d'effets de langue et d'effets sans rapport avec la traduction à l'aide de méthodes qualitatives et quantitatives.

- Proposer des améliorations aux processus de traduction de l'IB, en fonction des conclusions de l'étude.

Cette étude s'est déroulée en trois phases.

- Au cours de la première phase, les chercheurs ont utilisé des techniques quantitatives pour analyser les données des examens du groupe Sciences du Programme du diplôme de 2019 afin d'évaluer si des différences systématiques apparaissaient dans le niveau d'exigence des questions, représenté par la difficulté empirique de ces dernières, et pour évaluer l'amplitude de ces différences et déterminer si la langue source ou la langue cible est avantagée.
- La deuxième phase s'est appuyée sur la première dans la mesure où des réviseurs bilingues experts ont évalué un sous-ensemble des questions déterminées précédemment, dont le niveau d'exigence montrait des différences systématiques, pour évaluer si les versions des questions dans la langue source et dans les langues cibles présentaient des différences en fonction de critères linguistiques et de traduction clés.
- Au cours de la troisième phase, les chercheurs ont utilisé les conclusions tirées des deux premières phases pour développer un modèle explicatif visant à évaluer si les différences linguistiques et de traduction entre les versions des questions dans la langue source et dans les langues cibles étaient étroitement associées aux différences du niveau d'exigence dans les trois versions.

Approche méthodologique

L'étude a fait appel à des méthodes de pointe durant les trois phases.

- Au cours de la première phase, les chercheurs ont utilisé un domaine de la modélisation psychométrique appelé « théorie de la réponse aux éléments », et plus particulièrement le modèle logit multinomial à coefficients aléatoires, pour évaluer les différences du niveau d'exigence des différentes questions. Ils ont examiné les données de réponses des élèves ayant passé les évaluations de 2019 des matières du groupe Sciences du Programme du diplôme en anglais, en espagnol et en français, notamment les épreuves de physique au niveau moyen (NM) et au niveau supérieur (NS), de chimie (NM et NS) et de biologie (NM et NS). Les chercheurs ont analysé séparément chaque combinaison matière/niveau ainsi que la langue source (l'anglais) par rapport à la langue cible (l'espagnol d'une part et le français d'autre part). Ils ont donc effectué au total 12 analyses. Pour étudier les différences du niveau d'exigence, les chercheurs ont utilisé la technique du fonctionnement différentiel des éléments (*Differential Item Functioning*, ou DIF). Cette technique leur a permis d'effectuer des comparaisons statistiques entre les résultats des élèves des différents groupes de langues pour quantifier ces différences entre les versions anglaise, espagnole et française des questions. Trois modèles particuliers étaient adaptés aux

données de réponse : un modèle qui supposait l'absence de fonctionnement différentiel des éléments entre les groupes de langues et deux modèles qui supposaient l'existence de ce fonctionnement en intégrant un paramètre spécifique au groupe de langues ainsi qu'un facteur d'interaction entre le groupe et le paramètre du modèle de difficulté de l'élément (c'est-à-dire, la question). Dans les cas où le modèle avec fonctionnement différentiel des éléments correspondait mieux aux données de réponse, le facteur d'interaction a fourni une estimation de l'amplitude du fonctionnement différentiel des éléments (fonctionnement « nul », « faible », « modéré » ou « important ») au niveau de la question et indiqué si la langue source ou les langues cibles étaient avantagées. Les chercheurs ont ensuite comparé les estimations de fonctionnement différentiel des éléments avec d'autres propriétés psychométriques des questions pour chaque matière du Programme du diplôme examinée dans le cadre de cette étude, et un sous-ensemble de questions a été retenu pour les deux phases suivantes.

- Au cours de la deuxième phase, des experts ont effectué un examen qualitatif des questions dont le fonctionnement différentiel des éléments avait été déterminé comme « faible », « modéré » ou « important » lors de la phase précédente. Examiner toutes ces questions aurait exigé trop de ressources. C'est pourquoi les chercheurs ont sélectionné un sous-ensemble de questions dans trois matières (physique NM, chimie NS et biologie NM) en se basant sur plusieurs critères, notamment l'intégration de différents types de questions possibles (à choix multiple, à réponse construite) et un assemblage équilibré de questions avantageant et désavantageant le groupe de la langue source. Dix réviseurs bilingues et trilingues experts ont été recrutés en collaboration avec l'IB pour évaluer la comparabilité des versions traduites des questions sélectionnées avec leur version en anglais (deux dans chaque combinaison langue/matière). Les experts ont examiné les questions à l'aide d'une nouvelle enquête en 14 à 15 points développée par les chercheurs à partir d'un cadre de traduction et de vérification reconnu. Ces points couvraient huit critères de ce cadre ainsi que les processus internes de l'IB en matière d'exactitude des traductions entre la langue source et les langues cibles. La fiabilité interévaluateurs a été calculée pour les jugements des experts, et les réponses ont été rassemblées afin d'examiner les questions des évaluations de sciences du Programme du diplôme pour lesquelles un fonctionnement différentiel des éléments selon la langue avait été établi lors de la première phase. Le but était d'évaluer si ces questions faisaient ressortir des différences linguistiques et de traduction cohérentes avec ce fonctionnement différentiel. De plus, les variables réunies par les experts ont été utilisées lors de la troisième phase de l'étude pour contribuer au développement d'un modèle explicatif du fonctionnement différentiel des éléments.

- Au cours de la troisième phase, les chercheurs ont construit un modèle de fonctionnement différentiel des éléments selon la langue pour les trois matières étudiées lors de la deuxième phase. Ils ont procédé en deux étapes. Durant la première étape, la modélisation a porté uniquement sur le sous-ensemble d'éléments créé lors de la deuxième phase, de manière à pouvoir intégrer les variables réunies par les experts dans le modèle. Outre ces variables, la troisième phase a aussi intégré des indices pour les questions, basés sur le traitement du langage naturel (un domaine de la linguistique informatique), ainsi que des caractéristiques non linguistiques telles que la matière, l'épreuve (en tant que substitut du type d'élément) et la langue cible de la question. Les chercheurs ont calculé les indices de traitement du langage naturel pour chaque combinaison de matière, langue et question à l'aide de ReaderBench, un logiciel libre de traitement de textes multilingues. Des recherches précédentes sur ces indices ont montré qu'ils sont associés à la complexité textuelle et, donc, les chercheurs s'attendaient à ce que les différences entre ces indices dans les versions des questions dans la langue source et dans les langues cibles expliquent le fonctionnement différentiel des éléments selon la langue. Durant la seconde étape, toutes les questions pour lesquelles il existait une estimation de fonctionnement différentiel des éléments dans les trois matières ont été intégrées. Les variables réunies par les experts ont donc été retirées du modèle, et les travaux se sont concentrés sur la capacité d'explication des indices de traitement du langage naturel. L'approche de modélisation quantitative utilisée lors de cette troisième phase exigeait des quantités importantes de données afin de produire des estimations solides. Pour cette raison, l'analyse effectuée lors de cette deuxième étape a aussi intégré des questions de 2018 ainsi que leurs estimations de fonctionnement différentiel des éléments pour les trois mêmes matières. Les modèles explicatifs utilisés lors de cette phase viennent de l'apprentissage automatique. Trois modèles ont été appliqués (régression séquentielle, filet élastique et forêt aléatoire), car chacun a des avantages et des inconvénients. Ils peuvent permettre une interprétation plus transparente (régression séquentielle, filet élastique) ou être plus opaques, mais plus flexibles en matière d'interactions non linéaires et complexes entre les variables des modèles. Pour les deux étapes, les chercheurs ont commencé par évaluer les modèles en fonction de leur capacité d'explication et de leurs erreurs de prévision ; ils ont ensuite évalué le modèle présentant les meilleurs résultats en fonction des variables spécifiques qui étaient les plus importantes pour expliquer le fonctionnement différentiel des éléments selon la langue. Dans tous les cas, les modèles ont été appliqués en utilisant une approche de validation croisée pour améliorer la généralisation des résultats.

Principales conclusions

Les principales conclusions de la première phase de l'étude sont les suivantes.

- Les analyses ont montré qu'un modèle de fonctionnement différentiel des éléments convenait mieux que les autres dans toutes les combinaisons matière/langue, ce qui a prouvé que tous les examens du groupe Sciences du Programme du diplôme dans les trois langues présentaient un fonctionnement différentiel des éléments basé sur la langue.
- De manière plus positive pour le processus de traduction actuel de l'IB, seule une proportion faible, mais non négligeable des éléments présentait un fonctionnement différentiel des éléments « modéré » et « important » pour toutes les matières, et le fonctionnement différentiel le plus important avait tendance à être davantage répandu dans les éléments de réponses construites des épreuves 2 et 3. Dans l'ensemble, les épreuves de chimie comptaient la proportion la plus élevée de fonctionnement différentiel « modéré » et « important » au niveau de l'élément suivies par les épreuves de physique puis de biologie.
- La tendance générale indique que les questions ayant un fonctionnement différentiel important à l'avantage des langues cibles étaient enclines à être les éléments plus difficiles et les moins discriminants, ce qui était particulièrement le cas pour les questions à choix multiples. La relation entre les estimations de fonctionnement différentiel des éléments et les autres propriétés psychométriques des éléments a fourni des preuves montrant que certains fonctionnements différentiels dans les langues peuvent être dus à des problèmes généraux d'ajustement aux éléments plutôt qu'aux langues en tant que telles. En particulier, certains de ces fonctionnements différentiels peuvent être imputables au comportement de supposition (consistant à essayer de deviner les réponses), d'autant plus que les élèves ayant répondu dans les langues cibles avaient tendance à être, en moyenne, moins bons dans les différentes matières.
- Les questions des épreuves de physique NM ont été sélectionnées dans les autres phases de l'étude, car la taille de l'échantillonnage pour les épreuves de physique NS en français était très petite. Les estimations du fonctionnement différentiel des éléments pour la physique NM étaient donc plus fiables, malgré une amplitude généralement plus faible. Les questions des épreuves de chimie NS et de biologie NM ont elles aussi été sélectionnées, car l'amplitude des estimations pour ces deux matières était généralement plus grande que pour la chimie NM et la biologie NS.

Les principales conclusions de la deuxième phase de l'étude sont les suivantes.

- Les conclusions des experts et l'examen qualitatif des questions étaient très positifs pour le modèle de traduction actuel adopté par l'IB puisque, selon ces conclusions et cet examen, la majorité des versions cibles espagnole et française des questions étaient hautement comparables à leur version source anglaise. Certaines incohérences sont apparues dans des questions

spécifiques, mais ces incohérences étaient généralement mineures et non systématiques pour ce qui était de l'amplitude des estimations ou du groupe avantagé. Par exemple, les épreuves de chimie NS comptaient beaucoup plus de questions classées dans la catégorie des questions à fonctionnement différentiel des éléments « modéré » et « important », mais selon le jugement des réviseurs experts, les versions traduites de ces éléments étaient plus comparables à la version anglaise.

- Certains critères utilisés par les experts ont permis de montrer un écart entre les versions source et cible des questions. Le degré d'écart le plus constant et le plus élevé a été observé pour les critères « correspondances et structures » (ou « matpat », pour *matches and patterns*) et « exactitude de la formulation » (ou « word » pour *accuracy of wording)*, mais ces écarts avaient toujours tendance à être faibles en valeur absolue.
- Dans l'ensemble, les réviseurs experts ont pu utiliser de manière fiable la nouvelle enquête développée par les chercheurs pour évaluer les différences potentielles entre les versions source et cible des éléments. Ces résultats positifs en matière de fiabilité ont créé la confiance nécessaire pour que ces variables puissent être utilisées dans la modélisation effectuée lors de la troisième phase. Certains critères ont néanmoins fait ressortir une fiabilité systématiquement plus faible que le seuil de concordance de 70 % entre les matières et les langues, notamment les critères « formulation » et « longueur des propositions ». Il conviendra donc d'essayer de normaliser davantage la compréhension de leur signification pour les futures utilisations de cette enquête.

Les principales conclusions de la troisième phase de l'étude sont les suivantes.

- La façon dont les différences linguistiques et de traduction entre les versions des questions dans la langue source et dans les langues cibles expliquent les différences de difficulté entre les versions de ces questions dans les trois langues a conduit à des conclusions variées. Ainsi, aucune des variables axées sur la langue issues de l'examen réalisé par les experts lors de la deuxième phase ne s'est avérée être un indicateur solide du fonctionnement différentiel des éléments selon la langue, ce qui était toutefois cohérent avec les constatations descriptives effectuées pour ces variables lors de cette deuxième phase. Le manque de variation dans ces variables réunies par les experts explique probablement cette situation, au moins en partie.
- Dans une certaine mesure, les différences dans les indices de complexité textuelle du traitement du langage naturel entre les versions des éléments dans la langue source et dans les langues cibles ont permis d'expliquer les différents niveaux de fonctionnement différentiel selon la langue observés dans les éléments. Le modèle de forêt aléatoire, qui convenait le mieux pour les deux étapes, a donné de meilleurs résultats pour les petits sous-ensembles de la deuxième phase : il a représenté 11 % de la variance dans la variable de résultat du fonctionnement différentiel des

éléments selon la langue par rapport à 4 % pour les ensembles de données plus grands, qui comprenaient des éléments de 2019 et de 2018 pour les trois matières (physique NM, chimie NS et biologie NM).

- Les caractéristiques du traitement du langage naturel les plus importantes pour la prévision de la variable de résultat du fonctionnement différentiel des éléments selon la langue, issues du modèle de forêt aléatoire, ont pu être organisées en trois thèmes suivant leur ordre général d'importance dans le modèle : choix lexical, longueur des phrases et complexité structurelle.
    - Les indices de choix lexical représentent différents aspects de la façon dont les informations nouvelles ou mal connues contenues dans le texte peuvent poser des difficultés au lecteur, quelle que soit la langue. Plus une phrase est prévue ou prévisible pour le lecteur, plus cette phrase est facile à comprendre. Ces informations peuvent être des mots, des lettres, des phrases, voire des éléments de ponctuation.
    - Les indices de longueur de phrase font ressortir différents aspects de la manière dont, à mesure que la longueur d'une phrase augmente, la charge cognitive associée au traitement de cette phrase augmente et mettent en évidence la façon dont cela peut affecter la compréhension de la phrase par les lecteurs.
    - Les indices de complexité structurelle montrent comment différentes caractéristiques grammaticales et syntaxiques d'un texte peuvent se manifester à différents niveaux de complexité pour le lecteur.

Recommandations

Chaque phase de l'étude a donné lieu à des recommandations spécifiques. Les recommandations générales de la première phase sont notamment les suivantes.

- Examiner des éléments à choix multiple présentant des taux différentiels en matière de supposition entre les versions dans les diverses langues pour comprendre quelles caractéristiques de ces éléments peuvent entraîner une augmentation des suppositions, en général et dans une langue particulière.
- Examiner les éléments présentant un fonctionnement différentiel selon la langue « moyen » et « important » pour les trois autres matières qui n'ont pas été retenues dans les deux dernières phases de l'étude et pour les examens d'autres années.

Les recommandations générales de la deuxième phase sont notamment les suivantes.

- Garantir la normalisation des processus de traduction et d'assurance de la qualité dans les matières et entre les matières.
- Ouvrir l'évaluation en réduisant sa dépendance à la culture et à la langue. Il serait possible d'y parvenir en créant l'évaluation en deux langues sources (par exemple, en anglais et en espagnol)

puis en utilisant ces deux versions sources pour créer la version de l'évaluation en langue cible (par exemple, en français).

- Réfléchir à des procédures de révision des traductions et/ou d'assurance de la qualité qui permettent de résoudre les problèmes trouvés dans la version cible ou de vérifier s'ils se retrouvent dans la version source. Il se peut, par exemple, qu'un problème trouvé dans la version cible concerne aussi la version source et exige que les deux versions soient corrigées.
- Passer en revue les mots-consignes pour s'assurer que les listes de termes sont traduites de manière à ce que la formulation ne soit pas maladroite dans les langues cibles ou sans créer de différences de nuances dans la signification de ces termes entre les langues.
- Traduire les barèmes de notation pour effectuer des recherches plus approfondies sur une plus large gamme de matières afin d'évaluer si l'absence de traduction de ces barèmes a une incidence sur la validité des examens plurilingues.

Les recommandations générales de la troisième phase sont notamment les suivantes.

- Lors de la réflexion sur le choix des mots pendant la traduction, porter une attention particulière à la fréquence relative des mots signifiants (par exemple, les substantifs, les verbes, les adjectifs et les adverbes).
- Lors de la réflexion sur la longueur des phrases pendant l'élaboration des éléments et la traduction, toujours se demander si l'ajout de mots et de propositions rendra le texte plus clair ou plus complexe. Lorsque des phrases plus longues sont utilisées par souci de clarté, s'assurer que ce choix est cohérent entre leurs versions dans les différentes langues.
- Lors de la conception des éléments, éviter autant que possible les longues phrases complexes contenant de nombreux signes de ponctuation. Dans la mesure du possible, essayer d'utiliser des phrases plus courtes pour augmenter la clarté et réduire la charge cognitive associée au traitement des phrases longues.
- Lors du développement des éléments, faire attention à l'utilisation de catégories grammaticales pouvant accroître la complexité, par exemple les adverbes et les adjectifs. Si les catégories grammaticales sont utilisées pour apporter de la clarté, une attention particulière doit être portée à la fréquence relative de leur emploi entre les versions dans les différentes langues.
- Lors de la réflexion sur la structure des phrases, toujours se demander si l'ajout de mots et de propositions rendra le texte plus clair ou plus complexe. Lorsque des phrases plus longues sont utilisées par souci de clarté, s'assurer que ce choix est cohérent entre les versions dans les différentes langues.
- Les logiciels d'analyse de texte peuvent faciliter l'analyse syntaxique des phrases pour déterminer leurs éléments constitutifs, ce qui peut aider à établir des comparaisons concernant la complexité

structurelle des éléments. Dans la mesure du possible, la complexité relative des éléments doit être comparable entre les versions dans les différentes langues.

- Utiliser des logiciels de traitement du langage naturel tels que ReaderBench peut faciliter la prise en compte de toutes ces caractéristiques de complexité textuelle entre les langues. Utiliser un logiciel d'analyse de texte pour effectuer une analyse préliminaire des éléments peut aider à repérer ceux qui, parmi ces derniers, pourraient présenter des difficultés de lecture supplémentaires dans une langue spécifique.
- Combiner les recommandations des deuxièmes et troisièmes phases conduit à la recommandation d'ordre général finale suivante : développer simultanément les versions en langue source et en langue cible d'un examen. Par conséquent, toutes les différences que les experts et/ou l'examen par traitement du langage naturel ont trouvées entre les versions dans les différentes langues pourraient être traitées en apportant des changements dans la version source en anglais. Ces changements se propageraient dans les traductions, et il en résulterait une plus grande convergence linguistique entre toutes les versions.

Conclusion

La conclusion générale de cette étude est que les sciences n'ont pas perdu de sens en traduction pour les examens de 2019 du Programme du diplôme, car les six évaluations ont toutes présenté un degré élevé de comparabilité entre les versions anglaise, espagnole et française. Il semble que les processus de traduction actuels de l'IB qui font appel à la traduction, l'examen et la révision, et qui s'appuient sur l'expertise de l'organisation en matière de traduction et d'évaluation, sont des processus efficaces pour créer des évaluations de difficulté comparable entre ces trois langues. Néanmoins, l'étude a montré que le fonctionnement différentiel d'un nombre non négligeable d'éléments dans l'ensemble des six matières du groupe Sciences du Programme du diplôme était « modéré » et « important ». Il est donc clair qu'il est possible d'améliorer davantage la traduction de ces éléments.

La relation systématique entre les différences de difficulté des éléments dans les trois langues et les autres propriétés psychométriques de ces éléments a mis en évidence le lien entre, d'une part, la conception globale et le fonctionnement général des éléments et, d'autre part, les problèmes de traduction. En particulier, cette relation a montré que certains éléments méritent d'être examinés de manière plus approfondie pour ce qui est de la tendance plus marquée à la supposition dans certains groupes de langues. De plus, les experts ont suggéré que la traduction des éléments pourrait être plus précise sur le plan des correspondances et des structures au sein de l'élément ainsi que par rapport aux formulations comparables afin de transmettre les informations dans les versions traduites des éléments. Enfin, l'analyse par traitement du langage naturel des versions dans les différentes langues a fait ressortir une myriade de différences linguistiques subtiles entre ces versions, qui se sont avérées

être associées, dans une certaine mesure, au fonctionnement différentiel des éléments selon la langue.

Combiner l'analyse par traitement du langage naturel de la complexité textuelle des éléments entre les langues avec l'utilisation de techniques de modélisation d'apprentissage automatique, dans le but d'expliquer le fonctionnement différentiel des éléments (ou l'absence d'un tel fonctionnement) observé, a été une contribution hautement innovante de cette étude. Cette contribution a porté ses fruits en permettant de déterminer les différences linguistiques dans les éléments traduits associées au fonctionnement différentiel des éléments, différences qui seraient passées inaperçues si des méthodes plus conventionnelles avaient été utilisées. Cette approche sera certainement plus efficace si on l'applique aux disciplines du Programme du diplôme dont les examens et les éléments contiennent davantage de texte. Les indices de traitement naturel du langage portant sur la cohésion et le discours pourront alors être appliqués de manière pertinente. Nous pensons, en nous fondant sur les conclusions de l'étude, que l'utilisation de ces technologies d'intelligence artificielle pour prévoir et expliquer le fonctionnement différentiel des éléments basé sur la langue restera un domaine de recherche prolifique et instructif pour de nombreuses évaluations internationales et plurilingues.