Is Science Lost in Translation?
Language Effects in the International Baccalaureate Diploma Programme Science Assessments

Executive Summary

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

Oxford University Centre for Educational Assessment (OUCEA)

Department of Education- University of Oxford

Executive Summary

Context

International Baccalaureate (IB) programmes are offered in over 153 countries around the world. While the majority of the IB schools use English as a medium of instruction, many IB schools teach in either Spanish or French. Consequently, IB Diploma Programme (DP) assessments are administered in English, French and Spanish, and some are also offered in more than 75 other languages. Given the complexities involved in assessment translation and the IB's remit to offer multilingual assessments that are equivalent across the languages, key questions are raised concerning whether all three major language versions are comparable in terms of linguistic and by extension cognitive demands, as well as more broadly in terms of empirical item difficulty. These questions are important because they are relevant to the degree of comparability of IB scores. The equivalence of scores across language versions has significant implications for fairness in access to higher education and this is particularly the case for an international qualification like the IB DP. Consequently, this study provided a first investigation into this issue with respect to the comparability of 2019 DP Science assessments across the English, French and Spanish language versions.

Scope and objectives

The following broad aims drove the approach, methods and findings of the research study:

- Examining the trends and patterns in observed differences in student performance across the source (English) and target (French and Spanish) language versions of individual questions in the 2019 DP Science examinations to evaluate whether and to what extent questions were differentially demanding across the languages.
- Investigating the extent to which these observed differences were due to the translation of questions into the French and Spanish languages that resulted in changes in linguistic and cognitive demand.
- Developing a model to explain observed differences into a range of translation, language and non-translation related effects using both qualitative and quantitative methods.
- Propose improvements to IB translation processes based on the research findings.

The research was carried out in three phases:

- The first phase applied quantitative techniques to the 2019 DP Science examination data to evaluate whether there were systematic differences in the demand of the questions, as represented by their empirical difficulty, based on the examination language, as well as the magnitude of this difference and the language group it favoured.

- The second phase built on the first whereby a subset of the questions identified as showing systematic differences in demand were qualitatively evaluated by bilingual expert reviewers to evaluate whether the source and target language versions of the questions showed differences in terms of key linguistic and translation criteria.
- The third phase took the findings of the first two phases to develop an explanatory model to evaluate whether linguistic and translation differences between the source and target language versions of the questions were substantially associated with differences in demand across the language versions.

Methodological approach

The study applied cutting-edge methods during all three phases of the research:

- The first phase evaluated the differences in demand across the questions using an area of psychometric modelling known as Item Response Theory and specifically the Random Coefficients Multinomial Logit (RCML) model. This phase included response data from English, Spanish and French responding students from the 2019 DP Science assessments, including Physics Standard Level (SL) and Higher Level (HL), Chemistry SL and HL, and Biology SL and HL. Analyses were independently conducted for each of the subject-level combinations and for the source (English) versus the two target (French and Spanish) language versions, so 12 analyses were conducted in total. Demand differences were investigated using a technique called Differential Item Functioning (DIF), which statistically compared the performance of students from the different language groups to quantify any differences in demand across the language versions. Specifically, three models were fit to the response data, one that assumed there was no DIF between the language groups, and two that assumed there was DIF across the groups through the inclusion of a group specific parameter as well as an interaction term between the group and item (i.e., question) difficulty model parameter. In cases where the DIF model had better relative fit to the response data, this interaction term in the model provided an estimate of the magnitude of the DIF (no, small, moderate or large) at the question level, as well as indicating whether the source or target language group were advantaged. These DIF estimates were then compared with other psychometric properties of the questions for each of the DP subjects and a subset were carried forward into the next two phases of research.
- The second phase involved the qualitative, expert review of questions that were identified as having small, moderate and large DIF in Phase 1. As it would have been too resource intensive to review all such questions, a subset of questions was selected from three of the subjects (Physics SL, Chemistry HL and Biology SL) based on several criteria, including covering the different question types (multiple-choice and constructed response) and including a balance of questions

advantaging and disadvantaging the source language group. Ten bilingual/trilingual expert reviewers were recruited in collaboration with the IB to evaluate the comparability of translated versions of the selected questions to the English source version; two in each language-subject combination. The questions were expert reviewed using a 14- to 15-item survey that was newly developed by the researchers based on a renowned translation/verification framework. The survey items addressed eight key criteria from this framework and IB 'house' processes which relate to the veracity of the translations between source and target languages. The inter-rater agreement was calculated for the expert survey judgements and the responses were collated to evaluate whether the DP science questions found to have language DIF in Phase 1 showed linguistic and translation differences that were consistent with this DIF. Moreover, these collated expert review variables were carried forward into the third phase of the research study to contribute to an explanatory model for the DIF.

- The third phase of the research involved building an explanatory model of the language DIF for the same three subjects included in Phase 2. This was conducted in two steps. In the first step, only the subset of items selected in Phase 2 were modelled so that the expert review variables could be included in the model. In addition to the expert review variables, Phase 3 also included indices for the questions based on an area of computational linguistics known as Natural Language Processing (NLP), as well as non-linguistic features like the subject, paper (as a proxy for item type) and target language of the question. The NLP indices were calculated for each subject-language-question combination using an open source, multilingual text processing framework called *ReaderBench*. Previous research with these indices has shown that they are associated with textual complexity and so differences between these indices across the source and target language versions of the questions were expected to help explain language DIF. In the second step, all questions with DIF estimates from the three subjects were included and so the expert review variables were dropped from the model and the focus was on the explanatory power of the NLP indices. This second step was conducted as the quantitative modelling approach used in this phase required large amounts of data to produce robust estimates, and for this reason, the analysis also included 2018 questions and their language DIF estimates for the same three subjects. The explanatory models used in this phase come from machine learning. Specifically, three models (Stepwise regression, Elastic Net regression and Random Forest regression) were applied, as each has its advantages and disadvantages, including being more transparently interpretable (Stepwise regression and Elastic Net regression) versus being more opaque but more flexible in terms of non-linear and complex interaction relationships between the model variables. For both steps, the models were evaluated in terms of their prediction error and explanatory power, and the best performing model was evaluated in terms of the specific variables that were most important in the model for explaining the language DIF. In all cases, the models were applied using a cross-validation approach to enhance the generalizability of the findings.

Main findings

The following were the main findings from the first phase of the research:

- The analyses showed that one of the DIF models was the best fitting model across all subject-level-language combinations, providing evidence that language based DIF was present in all the DP Science examinations across the three language versions.
- More positively for current IB translation process, only a small but still substantial proportion of items showed moderate and large DIF across the subjects, and the larger DIF tended to be more prevalent in the constructed response items from Papers 2 and 3. Overall, the Chemistry subjects had the highest proportion of moderate and large DIF at the item level, followed by the Physics subjects and finally the Biology subjects.
- There was a general trend that the questions that showed significant DIF that advantaged the target languages tended to be the more difficult and less discriminating items, and this was particularly the case for the multiple-choice items. The relationship between the DIF estimates and these other psychometric properties of the items provided evidence that some of the DIF across the languages may be attributable to general fit issues with the items rather than language per se. In particular, some of this DIF may be attributable to guessing behaviour, particularly as the students responding in the target languages tended to be, on average, lower performing across the subjects.
- Physics SL was selected for further inclusion in the other phases of research, as the Physics HL sample size was very small for the French language group and so the DIF estimates for the former were more robust despite generally being smaller in magnitude. Chemistry HL and Biology SL were selected for further inclusion, as the magnitudes of DIF observed for these subjects were generally greater than their other level counterparts.

The following were the main findings from Phase 2 of the research:

- The findings from the expert and qualitative review of questions were very positive for the current translation model adopted by the IB, as the majority of questions were judged to be highly comparable between the French and Spanish target versions and the English source version. Some inconsistencies appeared in specific questions but these inconsistencies, overall, tended to be minor and not systematic with respect to the magnitude or group advantaged by the DIF. For example, the Chemistry HL papers had many more questions categorized as having medium to large language DIF but the translated versions of these items were, based on the judgement of the expert reviewers, more comparable to the English version.

- Of the expert review criteria that did show some deviation between the source and target versions of the questions, *matches and patterns* (matpat) and *accuracy of wording* (word) showed the most consistent and largest degree of deviation, although these deviations still tended to be small in absolute terms.
- Overall, the expert reviewers were able to reliably use the newly developed survey to evaluate the potential differences between source and target versions of the items. These favourable reliability results provided confidence for the use of these variables in the Phase 3 modelling. Nonetheless, some criteria showed consistently lower reliability than the 70% agreement threshold across the subjects and languages. These included the wording and length of clauses criteria, so future applications of this survey should look to enhance the standardized understanding of their meaning.

The following were the main findings from the third phase of the research:

- There were mixed findings regarding how the linguistic and translation differences between source and target language versions of questions explained differences in their difficulty across the language versions. Firstly, none of the language-focused variables from Phase 2's expert review were found to be substantial predictors of the language DIF, but this was consistent with the descriptive findings for these variables in Phase 2. This was likely, at least partially attributable to the lack of variation in these expert review variables.
- Differences in the NLP text complexity indices across the source and target language versions of the items were found to explain the different levels of language DIF observed across the items to some small extent. The performance of the Random Forest model, the best fitting model in both steps, was better for the smaller subset of items from Phase 2, accounting for 11% of the variance in the language DIF outcome variable as opposed to 4% for the larger dataset, which included 2019 and 2018 items for the three subjects (Physics SL, Chemistry HL and Biology SL).
- The most important NLP features for predicting the language DIF outcome variable from the Random Forest regression model could be organised into three themes, with the order following their general order of importance in the model: word choice, sentence length and structural complexity.
  - The word choice indices represent different aspects of how new or unfamiliar information in the text may present challenge for readers in any language. The more expected or predictable a sentence is for a reader, the easier that sentence is to understand. This information could be in the form of words, letters, sentences or even punctuation.

    - The sentence length indices reflect different aspects of how as the length of a sentence increases, the cognitive load associated with processing that sentence increases and this may affect the extent to which readers are able to understand the sentence.
    - The structural complexity indices reflect how different features of a text in terms of the grammatical and syntactical features can manifest in different levels of complexity for the reader.

Recommendations

Specific recommendations arose from each phase of the research. The broad recommendations from Phase 1 included:

- Review multiple-choice items that show differential rates of guessing across language versions to understand what features of these items may lead to increased guessing, in general, and in a specific language.
- Review items that show medium and large language DIF for the other three subjects that were not carried forward to the other two phases of research and for examinations from other calendar years.

Broad recommendations from the second phase included:

- Ensuring translation and quality assurance processes are standardized within and across subjects.
- Decentring the assessment by making it less culture- and dialect-based. This is done by creating two source language versions of the assessment (e.g., English and Spanish) and using both source versions to create a target version (e.g., French).
- Consider translation review and/or quality assurance procedures that enable issues identified in the target version to be reconciled or cross-checked with the source version. For instance, it may be the case that an issue identified in the target version is also relevant for the source version and would require both versions to be adapted.
- Review command terms to ensure that the lists of terms are translated into the target languages without introducing awkwardness in the language or nuanced difference in their meanings across languages.
- Translate mark schemes to conduct further research on a wider range of subjects to evaluate whether the lack of mark scheme translation is having an impact on the validity of the multilingual examinations.

The broad-level recommendations from Phase 3 included:

- When considering word choice during translation, specific attention should be paid to the relative frequency of content words (i.e., nouns, verbs, adjectives, and adverbs) in particular.

- When considering sentence length in item design and translation, always take heed of whether additional words and clauses will add to clarity or add to complexity. When using longer sentences for clarity, try to ensure this is consistent across language versions.
- As far as possible when designing items, avoid longer complex sentences with multiple punctuation marks within the sentence. Wherever possible, try to use shorter sentences to increase clarity and decrease the cognitive load associated with processing long sentences.
- When developing items, care should be taken when using parts of speech that may add to complexity such as adverbs and adjectives. In cases where these parts of speech are used to add clarity, specific attention should be paid to the relative frequency of their use across language versions.
- When considering sentence structure, always take heed of whether additional words and clauses will add to clarity or add to complexity. When using longer sentences for clarity try to ensure this is consistent across language versions.
- Textual analysis software can aid in parsing sentences into constituent parts. This can inform comparisons regarding the structural complexity of items. As far as possible, the relative complexity of items should be comparable across language versions.
- Accounting for all these features of text complexity across languages can be aided by the use of NLP software such as *ReaderBench*. Pre-screening items using textual analysis software can aid in identifying whether there are items that may present additional reading challenge in a specific language version.
- Combining the recommendations from Phases 2 and 3, a final broad level recommendation is to concurrently develop the source and target language versions of an examination. Consequently, any discrepancies between the language versions identified by expert and/or NLP review may be addressed by changes to the English source version and propagated through the translations, thereby resulting in greater linguistic convergence between all language versions.

Conclusion

The overarching conclusion from this research study was that science was not lost in translation for the 2019 DP Science examinations, as all six assessments showed a high degree of comparability across the English, French and Spanish language versions. It appears that the current IB translation processes involving forward translation and review and revision, drawing on both translation and IB assessment expertise, is effective in creating assessments with comparable difficulty across these three languages. Nonetheless, there were still a substantial number of items across all six DP science subjects that showed moderate and large language DIF and so it is clear that further improvements could be made to the translation of items.

The systematic relationship between the differential difficulty of items across languages and the items' other psychometric properties highlighted the connection between general item design/functioning and translation issues, and in particular, that some items warrant further investigation in terms of pronounced guessing behaviour by some language groups. Moreover, the expert review suggested that the translation of items could be more precise in terms of matches and patterns within the item, as well as with respect to comparable wording to convey information in the translated versions of items. Finally, NLP analysis of the different language versions of the items showed a myriad of subtle linguistic differences between them, which were shown to be associated with the language DIF to some extent.

The NLP analysis of item text complexity across languages combined with the use of machine learning modelling techniques to explain the language DIF (or lack thereof) observed for items was a highly innovative contribution of the current research study, which has borne fruit in terms of identifying linguistic differences in translated items that are associated with DIF that otherwise would have been missed by more conventional methods. This approach could be more effective when applied to DP subject areas where the examinations and items contain more text and so NLP indices concerned with cohesion and discourse can be meaningfully applied. Based on this study's findings, we believe that the use of these artificial intelligence technologies to predict and explain language-based DIF will continue to be a fruitful and informative area of research for various international and multilingual assessments.