

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

Final Report



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

Oxford University Centre for Educational Assessment (OUCEA)
Department of Education - University of Oxford



Contents

Acknowledgements	5
Executive Summary	6
Context	6
Scope and objectives	6
Methodological approach.....	7
Main findings	9
Recommendations	11
Conclusion.....	12
Introduction	14
Challenges of translating and adapting educational assessments.....	14
Empirical evidence of comparability issues for assessments in multiple languages	15
Methodological challenges in evaluating the comparability of language versions of assessments	16
Research Aims and Phases.....	16
Phase 1: Psychometrically evaluating DP Science examinations for DIF	18
Method	19
Results.....	25
Physics DIF by examination language	25
Chemistry DIF by examination language.....	29
Biology DIF by examination language	32
DIF and other psychometric properties	36
Discussion	38
Phase 2: Expert Review of selected DIF items	41
Methods.....	41
Item selection	41
Expert reviewers	42
The instrument	42
Data collection	43
Analysis	45
Results.....	45
Percentage agreement	45
Distance of reviewers' judgements from the neutral category	47
Discussion	60
Phase 3: Building a model to explain DIF across languages.....	62

Method	62
Quantifying differences in text complexity using NLP	62
Modelling DIF using a machine learning approach	69
Results.....	72
Explanatory models including expert review variables.....	72
Explanatory models including both 2018 and 2019 data	75
Discussion	77
Word choice.....	78
Sentence length	80
Structural complexity.....	81
Conclusion.....	82
Recommendations	84
Translation processes at IB	84
Recommendations from Phase 1.....	85
• Review multiple-choice items that show differential rates of guessing across language versions.....	85
• Review items that show medium and large language DIF for the other three subjects and other years.....	85
Recommendations from Phase 2.....	86
• Back-translation at the revision stage of the assessment	86
• Decentring the assessment	86
• House style – reviewing command terms	87
• Translation of mark scheme	87
Recommendations from Phase 3.....	87
• Account for textual complexity associated with word choice	88
• Account for textual complexity associated with sentence length.....	88
• Account for textual complexity associated with structural complexity	88
Conclusion	90
References	92

Final Report - Technical Documentation

Appendix 1 – Items that were collapsed or deleted for the main DIF analyses	96
Appendix 2 – DIF Estimates for link items across the Physics, Chemistry and Biology SL and HL examinations	102
Common-item linking DIF by examination levels.....	102
Appendix 3 – Covariate estimates and their effect sizes from the DIF plus covariate models.....	111
Appendix 4 – DIF Estimates for all items across the different subject, level and language combinations	124
Appendix 5 – Forest plots of DIF Estimates for all items across the different subject, level and language combinations.....	157
Appendix 6 – The Expert Review Instrument.....	193
Appendix 7 – Screenshot of instrument	199
Appendix 8 – Example of mark scheme	200
Appendix 9 – <i>ReaderBench</i> Tutorial.....	201

Acknowledgements

We would like to sincerely thank the International Baccalaureate for the support they have provided in conducting this research, including providing language experts for the expert review and helping with the extraction of text from the Diploma Programme Science examination papers. In particular, Rebecca Hamer has made significant contributions to this project across all phases and has been invaluable in overcoming the various challenges that this complex project combined with a global pandemic have inevitably thrown up. Finally, we would like to thank all the IB Diploma Programme students, teachers, markers, schools and the Assessment Division for their contributions to the examination data that made this project possible.

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.

Introduction

International Baccalaureate (IB) programmes are offered in over 153 countries around the world. While the majority of the IB schools opt for English as a medium of instruction, many IB schools teach in either Spanish or French. Consequently, IB assessments are administered in English, French and Spanish, with some also being offered in more than 75 languages. Given the complexities involved in assessment translation, key questions are whether all three major language versions of the IB assessments 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; in other words, the extent to which, for example, a score of 100 marks on an IB DP examination, say Biology SL, administered in English is equivalent to a score of 100 marks on the same assessment administered in French or Spanish. The equivalence of scores across language versions has significant implications for fairness in access to higher education where student scores and their conversions to grades 1 to 7 are treated as being comparable irrespective of the language of the assessment. These grades play a substantial role in determining the chances that students may pursue higher education, and particularly for the highly prestigious and competitive universities in various countries, which may attract applications from students of various national and linguistic backgrounds.

Further, fairness of assessments is a central criterion of assessment quality for the IB Assessment Division, which oversees the development, administration and marking of IB assessments and adopts a rigorous process of quality control to ensure that all IB examinations are valid, reliable, fair and manageable. There are various procedures put in place to ensure that IB assessments are of high quality; however, to date, there has been insufficient empirical evidence gathered regarding the comparability of IB assessments across languages. In this research project, we quantitatively and qualitatively examine the different language versions of the IB DP Science examinations, given the importance of this subject in secondary education, to rigorously evaluate the extent to which the language versions are comparable and provide equal opportunities of success for all students irrespective of the language version they sit.

Challenges of translating and adapting educational assessments

Hambleton (2005) highlighted challenges associated with the translation and adaptation of educational and psychological tests. He argued that no matter how rigorous the method of translation and adaptation is, bias in translation is inevitable at some stage. For instance, in forward translation, the model of translation adopted by the IB, a translator translates a test from a source language to a target language. However, translators typically have different levels of proficiency in the languages

they master and are often more comfortable in one language in the context of a specific subject (e.g., mathematics or science). These translator characteristics inevitably impact the quality of the translation. Moreover, unlike the backward translation model, forward translation does not include a process where the translated text is translated back to the source language, and consequently, inconsistencies in the translation can be totally missed. Having said that, IB currently adopts a number of quality control processes in order to address potential inconsistencies that might occur during forward translation, but given the stakes that IB DP assessments can have on students' future opportunities, it is important to examine the extent to which the current translation model and quality control processes adopted by the IB are effective in producing comparable language versions of assessments.

Empirical evidence of comparability issues for assessments in multiple languages

Various studies on international large-scale assessments, such as the Programme for International Student Assessment (PISA) and the Trends in Mathematics and Science Study (TIMSS), have pointed to issues of comparability of assessments across languages. For instance, Ercikan and Koh's (2005) study highlighted the lack of comparability of constructs between the English and French versions of the same TIMSS mathematics assessment. Huang, Wilson and Wang (2016), and El Masri et al. (2016) pointed to issues with the comparability of the English version of PISA science assessments with Chinese, French and Arabic versions of the same assessment. El Masri et al. (2016) argued that language is an inextricable part of what is being assessed in science which compounds the potential effects of translating science assessments. For instance, one of the examples the authors provide relates to the translation of scientific terms. These could be of high frequency in one language but of low frequency (and hence placing higher cognitive demands) in the translated language, or vice versa. Similar issues have been reported in translating science assessments from English to Spanish in the USA context (Solano-Flores & Nelson-Barber, 2001) and from English to Welsh in the British context (William, 1994). Scientific terminology is one of the many features of scientific language that have been identified in the literature as posing additional demand for students when reading science textual material (Halliday & Martin, 1993). Other features include long nominal sentences, everyday language, syntactic complexity, metaphors, etc.

Similarly, recent research on IB assessments has pointed to some inconsistencies in expectations and interpretations of assessment criteria, marking rubrics and standards across English and Spanish versions of assessments (e.g., Galache Ramos, 2017). However, this evidence has been focused on the comparability of marking across languages and did not examine the extent to which the translation of IB questions, mark schemes, etc., could impact students' performance on the assessments across languages.

Methodological challenges in evaluating the comparability of language versions of assessments

Studies investigating the comparability of items across groups typically use statistical techniques known as differential item functioning (DIF), which is a method that is employed in both Classical Test Theory and Item Response Theory frameworks (Penfield & Camilli, 2007; Zumbo, 2007). DIF techniques have been widely employed in examining the comparability of international assessments (e.g., Asil & Brown, 2016; Ercikan & Koh, 2005; Grisay, de Jong, Gebhardt, Berezner, & Halleux-Monseur, 2007; Grisay, Gonzalez, & Monseur, 2009; Hauger & Sireci, 2008; Huang, Wilson, & Wang, 2016; Kreiner & Christensen, 2014; Le, 2009; Oliveri, Olson, Ercikan, & Zumbo, 2012; Sandilands, Oliveri, Zumbo, & Ercikan, 2013; Wu & Ercikan, 2006; Xie & Wilson, 2008). Despite the popularity of these techniques, DIF only points to inconsistencies across groups in the expected performance of students on an item based on their performances on the whole assessment (i.e., their total score). However, DIF analysis does not identify the source of this inconsistency between groups. Many variables could have interacted and led to the inconsistency in performance of students across, for example, language groups, and the number of these variables becomes even larger in international assessments due to cultural differences, resources, quality of teaching, etc. Therefore, it is necessary to follow up any statistical analysis with in-depth qualitative analyses of questions (El Masri & Andrich, 2020; McGrane et al., 2014).

Research Aims and Phases

The following broad aims drove the approach, methods and findings of this project:

- Examining the trends and patterns in observed differences in student performance across the source (English) and target (French and Spanish) language variants of individual questions in DP Science examinations.
- Investigating the extent to which these observed differences are due to the translation of questions into the French and Spanish languages resulting in a change in demand.
- Developing a model separating observed differences into a range of translation, language and non-translation related effects using both qualitative, expert review of the questions and quantitative analysis methods, including computational linguistics and machine learning methods.
- Propose improvements to IB processes addressing differences in student performance on individual examination questions that can be attributed to existing IB translation practices.

The research was carried out in three phases. The first phase applied quantitative, psychometric techniques to the 2019 DP Science examination data to evaluate whether there were systematic differences in the demands of the questions based on the examination language, and further, quantified the magnitude of the demand difference and the group it favoured for each of the

questions across the examinations and language versions. The second phase built on the first whereby a subset of DP Science examination questions identified as showing systematic differences in demand was evaluated by bilingual expert reviewers using a survey based on a newly developed expert review framework. The survey included questions to evaluate whether the source and target language versions of the questions showed differences in terms of key linguistic and translation criteria providing potential explanations for the observed differences.

The third and final phase used 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 could be used to predict the magnitudes of differences in demand across the language versions. In addition to the variables from the expert review, this model also included variables from an area of computational linguistics known as Natural Language Processing (NLP), which was used to quantify differences between source and target versions of questions in terms of features that are known from previous research to correspond with the reading demand of text. The specific expert review and NLP features that showed a substantial association with language-based demand differences were identified using a machine learning modelling approach.

The findings from all three phases of the research study were used to evaluate the comparability of the English, French and Spanish versions of the DP Science examinations to highlight the current success of, and make recommendations for future improvements to the existing IB translation processes. This included highlighting the linguistic and other features that are most strongly associated with empirical differences in demand across the source and target languages, which may then be given specific attention during the translation of future IB science assessments.

Phase 1: Psychometrically evaluating DP Science examinations for DIF

The psychometric evaluation of differences in demand of examination questions across different groups of students, e.g., those responding to different languages of the exam, was carried out at the overall examination level and at the individual question level. There are various ways this can be psychometrically approached. This study adopted an approach, consistent with Huang et al. (2016), where different models were applied to the examination response data to evaluate whether there were differences in demand at both the overall examination and individual question levels. The first of these models did not include a group (e.g., examination language) by question parameter for each question and so assumed that there were no differences in the demand of the questions across the groups. The second model did include a group-by-question parameter for each question and so allowed for the estimation of differences in the demand of the questions across the groups. When the second model shows greater correspondence with the response data, which is typically referred to as better ‘fit’, than the first model, then this is consistent with there being overall substantial differences in the demand of the questions across the groups. These differences then can be investigated at the question level in terms of the group-by-question parameter estimates to identify the specific questions that show the greatest differences in demand.

Finally, a third model was added that was similar to the second model but additionally added group-specific parameters that may confound the evaluation of the group-by-question estimates in the second model. For example, if there was a disproportionate number of males than females in one of the language groups, and gender was associated with differential performance on the exam, then the second model, which only includes a parameter for language, may have led to the conclusion that there were systematic differences in question demand based on examination language, when the differences actually were (at least partially) attributable to gender. On the other hand, this third type of model may account for this potential confound through the additional inclusion of a gender-specific parameter. Moreover, if the third model showed greater fit to the response data than the second model, then the third model’s group-by-question estimates should be used to evaluate differences in demand for the grouping factor of interest, in our case examination language, so that the estimates are not confounded by other group factors.

This approach to evaluating differences in examination question demand was applied to evaluate the following research questions:

- Do the DP Science examinations, overall, show differences in question demand across the English (source) and French and Spanish (target) language versions?

- Which DP Science examination questions and question types demonstrate systematic differences in demand in the target languages compared to the source language version, and what are the magnitudes of these differences?
- Are there any patterns in the psychometric properties (difficulty and fit) of questions that show differential demand between the source and target languages?

The first phase's methodological approach and the findings with respect to the above research questions are now elaborated in more specific detail below.

Method

To evaluate these three research questions, the full set of questions from the 2019 DP Physics, Chemistry and Biology SL and HL examinations (six in total) was evaluated for DIF between the English, French and Spanish versions using an Item Response Theory (IRT) framework. Given the use of IRT, the examination questions will now be referred to as items and demand will be referred to as difficulty. Similarly the source language group (English) will be referred to as the reference group and the target language groups (French and Spanish) as focal groups.

Differential Item Functioning – DIF – is a psychometric technique used to evaluate whether examination questions are systematically easier or more difficult for certain group(s) of students sitting the exam. DIF works by comparing students from these different groups who have the same overall performance on the exam (as reflected in their total mark) to establish whether any item(s) are systematically more difficult/easy for one group despite this common level of overall performance. In this way, DIF does *not* reflect the average difference in performance on an item by the group. Moreover, DIF is normally evaluated in terms of a 'reference group' and a 'focal group' whereby, as the names suggest, the latter are compared with the former. In this project, the French and Spanish examination groups will be the two focal groups.

Specifically, like the approach taken by Huang et al. (2016) and as described in general terms in the introduction to this phase of the research, we employed the Rasch model-based random coefficient

The *random coefficient multinomial logit – RCML* – model is a mash up of a Rasch model and a type of regression model that is commonly used in economics. As per the typical presentation of the Rasch model, the log-odds of a correct response to an item is modelled as a trade-off of the item's difficulty (estimated in terms of how many people obtained a correct response) and the person's ability (estimated in terms of their overall score on the exam), whereby the more difficult an item is relative to a student's ability, the less likely it is that they will obtain a correct response. However, as the RCML is a flexible, regression-type model, you can also add other predictors to the model like a salient group factor(s) and a group-by-item difficulty interaction term, which provides an estimate of DIF because this estimate would be equal to zero if the item's difficulty did not vary across the groups and is non-zero when the difficulty is systematically different across the groups.

multinomial logit (RCML) model DIF detection procedure, which Paek and Wilson (2011) term the Rasch DIF model. The latter showed that the Rasch DIF model can be effectively used for samples sizes as small as 100, and it is not contingent on the different groups having similar average performance or distributions of performance. All analyses were carried out using the Test Analysis Modules (TAM) package in the R statistical software (Kiefer, Robitzsch, & Wu, 2020), which is now routinely used for psychometric analyses in research contexts (e.g., Robinson, Johnson, Walton, & MacDermid, 2019).

The Rasch DIF model DIF detection procedure involves comparing the performance of students based on the interactions between a number of variables. In order to do this, three models were estimated:

1. A 'no DIF' model.

In this model, only the usual parameters in the Rasch model, i.e., item difficulty and person ability, were estimated and the item difficulties are assumed to be the same across the source and target examination languages.

2. A 'DIF' model.

In this model, the group variable (i.e., examination language) was added as a parameter to the model, as well as its interaction with item difficulty. The interaction between the group variable and item difficulty represents the DIF variable. In cases where there are no differences in item difficulties across the groups, these interaction estimates will not improve model fit. Moreover, the inclusion of the group factor as a covariate in the model means that the item-by-group (DIF) interaction estimates are not contaminated by differences in average performance across the groups because these differences are partialled out in the estimation. We will refer to this second model throughout this report as the 'DIF' model and the group variable of interest will be the examination language. These two models were statistically compared following the procedure described below to establish whether adding the DIF parameters led to better overall model fit.

3. A 'DIF+covariates' model.

In this model, additional group factors, termed covariates, were added to the model. Covariates are additional factors that are added to a model so as not to confound the estimates of the variables that are of central interest (in this case examination language and its interaction with item difficulty). In this case, we added the covariates gender, sub-region as a proxy for cultural differences, and first/second language match¹ between the student and the examination language. While, in theory, an indefinite number of covariates may be added to the model, each additional covariate adds to the

¹ In cases where the exam language could not be matched with the student's first or second language, this was coded as 'no match'.

model complexity and processing time, and so these three covariates were considered the most salient while also maintaining the tractability of the model.

Model comparison was carried out using a number of statistical criteria. The first of these was a log-likelihood ratio test, whereby the difference in the compared models' log-likelihood (which is closer to zero with better model fit) is the quantity of interest, the difference in the number of model parameters being estimated is the degrees of freedom for the significance test, and the statistical significance is determined relative to a chi-square distribution. In addition, the models were compared in terms of the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), where, again, values closer to zero represent better relative model fit. Both these criteria quantify the deviance between the model predictions and the actual data and add a penalty for the number of parameters in the model, whereby the BIC's penalty is more severe than the AIC's. This penalty is added as models with more parameters almost always fit the data better despite the extra parameters adding no value in terms of generalisability of the model; what is sometimes referred to as 'overfitting'. Thus, the AIC and BIC attempt to minimise the possibility of overfitting to varying degrees and we will thus refer to both when evaluating model fit.

The Rasch DIF model also provided DIF diagnostics at the item level. Specifically, the interaction estimate from the DIF model represents the magnitude of the difference in difficulty for an item between the focal (French or Spanish) group and the reference (English) group. This difference is then statistically evaluated for significance using a Wald test, whereby the estimate is divided by its standard error and then evaluated in terms of the z-distribution. Given that many DIF estimates are being simultaneously evaluated for statistical significance, a Bonferroni correction is applied to control for the increased likelihood of concluding that an item has significant DIF by random chance alone, which involves dividing the critical p-value (.05) by the number of comparisons, i.e., the number of items being evaluated for DIF in the analysis. In addition, as the interaction estimates are standardized, they may be interpreted in a manner akin to effect sizes, whereby even though an item may have statistically significant DIF, the magnitude of the DIF may be so small that it is not considered to be of practical significance. Therefore, after determining whether the items display statistically significant DIF or not, using the cut-offs provided by Paek and Wilson (2011), the significant DIF estimates were categorised into small (A+/A-), moderate (B+/B-) and large (C+/C-) effect size groups, whereby the valence (+/-) reflected whether the item was systematically less or more difficult for the focal group given a comparable level of performance across the whole exam. Typically, only items that have moderate or large DIF are taken to be practically significant when considering the comparability of an assessment across groups.

Finally, the Rasch DIF model also provided information about other psychometric properties of the examination, including, as addressed above, the difficulty of the items across the whole cohort, as well as statistics that quantify how well the items fit the Rasch model. The most used fit statistic for the Rasch model is the infit statistic. This statistic has an expected value of 1, and as the value gets closer to 0, it provides evidence that the item discriminates more than model expectation (i.e., the average discrimination of all modelled items), and as the value becomes larger than one, it provides evidence that the item discriminates less than model expectation. This ‘under-discrimination’ means that the lower ability students were getting the item correct more often than you would expect, or vice-versa regarding the higher ability students, and may be indicative of guessing for multiple-choice items or some kind of ‘clue’ in the item for lower ability students, or conversely that some feature of the item is ‘tricking up’ the higher ability students. Typically, under-discrimination is considered the more problematic form of misfit because of its association with factors like guessing or confusing/ambiguous aspects of the question. These psychometric properties of item difficulty and fit were compared with the DIF estimates to evaluate such questions as whether easier or more difficult items tended to show DIF across the languages, and whether the DIF also tended to correspond with other item fit issues like over or under-discrimination.

The analyses were carried out in the following steps:

1. The 2019 examination data for the six DP subjects and the three languages were prepared, cleaned and recoded for analysis in R². Table 1 below provides a summary of the sample sizes for the different subject-level-language combinations. It should be noted that the sample sizes for the French cohort in both Physics levels were quite small (<100) and so the findings for these groups in the report should be interpreted with due caution.
2. DIF analyses by examination language³ were then conducted separately for each of the six subjects. Moreover, separate sets of analyses were carried out with French as the focal group and then Spanish as the focal group vs. English across each of the six subjects – meaning there were 12 sets of analyses conducted in total. As the sample sizes were much smaller for the French and Spanish cohorts in certain subjects, some further data preparation had to be carried out,

² As part of this, in cases where successive integer scoring was not used, items given ordered marks beyond just incorrect and correct were recoded so that their scoring commenced from 0 (lowest performance) and proceeded in successive integers (i.e., 0, 1, 2, 3, etc.) to reflect ordered performance levels on the item, as this successive integer scoring is a requirement of the Rasch Model.

³ DIF analyses were also conducted to evaluate whether the SL and HL examinations for the three subjects could be simultaneously analysed using the questions that are common across the two levels as links, however, the results suggested that the difficulty of these common questions systematically varied across the two levels and so the levels were separately analysed to evaluate DIF by exam language. The results of these DIF analyses in terms of the common linking items between the levels are presented in Appendix 2.

particularly as both Papers 2 and 3 had optionality and so some items in these papers had few responses in either focal language. This involved ensuring that each score category for each item had at least five observations across the language groups to ensure that any observed DIF was not a consequence of a small or zero response rate in a category. Table 2 provides a summary of the number of items that had to have their categories collapsed⁴ or be completely deleted to ensure this minimum response across the two focal languages – the full set of items can be found in Appendix 1. As can be seen, there were far more items either collapsed or deleted for the French cohort across the different subjects given their smaller sample sizes.

3. Each set of DIF analyses involved fitting the ‘no DIF’, ‘DIF’ and ‘DIF+covariates’ models to the data and comparing the overall model fit using the different criteria, as well as evaluating the item-level DIF statistics from the best fitting model and quantifying the number of items with small, moderate and large DIF, including across item type (multiple-choice items in Paper 1 vs. constructed response items in Papers 2 and 3).
4. Finally, the 12 sets of DIF estimates from the best fitting model were correlated with their respective item difficulty estimates and infit statistics to evaluate whether there was any systematic relationship between these psychometric properties and DIF by examination language relative to the English source language.

⁴ Collapsing involves combining adjacent score categories so there are at least 5 observations in each of the score categories for an item, e.g., combining the ‘2’ and ‘3’ score categories of an item into ‘2’ in the case that fewer than 5 students in that group scored a ‘3’ for the item.

Table 1.

Sample sizes of the cohorts for each of the subject-level-language combinations.

Subject	Level	Language	N
Physics	SL	English	4403
		French	84
		Spanish	1188
	HL	English	9020
		French	53
		Spanish	431
Chemistry	SL	English	6075
		French	234
		Spanish	1544
	HL	English	10491
		French	159
		Spanish	271
Biology	SL	English	7922
		French	253
		Spanish	3077
	HL	English	12158
		French	230
		Spanish	754

Table 2.

The number (n) of items that had to be collapsed or deleted to ensure the minimum category response rate across the different subjects and focal languages.

Subject	Level	Language	Collapsed Items (n)	Deleted Items (n)
Physics	SL	French	16	19
		Spanish	1	0
	HL	French	38	49
		Spanish	6	2
Chemistry	SL	French	5	11
		Spanish	4	0
	HL	French	4	39
		Spanish	10	9
Biology	SL	French	12	7
		Spanish	0	0
	HL	French	7	31
		Spanish	6	0

Results

Physics DIF by examination language

Exam-level DIF

For SL French vs. English, the 'no DIF' model had statistically significantly worse fit than the 'DIF' model, $\chi^2(96) = 632.69, p < .001$. Similarly, the 'no DIF' model had statistically significantly worse fit compared to the 'DIF+covariates' model, $\chi^2(109) = 1014.13, p < .001$. For the 'DIF' model compared to the 'DIF+covariates', the latter had statistically significantly better fit according to the likelihood-ratio test, $\chi^2(13) = 381.44, p < .001$. These results were also confirmed by the AIC, however, only the 'DIF+covariates' model had better fit than the 'no DIF' model according to the BIC (see Table 4 below).

For SL Spanish vs. English, the 'no DIF' model had statistically significantly worse fit than the 'DIF' model, $\chi^2(115) = 2284.46, p < .001$. Similarly, the 'no DIF' model had statistically significantly worse fit compared to the 'DIF+covariates' model, $\chi^2(127) = 2824.82, p < .001$. For the 'DIF' model compared to the 'DIF+covariates', the latter had statistically significantly better fit according to the likelihood-ratio test, $\chi^2(12) = 540.36, p < .001$. This pattern of results was also confirmed by both the AIC and BIC (see Table 3).

For HL French vs. English, the 'no DIF' model had statistically significantly worse fit than the 'DIF' model, $\chi^2(121) = 536.27, p < .001$. Similarly, the 'no DIF' model had statistically significantly worse fit compared to the 'DIF+covariates' model, $\chi^2(132) = 986.14, p < .001$. For the 'DIF' model compared to the 'DIF+covariates', the latter had statistically significantly better fit according to the likelihood-ratio test, $\chi^2(11) = 449.87, p < .001$. These results were also confirmed by the AIC. However, the BIC favoured the 'no DIF' model relative to the other two models (see Table 3).

For HL Spanish vs. English, the 'no DIF' model had statistically significantly worse fit than the 'DIF' model, $\chi^2(168) = 1313.91, p < .001$. Similarly, the 'no DIF' model had statistically significantly worse fit compared to the 'DIF+covariates' model, $\chi^2(179) = 1718.31, p < .001$. For the 'DIF' model compared to the 'DIF+covariates', the latter had statistically significantly better fit according to the likelihood-ratio test, $\chi^2(11) = 404.40, p < .001$. These results were also confirmed by the AIC, however, only the 'DIF+covariates' model had better fit than the 'no DIF' model according to the BIC (see Table 3).

Table 3.

Model comparison statistics for the models with and without DIF parameters, and with and without covariates for Physics across the different level and language combinations.

Level	Language	Model	Log Likelihood	AIC	BIC	No. of parameters
SL	French	No DIF	-226294.91	452881.83	453834.65	146
		DIF	-225978.57	452441.14	454020.47	242
		DIF + cov.	-225787.85	452085.70	453749.87	255
	Spanish	No DIF	-300487.25	601326.50	602520.13	176
		DIF	-299345.02	599272.03	601245.60	291
		DIF + cov.	-299074.84	598755.68	600810.63	303
HL	French	No DIF	-554001.23	1108366.47	1109661.13	182
		DIF	-553733.10	1108072.20	1110227.59	303
		DIF + cov.	-553508.16	1107644.33	1109877.97	314
	Spanish	No DIF	-684321.10	1369168.21	1371049.79	263
		DIF	-683664.15	1368190.30	1371273.80	431
		DIF + cov.	-683461.95	1367807.90	1370970.10	442

On balance, it appears that the ‘DIF+covariates’ model has the best fit for each of the levels and languages, even when penalised for the extra parameters in the model⁵. Thus, given the superior fit of the DIF model versus the no DIF model in all cases, it appears that for all four Physics level and language combinations, there is substantial DIF at the overall examination level that warrants further evaluation at the item level, and this evaluation is presented in the next subsection. Moreover, given the superior fit of the ‘DIF+covariates’ model, the z-score⁶ DIF estimates from this model were used throughout the rest of the study for each of the Physics level-language combinations, as these estimates are not confounded by any average differences in these covariates across the examination

⁵ The one partial exception to this was for the HL French vs. English models but given that both the likelihood-ratio test and AIC provided support for the ‘DIF+covariates’ model, it was considered the best model for this cohort.

⁶ To calculate the standardised z-score from the DIF estimate, the estimate is divided by its standard error.

language groups⁷. The full set of covariate estimates from the 'DIF+covariates' model across all the subject-level-language combinations are presented in Appendix 3.

Item-Level DIF

Table 4 provides a breakdown of the frequencies and percentages of the different levels (small, moderate and large) of DIF observed in Physics for the different level-language and paper combinations. The complete set of item-level DIF estimates may be viewed in Appendix 4 and the forest plots of the item-level DIF estimates across the three papers for each subject are presented in Appendix 5.

Across the four level-language combinations, the HL French vs. English model items showed the most substantial amount of DIF, with 58% of the items found to have moderate or large DIF. However, these findings must be interpreted cautiously given the very small sample size of the French cohort. For the other level-language combinations, where the sample size of the non-English cohorts was substantially larger (except for French SL), most items were found to either have non-significant DIF or only small DIF effect sizes – overall, 75% of Physics items were found to have either no statistically significant DIF or only small DIF (A+/A-). Moderate DIF was observed for 12% of the items and large DIF was observed for 14% of items. This moderate and large DIF was observed in both multiple-choice (Paper 1) and constructed response (Papers 2 and 3) items, but on balance, more practically significant DIF was observed for the latter papers.

⁷ It should be noted that despite the 'DIF+covariates' model consistently showing the best fit to the data for all subject-level-language combinations, the correlations between the DIF estimates obtained from this model and the 'DIF' (without covariates) model were extremely high, ranging from .98 to essentially 1, i.e., adding the additional covariates to the models lead to extremely small changes in the DIF estimates and so either could have been used in Phases 2 and 3 of the research.

Table 4.

Frequencies (top tier) and percentages (bottom tier) of different DIF levels observed for the Physics items across the different level-language-paper combinations.

DIF Category								
	A-	A+	B-	B+	C-	C+	No	Total
Physics	113	79	33	24	29	41	181	500
HL	48	38	22	16	28	29	108	289
French	20	8	15	5	24	27	22	121
Paper 1	10	4	6	1	11	2	5	39
Paper 2	7	3	6	2	5	18	6	47
Paper 3	3	1	3	2	8	7	11	35
Spanish	28	30	7	11	4	2	86	168
Paper 1	6	8	2	1	2	0	20	39
Paper 2	16	9	3	3	1	1	18	51
Paper 3	6	13	2	7	1	1	48	78
SL	65	41	11	8	1	12	73	211
French	30	15	6	3	1	7	34	96
Paper 1	12	6	1	1	0	0	10	30
Paper 2	14	3	1	0	0	2	9	29
Paper 3	4	6	4	2	1	5	15	37
Spanish	35	26	5	5	0	5	39	115
Paper 1	14	6	0	1	0	0	9	30
Paper 2	11	8	3	0	0	1	6	29
Paper 3	10	12	2	4	0	4	24	56
	A-	A+	B-	B+	C-	C+	No	
Physics	23%	16%	7%	5%	6%	8%	36%	
HL	17%	13%	8%	6%	10%	10%	37%	
French	17%	7%	12%	4%	20%	22%	18%	
Paper 1	26%	10%	15%	3%	28%	5%	13%	
Paper 2	15%	6%	13%	4%	11%	38%	13%	
Paper 3	9%	3%	9%	6%	23%	20%	31%	
Spanish	17%	18%	4%	7%	2%	1%	51%	
Paper 1	15%	21%	5%	3%	5%	0%	51%	
Paper 2	31%	18%	6%	6%	2%	2%	35%	
Paper 3	8%	17%	3%	9%	1%	1%	62%	
SL	31%	19%	5%	4%	0%	6%	35%	
French	31%	16%	6%	3%	1%	7%	35%	
Paper 1	40%	20%	3%	3%	0%	0%	33%	
Paper 2	48%	10%	3%	0%	0%	7%	31%	
Paper 3	11%	16%	11%	5%	3%	14%	41%	
Spanish	30%	23%	4%	4%	0%	4%	34%	
Paper 1	47%	20%	0%	3%	0%	0%	30%	
Paper 2	38%	28%	10%	0%	0%	3%	21%	
Paper 3	18%	21%	4%	7%	0%	7%	43%	

Chemistry DIF by examination language

Exam-level DIF

For SL French vs. English, the ‘no DIF’ model had statistically significantly worse fit than the ‘DIF’ model, $\chi^2(112) = 472.01$, $p < .001$. Similarly, the ‘no DIF’ model had significantly worse fit compared to the ‘DIF+covariates’ model, $\chi^2(124) = 884.44$, $p < .001$. For the ‘DIF’ model compared to the ‘DIF+covariates’, the latter had statistically significantly better fit according to the likelihood-ratio test, $\chi^2(12) = 412.43$, $p < .001$. These results were also confirmed by the AIC. However, the BIC favoured the ‘no DIF’ model relative to the other two models (see Table 5 below).

Table 5.

Model comparison statistics for the models with and without DIF parameters, and with and without covariates for Chemistry across the different level and language combinations.

Level	Language	Model	Log Likelihood	AIC	BIC	No. of parameters
SL	French	No DIF	-327441.50	655194.99	656248.17	156
		DIF	-327205.49	654946.98	656756.29	268
		DIF + cov.	-326999.28	654558.55	656448.88	280
	Spanish	No DIF	-400318.20	800988.40	802209.77	176
		DIF	-398192.62	796983.24	799058.17	299
		DIF + cov.	-397785.91	796195.82	798360.97	312
HL	French	No DIF	-749837.62	1500077.24	1501539.42	201
		DIF	-749386.13	1499454.26	1501934.87	341
		DIF + cov.	-749084.96	1498875.91	1501443.82	353
	Spanish	No DIF	-802712.83	1605925.66	1607746.91	250
		DIF	-801842.05	1604524.11	1607583.80	420
		DIF + cov.	-801576.74	1604017.47	1607164.59	432

For SL Spanish vs. English, the ‘no DIF’ model had statistically significantly worse fit than the ‘DIF’ model, $\chi^2(123) = 4251.17$, $p < .001$. Similarly, the ‘no DIF’ model had significantly worse fit compared to the ‘DIF+covariates’ model, $\chi^2(136) = 5064.58$, $p < .001$. For the ‘DIF’ model compared to the ‘DIF+covariates’, the latter had statistically significantly better fit according to the likelihood-ratio test,

$\chi^2(13) = 813.42, p < .001$. This pattern of results was also confirmed by both the AIC and BIC (see Table 5).

For HL French vs. English, the 'no DIF' model had statistically significantly worse fit than the 'DIF' model, $\chi^2(140) = 902.98, p < .001$. Similarly, the 'no DIF' model had significantly worse fit compared to the 'DIF+covariates' model, $\chi^2(182) = 2272.19, p < .001$. For the 'DIF' model compared to the 'DIF+covariates', the latter had statistically significantly better fit according to the likelihood-ratio test, $\chi^2(12) = 602.35, p < .001$. These results were also confirmed by the AIC, however, only the 'DIF+covariates' model had better fit than the 'no DIF' model according to the BIC (see Table 5).

For HL Spanish vs. English, the 'no DIF' model had statistically significantly worse fit than the 'DIF' model, $\chi^2(170) = 1741.56, p < .001$. Similarly, the 'no DIF' model had significantly worse fit compared to the 'DIF+covariates' model, $\chi^2(182) = 2272.19, p < .001$. For the 'DIF' model compared to the 'DIF+covariates', the latter had statistically significantly better fit according to the likelihood-ratio test, $\chi^2(12) = 530.63, p < .001$. This pattern of results was also confirmed by both the AIC and BIC (see Table 5).

On balance, it appears that the 'DIF+covariates' model has the best fit for each of the levels and languages, even when penalised for the extra parameters in the model⁸. Thus, given the superior fit of the DIF model versus the no DIF model in all cases, it appears that for all four Chemistry level and language combinations, there is substantial DIF at the overall examination level that warrants further evaluation at the item level, which will be presented in the next subsection. Moreover, given the superior fit of the 'DIF+covariates' model, the z-score DIF estimates from this model were used throughout the rest of the study for each of the Chemistry level-language combinations. The 'DIF+covariates' models' covariate estimates are presented in Appendix 3.

Item-Level DIF

Table 6 provides a breakdown of the frequencies and percentages of the different levels (small, moderate and large) of DIF observed in Chemistry for the different level-language and paper combinations. The complete set of DIF estimates may be viewed in Appendix 4 and the forest plots of the DIF estimates across the three papers for each subject are presented in Appendix 5.

⁸ The two partial exceptions to this were for SL and HL French vs. English models but given that both the likelihood-ratio tests and AICs provided support for the 'DIF+covariates' model, it was considered the best model for this cohort.

Table 6.

Frequencies (top tier) and percentages (bottom tier) of different DIF levels observed for the Chemistry items across the different level-language-paper combinations.

DIF Category								
	A-	A+	B-	B+	C-	C+	No	Total
Chemistry	109	114	41	46	36	36	163	545
HL	64	65	24	28	25	25	79	310
French	32	32	11	10	10	12	33	140
Paper 1	5	11	2	5	1	6	9	39
Paper 2	19	11	3	4	5	4	10	56
Paper 3	8	10	6	1	4	2	14	45
Spanish	32	33	13	18	15	13	46	170
Paper 1	3	10	1	9	0	2	14	39
Paper 2	15	7	3	4	5	9	13	56
Paper 3	14	16	9	5	10	2	19	75
SL	45	49	17	18	11	11	84	235
French	25	19	11	5	1	6	45	112
Paper 1	6	4	4	1	0	0	15	30
Paper 2	11	6	1	1	1	0	12	32
Paper 3	8	9	6	3	0	6	18	50
Spanish	20	30	6	13	10	5	39	123
Paper 1	4	3	2	3	1	1	16	30
Paper 2	4	11	0	5	2	2	8	32
Paper 3	12	16	4	5	7	2	15	61
	A-	A+	B-	B+	C-	C+	No	
Chemistry	20%	21%	8%	8%	7%	7%	30%	
HL	21%	21%	8%	9%	8%	8%	25%	
French	23%	23%	8%	7%	7%	9%	24%	
Paper 1	13%	28%	5%	13%	3%	15%	23%	
Paper 2	34%	20%	5%	7%	9%	7%	18%	
Paper 3	18%	22%	13%	2%	9%	4%	31%	
Spanish	19%	19%	8%	11%	9%	8%	27%	
Paper 1	8%	26%	3%	23%	0%	5%	36%	
Paper 2	27%	13%	5%	7%	9%	16%	23%	
Paper 3	19%	21%	12%	7%	13%	3%	25%	
SL	19%	21%	7%	8%	5%	5%	36%	
French	22%	17%	10%	4%	1%	5%	40%	
Paper 1	20%	13%	13%	3%	0%	0%	50%	
Paper 2	34%	19%	3%	3%	3%	0%	38%	
Paper 3	16%	18%	12%	6%	0%	12%	36%	
Spanish	16%	24%	5%	11%	8%	4%	32%	
Paper 1	13%	10%	7%	10%	3%	3%	53%	
Paper 2	13%	34%	0%	16%	6%	6%	25%	
Paper 3	20%	26%	7%	8%	11%	3%	25%	

Across the four level-language combinations, the HL models' items showed the most substantial amount of DIF, with 31% of the items found to have moderate or large DIF for the French vs. English model, and 36% of the items found to have moderate or large DIF for the Spanish vs. English model.

For the SL, most items were found to either have non-significant DIF or only small DIF effect sizes – 76% of SL Chemistry items were found to have either no statistically significant DIF or only small DIF, compared with 67% of HL Chemistry items and 71% of all Chemistry items. Overall, moderate DIF was observed for 16% of items and large DIF was found for 14% of items. This moderate and large DIF was observed in both multiple-choice (Paper 1) and constructed response (Papers 2 and 3) items, but on balance, the large DIF was more prevalent for the constructed response items.

Biology DIF by examination language

Exam-level DIF

For SL French vs. English, the 'no DIF' model had statistically significantly worse fit than the 'DIF' model, $\chi^2(99) = 423.61, p < .001$. Similarly, the 'no DIF' model had significantly worse fit compared to the 'DIF+covariates' model, $\chi^2(111) = 717.75, p < .001$. For the 'DIF' model compared to the 'DIF+covariates', the latter had statistically significantly better fit according to the likelihood-ratio test, $\chi^2(12) = 294.15, p < .001$. These results were also confirmed by the AIC. However, the BIC favoured the 'no DIF' model relative to the other two models (see Table 7 below).

For SL Spanish vs. English, the 'no DIF' model had statistically significantly worse fit than the 'DIF' model, $\chi^2(106) = 8910.49, p < .001$. Similarly, the 'no DIF' model had significantly worse fit compared to the 'DIF+covariates' model, $\chi^2(119) = 9530.349, p < .001$. For the 'DIF' model compared to the 'DIF+covariates', the latter had statistically significantly better fit according to the likelihood-ratio test, $\chi^2(13) = 619.86, p < .001$. This pattern of results was also confirmed by both the AIC and BIC (see Table 7).

For HL French vs. English, the 'no DIF' model had statistically significantly worse fit than the 'DIF' model, $\chi^2(104) = 455.93, p < .001$. Similarly, the 'no DIF' model had significantly worse fit compared to the 'DIF+covariates' model, $\chi^2(115) = 689.63, p < .001$. For the 'DIF' model compared to the 'DIF+covariates', the latter had statistically significantly better fit according to the likelihood-ratio test, $\chi^2(11) = 233.70, p < .001$. These results were also confirmed by the AIC. However, the BIC favoured the 'no DIF' model relative to the other two models (see Table 7).

Table 7.

Model comparison statistics for the models with and without DIF parameters, and with and without covariates for Biology across the different level and language combinations.

Level	Language	Model	Log Likelihood	AIC	BIC	No. of parameters
SL	French	No DIF	-411630.23	823604.45	824810.05	172
		DIF	-411418.42	823378.84	825278.37	271
		DIF + cov.	-411271.35	823108.70	825092.34	283
	Spanish	No DIF	-555730.70	1111841.40	1113229.53	190
		DIF	-551275.46	1103142.91	1105305.46	296
		DIF + cov.	-550965.53	1102549.05	1104806.58	309
HL	French	No DIF	-790484.32	1581364.65	1582834.76	198
		DIF	-790256.36	1581116.71	1583359.00	302
		DIF + cov.	-790139.51	1580905.01	1583228.98	313
	Spanish	No DIF	-860522.01	1721554.02	1723458.15	255
		DIF	-859280.66	1719341.32	1722253.51	390
		DIF + cov.	-859100.12	1719004.23	1722006.03	402

For HL Spanish vs. English, the ‘no DIF’ model had statistically significantly worse fit than the ‘DIF’ model, $\chi^2(135) = 2482.70$, $p < .001$. Similarly, the ‘no DIF’ model had significantly worse fit compared to the ‘DIF+covariates’ model, $\chi^2(147) = 2843.79$, $p < .001$. For the ‘DIF’ model compared to the ‘DIF+covariates’, the latter had statistically significantly better fit according to the likelihood-ratio test, $\chi^2(12) = 361.09$, $p < .001$. This pattern of results was also confirmed by both the AIC and BIC (see Table 7).

On balance, it appears that the ‘DIF+covariates’ model has the best fit for each of the levels and languages, even when penalised for the extra parameters in the model⁹. Thus, given the superior fit of the DIF model versus the no DIF model in all cases, it appears that for all four Biology level and

⁹ The two partial exceptions to this were for SL and HL French vs. English models but given that both the likelihood-ratio tests and AICs provided support for the ‘DIF+covariates’ model, it was considered the best model for this cohort.

language combinations, there is substantial DIF at the overall examination level that warrants further evaluation at the item level, which will be presented in the next subsection. Moreover, given the superior fit of the 'DIF+covariates' model, the z-score DIF estimates from this model were used throughout the rest of the study for each of the Biology level-language combinations. The 'DIF+covariates' models' covariate estimates are presented in Appendix 3.

Item-Level DIF

Table 8 below provides a breakdown of the frequencies and percentages of the different levels (small, moderate and large) of DIF observed in Biology for the different level-language and paper combinations. The complete set of DIF estimates may be viewed in Appendix 4 and the forest plots of the DIF estimates across the three papers for each subject are presented in Appendix 5.

Across the four level-language combinations, the SL models' items were found to have slightly more practically significant DIF, with 15% of the items found to have moderate or large DIF for the French vs. English model, and 26% of the items found to have moderate or large DIF for the Spanish vs. English model.

For the HL, most items were found to either have non-significant DIF or only small DIF effect sizes – 83% of HL Biology items were found to have either no statistically significant DIF or only small DIF, compared with 80% of SL Biology items and 82% of all Biology items. Overall, moderate DIF was observed for 10% of items and large DIF was found for 9% of items. This moderate and large DIF was observed in both multiple-choice (Paper 1) and constructed response (Papers 2 and 3) items, but on balance, the large DIF was slightly more prevalent for the constructed response items.

Table 8.

Frequencies (top tier) and percentages (bottom tier) of different DIF levels observed for the Biology items across the different level-language-paper combinations.

DIF Category								
	A-	A+	B-	B+	C-	C+	No	Total
Biology	97	111	22	21	17	21	155	444
HL	57	70	11	14	8	7	72	239
French	29	39	3	2	2	1	28	104
Paper 1	12	13	1	1	1	0	12	40
Paper 2	7	14	0	1	1	0	5	28
Paper 3	10	12	2	0	0	1	11	36
Spanish	28	31	8	12	6	6	44	135
Paper 1	8	6	5	3	0	3	15	40
Paper 2	9	13	0	1	1	0	4	28
Paper 3	11	12	3	8	5	3	25	67
SL	40	41	11	7	9	14	83	205
French	16	22	4	3	2	6	46	99
Paper 1	5	7	3	1	1	0	13	30
Paper 2	6	10	1	1	0	1	6	25
Paper 3	5	5	0	1	1	5	27	44
Spanish	24	19	7	4	7	8	37	106
Paper 1	9	2	3	2	6	0	8	30
Paper 2	8	7	2	0	1	2	5	25
Paper 3	7	10	2	2	0	6	24	51
	A-	A+	B-	B+	C-	C+	No	
Biology	22%	25%	5%	5%	4%	5%	35%	
HL	24%	29%	5%	6%	3%	3%	30%	
French	28%	38%	3%	2%	2%	1%	27%	
Paper 1	30%	33%	3%	3%	3%	0%	30%	
Paper 2	25%	50%	0%	4%	4%	0%	18%	
Paper 3	28%	33%	6%	0%	0%	3%	31%	
Spanish	21%	23%	6%	9%	4%	4%	33%	
Paper 1	20%	15%	13%	8%	0%	8%	38%	
Paper 2	32%	46%	0%	4%	4%	0%	14%	
Paper 3	16%	18%	4%	12%	7%	4%	37%	
SL	20%	20%	5%	3%	4%	7%	40%	
French	16%	22%	4%	3%	2%	6%	46%	
Paper 1	17%	23%	10%	3%	3%	0%	43%	
Paper 2	24%	40%	4%	4%	0%	4%	24%	
Paper 3	11%	11%	0%	2%	2%	11%	61%	
Spanish	23%	18%	7%	4%	7%	8%	35%	
Paper 1	30%	7%	10%	7%	20%	0%	27%	
Paper 2	32%	28%	8%	0%	4%	8%	20%	
Paper 3	14%	20%	4%	4%	0%	12%	47%	

DIF and other psychometric properties

As shown in Table 9 below, there was quite a systematic relationship between the DIF estimates and the difficulty and infit statistics for the Physics items, and this was particularly pronounced for the French vs. English models and for Paper 1, i.e., the multiple-choice items. These consistently negative correlations indicate that the items that show greater DIF in favour of the focal language (French or Spanish) tend to be the more difficult items, which also show lower discrimination compared to the average discrimination level of all items. The negative correlations were moderate to large for Paper 1 but tended to be small for Papers 2 and 3.

Table 9.

Correlations between the DIF estimates and the Physics items' difficulty and infit statistics across the different level-language-paper combinations.

Level – Language - Paper	Difficulty correlation	Infit correlation
SL		
French	-0.27	-0.26
Paper 1	-0.74	-0.58
Paper 2	0.02	-0.11
Paper 3	-0.25	-0.34
Spanish	0.04	-0.40
Paper 1	-0.48	-0.80
Paper 2	0.19	-0.45
Paper 3	0.15	-0.24
HL		
French	-0.42	-0.48
Paper 1	-0.94	-0.65
Paper 2	-0.23	-0.53
Paper 3	-0.34	-0.12
Spanish	-0.13	-0.29
Paper 1	-0.57	-0.60
Paper 2	0.05	-0.02
Paper 3	-0.03	-0.28

For the Chemistry items, the correlations between difficulty and infit and the DIF estimates were far less systematic, and particularly for the SL items, as shown in Table 10 below. The correlations were, again, consistently negative for the HL items and ranged from small to moderate in size. Unlike Physics, the sizes of the negative correlations were reasonably similar across all three papers.

Table 10.

Correlations between the DIF estimates and the Chemistry items' difficulty and infit statistics across the different level-language-paper combinations.

Level – Language - Paper	Difficulty correlation	Infit correlation
SL		
French	0.09	-0.15
Paper 1	0.12	-0.10
Paper 2	0.05	-0.16
Paper 3	-0.05	-0.21
Spanish	0.05	-0.29
Paper 1	-0.24	-0.46
Paper 2	0.13	-0.41
Paper 3	0.14	-0.22
HL		
French	-0.48	-0.38
Paper 1	-0.49	-0.48
Paper 2	-0.43	-0.18
Paper 3	-0.38	-0.46
Spanish	-0.34	-0.17
Paper 1	-0.16	-0.25
Paper 2	-0.25	-0.24
Paper 3	-0.39	-0.03

For the Biology items, the patterns of correlation between difficulty and infit statistics and the DIF estimates were somewhere in between the patterns for the other two subjects, as shown in Table 11 below. Overall, there was little evidence of a systematic relationship between DIF estimates and item difficulty, although there were moderate negative correlations for all three papers in the HL French vs. English model and for Paper 1 in the SL French vs. English model. There was a more systematic relationship between the DIF estimates and the infit statistics showing that items with larger DIF in favour of the focal languages had lower discrimination, and this was particularly the case for the multiple-choice items from Paper 1.

Table 11.

Correlations between the DIF estimates and the Biology items' difficulty and infit statistics across the different level-language-paper combinations.

Level – Language - Paper	Difficulty correlation	Infit correlation
SL		
French	-0.08	-0.25
Paper 1	-0.32	-0.41
Paper 2	0.02	-0.25
Paper 3	-0.12	-0.18
Spanish	0.11	-0.64
Paper 1	-0.19	-0.92
Paper 2	-0.17	-0.58
Paper 3	0.13	-0.51
HL		
French	-0.30	-0.37
Paper 1	-0.48	-0.63
Paper 2	-0.20	-0.28
Paper 3	-0.39	-0.31
Spanish	0.01	-0.25
Paper 1	0.03	-0.51
Paper 2	-0.14	-0.32
Paper 3	0.05	-0.13

Discussion

Generally, the findings from Phase 1 of this research project are very positive for the IB's current translation processes of DP Science examinations from English to French and Spanish. Nonetheless, in response to research questions 1 and 2 of this phase of the research, the analyses showed that the 'DIF+covariates' model was, on balance, the best fitting model across all subject-level-language combinations, and a small but substantial proportion of items showed moderate and large DIF across the subjects, which 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. Moreover, regarding research question 3, there was a general trend that the items that showed significant DIF in favour of the focal languages tended to be the more difficult and less discriminating items, and this was particularly the case for the multiple-choice items.

The systematic relationship between the DIF estimates and these other psychometric properties of the items provide 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, the large negative correlations observed between difficulty and infit statistics and the DIF estimates for the multiple-choice items

suggest that some of this DIF may be attributable to guessing behaviour, particularly as the students from the focal languages tended to be, on average, lower performing across the subjects. For the constructed response items, the smaller but still significant negative correlations may be attributable to other fit issues with the items.

A key limitation of the current analyses were the small sample sizes for the French cohort in some of the subjects, and particularly Physics, which means that those results should be interpreted with caution. This is particularly the case for Paper 3, which contains a high degree of optionality and so many items that were not deleted or collapsed because they met the criteria of having at least five observations for each category but still had relatively small numbers of observations in their scoring categories. Moreover, although we have included what we believe are the most salient covariates in the 'DIF+covariates' models, it is possible that the DIF observed in terms of the examination language groups may be a proxy for other causal factors, such as, for example, systematic differences in the socio-economic backgrounds of students taking the examinations in the focal languages. Nonetheless, there was an extremely high correlation in the DIF estimates between the two DIF models with and without covariates, which provides confidence in the robustness of the estimates.

Overall, the results of Phase 1 are a good news story for IB and its translation processes in the DP Science examinations, as they have stood up well to the scrutiny of some intensive statistical evaluations of their comparability across these three languages. Nonetheless, the analyses have identified many items that show moderate to large DIF across the languages, which were further investigated both in terms of expert qualitative evaluation of their comparability according to key linguistic and translation criteria, and quantitative comparisons by way of differences in NLP features across the source and target languages to establish an explanatory model in Phases 2 and 3 of the research for the systematic differences in difficulty across the languages observed for some items in the current phase.

Due to time and resource constraints, it was not possible to conduct Phase 2's expert review and Phase 3's explanatory modelling on all examinations and items within the examinations, a subset of examinations and items were selected for these further phases. It was important to maintain the representation of all three subjects (Physics, Chemistry and Biology) and so one level (Standard or Higher) was selected from each subject. For Physics, the SL was selected as the sample size for the French cohort was larger and so their DIF estimates will be more reliable. For Chemistry, both levels had adequate sample sizes for both the focal language groups, and so the HL was selected, as more DIF was observed for this exam. Moreover, selection of the HL for this subject means that both levels continue to be represented in the later phases of the research. For Biology, the SL was selected, as,

again, the sample sizes were adequate and more DIF was observed for this level. The rationale for the selection of the specific items for expert review from these three subject-level combinations is provided in the next section.

Phase 2: Expert Review of selected DIF items

This section overviews the expert review study that was carried out to evaluate the comparability of a subset of the French and Spanish (target) versions of the DP science items to the English (source) version. The main aim of the expert review was to identify possible variables that potentially explain the DIF observed in Phase 1 of the research.

Phase 2 addressed the following research questions:

- Are there differences in terms of key linguistic and translation criteria in the translated versions of DP Science examination questions that show differential difficulty across the source and target languages?
- Are the expert reviewers able to reliably evaluate differences in the source and target language versions of the DP Science examination questions using the newly developed survey and expert review framework?

The following section describes the methods adopted in the expert review study, which included the selection of the subsets of items, the recruitment of expert reviewers, and the development of a survey instrument used by the experts to review the source and target versions of the items.

Methods

Item selection

Items were selected from all three papers (1, 2 and 3) of the 2019 Physics SL, Chemistry HL and Biology SL examinations. Thirty-five to 40 items were selected for both the French and Spanish versions (71 to 79 per exam) of the three examinations according to the following criteria:

- *Text heaviness of the item:* Select items that are most text-heavy to maximise the chance that the DIF is due to language issues.
- *A balance of items taken from the 3 papers:* Given that Paper 1 is a multiple-choice paper and so inherently less text-heavy in comparison with the constructed response items in Papers 2 and 3, items from Paper 1 should be judiciously chosen.
- *Range in DIF magnitude:* All the items included have substantial DIF, meaning the items categorized as having small, moderate or large DIF, and while the focus is on the inclusion of items with moderate and large DIF, the sample also includes small DIF items so that there is a full representation of different DIF levels to avoid restriction of range issues in the Phase 3 modelling.
- *A balance in the direction of DIF:* Select, as appropriate, a comparable number of DIF items in each subject that are advantaging and disadvantaging the target language (i.e., + DIFs \approx - DIFs)

Table 12 below details the characteristics of the items selected across the three subjects, showing the prevalence of items selected from the three papers, the two item-types, and the six substantial DIF categories.

Table 12.
Characteristics of selected items for the three subjects.

		Bio SL		Chem HL		Phys SL	
		French	Spanish	French	Spanish	French	Spanish
No. items		35	40	40	39	36	35
Paper	1	12	15	15	6	10	8
	2	16	13	19	19	7	9
	3	7	12	6	14	19	17
Type of item	MCQ	12	15	15	6	10	8
	CR	23	25	25	33	26	26
DIF category	A-	15	13	7	3	13	10
	A+	10	12	5	2	13	13
	B-	5	3	6	9	5	5
	B+	3	3	9	6	2	4
	C-	0	3	8	9	3	3
	C+	2	6	5	10	0	0

Expert reviewers

Ten expert reviewers were recruited in collaboration with the IB to evaluate the comparability of translated versions of the items to the English source version; two in each language-subject combination. The expert reviewers included IB subject specialists who teach/examine or have recently taught/examined DP science and two Oxford researchers. All expert reviewers were at least bilingual, i.e., they had a high proficiency in English and one of the two target languages. Two reviewers were trilingual and acted as expert reviewers for the French and Spanish versions relative to the English source versions of the Chemistry and Physics items, respectively.

The instrument

The instrument is a 14- to 15- item survey (See Appendix 6) developed based on a framework adapted from cApStAn's translation/verification framework for evaluating test translation (Ferrari & Dept, 2020) and informed by El Masri et al.'s research (2016; 2017) on the translation and adaption of science examination questions and predicting item difficulty (see Table 13). The adapted framework initially included seven variables; however, it was refined together with the instrument following feedback from IB and the subsequent addition of the 'house style' variable.

Data collection

Expert reviewers were required to complete a survey in which they were asked to make judgements on the eight variables of the adapted framework by comparing the target version (French or Spanish) of the pre-selected items (and stimulus material if applicable) to the source version (English) of the items (and stimulus material if applicable). Expert reviewers were invited to an online workshop (2 hours) offered on two different dates to accommodate the experts' timetables and time zones. The workshop session included an introduction to the framework and a description of the variables followed by a training session where the reviewers completed a demo version of the online instrument and rated three example items (French or Spanish versions) that varied in question type and subject.

The survey was administered online via the Qualtrics software. Reviewers were provided with individual links that gave them access to the instrument and the set of items that matched their subject and language expertise. The instrument included the English version of a subset of examination items and their mark schemes as well as the French or Spanish version of the same items. Experts reviewing the Spanish version of the biology items were also provided with the corresponding mark schemes in Spanish, however, translated mark schemes were not available for the other subjects. The examination items were followed by the survey questions (see Appendix 6). Reviewers reviewed one examination item at a time and were required to complete all survey questions for each examination item they rated. The survey allowed reviewers to pause the rating whenever they needed to and complete the survey at a later time.

Table 13.
Description of the eight variables targeted in the survey.

Variable	Description
Added/ Missing information	Added: Information is present in the target version but not in the source version (e.g., explanation between brackets). Missing: Information present in the source version but omitted in the target version.
Matches & Patterns	1. A literal match (repetition of the same word or phrase) or a synonymous match (use of a synonym or a paraphrase) in the source version is not reflected in the target version. Most important: literal or synonymous match between stimulus and item and between a question stem and response options. 2. A pattern in multiple choice items is not reflected in the target version (e.g., all but one response option start with the same word, or proportional length of responses options.)
Register/ Wording	This category is typically used for vague or inaccurate, or not quite fluent translations. 1. Register: difference in the level of terminology (scientific item vs. familiar item) or level of language (formal vs. casual; standard vs. idiomatic) in the target version versus the source version. 2. Wording: inappropriate or less than optimal choice of vocabulary or wording in target to fluently convey the same information as in the source version.
Grammar/ Syntax	1. Grammar: Grammatical mistake that could affect comprehension or equivalence (e.g., incorrect subject-verb agreement) 2. Syntax: Syntax-related deviation from the source, e.g., a long (source) sentence is split into two (target) sentences or two (source) sentences are merged into a single (target) one; or another syntactic problem due to, e.g., overly literal translation of the source.
Layout/ Format consistency	A deviation or defect in layout or formatting: disposition of text and graphics, item labels, question numbering, styles (boldface , <u>underlining</u> , <i>italics</i> , UPPERCASE), legibility of captions, tables, number formatting (decimal separators, “five” vs. “5”), etc.
House style*	House style is a set of rules adopted by an organization that define how all assessments should be formatted (e.g., font style, size, language conventions, etc.). The application of house style could lead to awkward wording, confusion and/or more demand in the target version of an item.
Depth of knowledge*	Translation of the command terms (or the stem in a multiple-choice item) leads to an item that is of a greater or lesser conceptual depth than in the target version.
Mark scheme*	1. The source version of the mark scheme is not translated into the target version potentially leading to a misinterpretation or more or less severe interpretation of the mark scheme by the marker of the target version. 2. The translation of the item’s mark scheme is poor (vague, inaccurate or not fluent). 3. The translation of the item’s mark scheme and expected response leads to a more, or less cognitively demanding item.

* Not part of cApStAn’s verification framework.

Analysis

Percentage agreement between two expert reviewers was computed for each subject-language combination and is reported at the overall survey level and at the survey item level. Later, for all but question 12 in the survey, disparate judgements were reconciled by taking the average of the two judgements. Given that the scale of response type in question 12 was nominal rather than ordinal, reconciliation of expert reviewers' judgements required the adjudication by one of the researchers who selected the most appropriate response type based on her judgement.

To assess the extent to which the target version of each selected item deviated from the source version, the distance between the mean rating of each survey item and the 'neutral response category' (i.e., the response category in the survey describing the language versions of the items as the same or very similar for that criterion) was computed. The next section outlines the results of these analyses.

Results

Percentage agreement

The overall percentage agreement between expert reviewers was substantial to high ranging between 74.51% to 86.15% across the subject-language combinations. The mean of the absolute difference between judgements at survey level was small ranging between 0.16 and 0.28 (see Table 14).

The consistency between reviewers at survey level for all six subject-language combinations provides a reassuring picture in terms of the functioning of the survey. Nonetheless, it is worth noting that the consistency of expert reviewers' judgements varied across survey questions across different subjects with the percentage agreement lower than 70% across several subject-language combinations in question 5 (wording) and question 8 (length of clauses) and with mean difference of rating exceeding 0.5 in many cases (i.e., the mid-point between two consecutive response categories in the survey). This is partly due to challenges that reviewers seemed to have experienced when making judgements about two consecutive response categories such as *somewhat accurately* or *mostly accurately* (referring to wording) or between *somewhat longer* or *much longer* (referring to length of clauses). Moreover, reviewers may have adopted different ways of judging the length of a clause with some counting words while others may have been relying on visual comparisons and some comparing the lengths of sentences rather than clauses.

Table 14.

Percentage agreement and mean difference of rating at subject-language combination.

Variable code (survey question number)	Biology (SL)				Chemistry (HL)				Physics (SL)			
	French		Spanish		French		Spanish		French		Spanish	
	% agree	Mean diff	% agree	Mean diff	% agree	Mean diff	% agree	Mean diff	% agree	Mean diff	% agree	Mean diff
info (q1)	85.71	0.14	100.00	0.00	87.50	0.15	92.31	0.08	80.56	0.19	71.43	0.29
matpat (q2)	85.71	0.20	77.50	0.23	75.00	0.30	71.79	0.38	61.11	0.42	40.00	0.71
sreg (q3)	94.29	0.06	85.00	0.15	100.00	0.00	89.74	0.10	86.11	0.14	97.14	0.03
freg (q4)	82.86	0.17	82.50	0.18	80.00	0.20	84.62	0.15	100.00	0.00	97.14	0.03
word (q5)	74.29	0.37	70.00	0.40	62.50	0.38	53.85	0.67	55.56	0.50	54.29	0.66
gram (q6)	100.00	0.00	100.00	0.00	95.00	0.05	87.18	0.13	94.44	0.06	91.43	0.09
nclaus (q7)	100.00	0.00	100.00	0.00	100.00	0.00	100.00	0.00	100.00	0.00	91.43	0.09
lclaus (q8)	42.86	0.60	65.00	0.35	82.50	0.18	56.41	0.44	55.56	0.44	62.86	0.37
layfor (q9)	88.57	0.11	92.50	0.08	82.50	0.20	94.87	0.05	66.67	0.33	65.71	0.40
house (q10)	57.14	1.00	95.00	0.05	85.00	0.20	82.05	0.26	88.89	0.11	80.00	0.20
dok (q11)	57.14	0.43	65.00	0.38	90.00	0.10	92.31	0.08	97.22	0.03	85.71	0.14
msres (q12)	85.71	0.14	100.00	0.00	82.50	0.25	58.97	0.59	77.78	0.44	77.14	0.46
mseng (q13)	94.29	0.06	100.00	0.00	97.50	0.03	56.41	0.33	94.44	0.06	54.29	0.20
msacc (q14)			60.00	0.48								
msdem (q15)			80.00	0.20								
Overall	80.66	0.25	84.83	0.17	86.15	0.16	78.50	0.25	81.41	0.21	74.51	0.28

Info = added/missing information; matpat = matches and patterns; sreg = scientific register; freg = formal register; word = accuracy of wording; gram = grammatical mistakes; nclaus = number of clauses; lclaus = length of clauses; layfor = layout/format; house = house style; dok = depth of knowledge; msres = type of response expected in mark scheme; mseng = use of english source of mark scheme; msacc = level of accuracy of mark scheme; msdem = conceptual demand of the mark scheme.

Another question that had a lower percentage agreement, especially in Chemistry HL for the Spanish target language and to a lesser extent Physics SL was question 12, for type of response expected. There are multiple possible explanations for the inconsistency observed; one being making judgements based on the source rather than the target version. This has sometimes led reviewers to select different response categories. For example, the response of one of the questions in physics was *magnifying lens* (i.e., a phrase) in English and *loupe* (i.e., a word) in French. Moreover, some responses in Chemistry and Physics (e.g., chemical formulas and equations) were hard to match to any of the eight response categories included in question 12 of the survey. Moreover, expert reviewers did not agree on the way to interpret the length of an expected response based on the mark scheme provided. Mark schemes provide elements of a correct response that are awarded credit and these could constitute one sentence or several based on the judgement of the reviewer (see Appendix 8 for an example). This would hence lead to different response categories in the survey.

Distance of reviewers' judgements from the neutral category

The extent to which the target versions of the science examination items were comparable to the English source version was evaluated at survey item level for each subject-language combination by comparing the extent to which the average rating between the two expert reviewers deviated from the neutral response category (i.e., the response category in the survey indicating no differences between the source and target versions) for each variable. The variable 'type of response expected in the mark scheme' (msres) will not be discussed further in this section since it is a nominal rather than an ordinal variable. In addition, given the reason outlined in the previous section, the results of the 'length of clauses' (lclaus) variable will not be commented on in this section. The results for each subject-language combination are presented below with illustrative examples for some variables where the distance from the neutral category was above 0.20 or below -0.20.

Biology SL French

As presented in Table 16, the French versions of the selected 35 biology items were highly comparable to their respective English versions. For most variables targeted by the survey questions, the distance between expert reviewers' rating and the neutral category was less than 0.10 in absolute value (Table 15). For four variables, the distance exceeded 0.20 (in absolute value): matches and patterns, length of clauses, house style and depth of knowledge.

Table 15.

Distance from neutral category at survey item level in the French/English comparison of Biology SL items.

Survey question	number of categories	neutral category	mean rating	distance from neutral category
<i>info</i>	5	3	2.99	-0.01
<i>matpat</i>	4	4	3.76	-0.24
<i>sreg</i>	3	2	2.03	0.03
<i>freg</i>	3	2	2.09	0.09
<i>word</i>	4	4	3.81	-0.19
<i>gram</i>	3	1	1.00	0.00
<i>nclaus</i>	3	2	2.00	0.00
<i>lclaus</i>	5	3	3.40	0.40
<i>layfor</i>	4	4	3.94	-0.06
<i>house</i>	4	4	3.50	-0.50
<i>dok</i>	5	3	3.21	0.21
<i>mseng</i>	4	1	1.03	0.03

The following discussion provides examples of Biology SL items where the French version was less comparable to the source version based on the expert reviewers' judgement on some of the variables outlined above.

Matches and patterns

Matches and patterns are nearly exactly the same across versions of items. Some exceptions include the item in Figure 1 below.

Figure 1. Comparability of the English and French versions of item 25 of Paper 1, Biology SL in terms of matches and patterns.

25.	What causes the <u>atrioventricular</u> valves to close during a heartbeat?
A.	Pressure in the <u>atria</u> is higher than in the ventricles.
B.	Pressure in the atria is lower than in the ventricles.
C.	Pressure in the arteries is higher than in the ventricles.
D.	Pressure in the arteries is lower than in the ventricles.
25.	Qu'est-ce qui provoque la fermeture des valvules <u>auriculo-ventriculaires</u> durant un battement cardiaque ?
A.	La pression dans les <u>oreillettes</u> est plus élevée que dans les ventricules.
B.	La pression dans les oreillettes est plus faible que dans les ventricules.
C.	La pression dans les artères est plus élevée que dans les ventricules.
D.	La pression dans les artères est plus faible que dans les ventricules.

In the item above, the matching between *atrioventricular* in the stem and *atria* in the first and second options are not mirrored in the French version of the item. The term *auriculo-ventriculaires* is not partially composed of the term *oreillette(s)* (meaning atrium in French). The absence of match between stem and option in this multiple choice is not due to poor translation. Translators had to use the correct scientific terms when translating atrioventricular and atria into French to maintain the level of formality and scientific register of the question in the source version. Not maintaining matches and patterns in the target version could potentially lead to higher language demands for students in the target group. However, in this particular item, the impact is likely to be small as students tend to focus on scientific vocabulary (Halliday & Martin, 1993). Indeed, the DIF magnitude was small and in favour of English language group (category A+).


House Style

The application of house style results in translations that are mostly to entirely clear (rating = 3.50). Some unusual translations include command terms such as the command verb *state* is translated into *exprimez* which means 'express' rather than *nommez* or *identifiez* in item 4a of Paper 2 of Biology SL (Figure 2). Other items include *outline* in English being translated into *résumez*, which means 'summarise'. *Summarise* has a different nuance to the verb *outline*. Also, the verb *sketch* is translated in some items into *représentez*, with the latter not necessarily referring to the production of a drawing. The question in Figure 2 was flagged for DIF with small magnitude (A+). It is hard to confirm whether the command term *state* provided a slight advantage to students completing the question in English without collecting additional data such as students' responses to this question.

Figure 2. Comparability of the English and French versions of item 4a of Paper 2, Biology SL in terms of command terms.


4. (a) The images show parts of plants belonging to two different phyla.

Plant X



[Source: <https://hiveminer.com>]

Plant Y

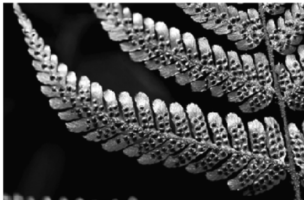


[Source: <https://classconnection.s3.amazonaws.com>]

State the phylum of plant X and of plant Y. [2]


l. (a) Les images représentent des parties de plantes appartenant à deux embranchements différents.

Plante X



[Source : <https://hiveminer.com>]

Plante Y



[Source : <https://classconnection.s3.amazonaws.com>]

Exprimez l'embranchement de la plante X et celui de la plante Y. [2]

Biology SL Spanish

The results in Table 5 suggest that the Spanish versions of the selected 40 biology items were highly comparable to their respective English versions. For most variables targeted by the survey questions, the distance between expert reviewers' rating and the neutral category was less than 0.10 in absolute value (Table 16). For three variables, the distance exceeded 0.20 (in absolute value): wording, length of clauses and mark scheme accuracy.

Table 16.

Distance from neutral category at survey item level in the Spanish/English comparison of Biology SL items.

Survey question	Number of categories	neutral category	mean rating	distance from neutral category
<i>Info</i>	5	3	3.00	0.00
<i>matpat</i>	4	4	3.84	-0.16
<i>Sreg</i>	3	2	2.08	0.08
<i>Freg</i>	3	2	2.09	0.09
<i>Word</i>	4	4	3.78	-0.23
<i>Gram</i>	3	1	1.00	0.00
<i>Nclaus</i>	3	2	2.00	0.00
<i>Lclaus</i>	5	3	3.28	0.28
<i>Layfor</i>	4	4	3.96	-0.04
<i>House</i>	4	4	3.98	-0.02
<i>Dok</i>	5	3	3.18	0.18
<i>Mseng</i>	4	1	1.00	0.00
<i>Msacc</i>	4	4	3.34	-0.66
<i>Msdem</i>	5	3	3.04	0.04

The following discussion provides examples of Biology SL items where the Spanish version was less comparable to the source version based on the expert reviewers' judgement on some of the variables outlined above.

Wording and mark scheme accuracy

The choice of wording in the Spanish version conveys the meaning of the information in the source version entirely accurately. However, some translated items and their mark scheme were judged as *somewhat accurate*. Below is an example of a Biology SL item and its mark scheme where the term *stroke*, which in the context of the question means a sudden interruption of blood flow to the brain, was translated to *traumatismo* in Spanish, which refers to *traumatism*. *Trauma* is a more general term which does not necessarily refer to a brain injury and could refer to other serious physical injuries or deep psychological distress. Translating a term with a specific connotation in one language to another one with a broader meaning could be misleading to some students.

Figure 3. Comparability of the English and Spanish versions of item 4c of Paper 3, Biology SL in terms of accuracy of wording and mark scheme.

Item

(c)	It has been suggested that cinnamon might be of benefit to patients who are recovering from a stroke. Suggest one advantage of adding cinnamon to the diet of a patient who has suffered a stroke.	[1]
(c)	Se ha sugerido que la canela podría ser beneficiosa para los pacientes que se están recuperando de un traumatismo. Sugiera una ventaja de la adición de canela a la dieta de un paciente que haya sufrido un traumatismo.	[1]

Mark scheme

4.	c	a. reorganization of brain function through plasticity «which is enhanced by cinnamon» ✓ b. cinnamon helps to form new neural pathways to replace the ones that were lost «due to the stroke» ✓	1 max
4.	c	a. función de reorganización del cerebro en virtud de la plasticidad «que se ve incrementada por la canela» ✓ b. la canela ayuda a formar nuevas rutas neuronales para sustituir a las que se han perdido «debido al traumatismo» ✓	1 máx.

Another translation issue that could emerge in this item relates to other connotations of the term *stroke*. It is a term used in everyday language with multiple meanings including the act of hitting someone or something, the mark left by a pen or a paintbrush, the act of moving one's hand over a surface with light pressure, etc. Scientific terms that have everyday language meanings can often result in ambiguities for some students especially readers with low proficiency. The item in Figure 3 was flagged for large DIF favouring students completing the Spanish version of the item (C-). With the current data only, it is hard to attribute the large DIF and its direction simply to the inaccuracy of the term or the multiple meaning that the term *stroke* has. In this case, it would have helped to examine students' responses and see whether the translation of *stroke* into *traumatismo* provided an advantage to Spanish speakers or whether other non-linguistic factors were involved.

Chemistry HL French

The French versions of the selected 40 Chemistry HL items were highly comparable to their respective English versions. The distance between expert reviewers' rating and the neutral category was less than 0.10 in absolute value for most variables targeted by the survey questions (Table 17) and was higher than 0.20 for only one variable – wording.

Table 17.

Distance from neutral category at survey item level in the French/English comparison of chemistry items.

Survey question	number of categories	neutral category	mean rating	distance from neutral category
<i>info</i>	5	3	3.03	0.02
<i>matpat</i>	4	4	3.83	-0.18
<i>sreg</i>	3	2	2.00	0.00
<i>freg</i>	3	2	2.08	0.08
<i>word</i>	4	4	3.76	-0.24
<i>gram</i>	3	1	1.03	0.02
<i>nclaus</i>	3	2	2.00	0.00
<i>lclaus</i>	5	3	3.14	0.14
<i>layfor</i>	4	4	3.90	-0.10
<i>house</i>	4	4	3.88	-0.13
<i>dok</i>	5	3	3.03	0.02
<i>mseng</i>	4	1	1.01	0.01

Below is an example of an item where the French translation was rated as mostly (rather than entirely) accurate. This is likely due the addition of the information *l'ion* (the ion) in the French version of the item, which did not appear in the English version. This addition might provide additional cues to students completing the French version of the item by activating the correct schemas. Indeed, this item was flagged for large DIF in favour of the French-speaking group (C-). However, in the absence of additional information, it is difficult to establish that the addition of the term *ion* in the French version has provided an advantage for students sitting the item in French. DP students sitting a HL examination in Chemistry are less likely to need to be primed with the term *ion* preceding the symbol of the ion.

Figure 4. Comparability of the English and French versions of item 4 of Paper 1, Chemistry HL in terms of accuracy of wording.

4. Which is correct for $^{34}_{16}\text{S}^{2-}$?			
	Protons	Neutrons	Electrons
A.	16	18	14
B.	18	16	18
C.	16	18	16
D.	16	18	18

4. Quels nombres correspondent à l'ion $^{34}_{16}\text{S}^{2-}$?			
	Protons	Neutrons	Électrons
A.	16	18	14
B.	18	16	18
C.	16	18	16
D.	16	18	18

Another distinction between the two items is in the amount of information in the stem in each version of the item. The back-translation of the French version will be *Which numbers correspond to the ion $^{34}_{16}\text{S}^{2-}$?* There is certainly more information in the French version than the English version but again it is hard to establish whether that additional information led to a substantial advantage for French-speaking examinees.

Chemistry HL Spanish

The Spanish versions of the selected 39 Chemistry items were highly comparable to their respective English versions. The distance between expert reviewers' rating and the neutral category was less than 0.10 in absolute value for most variables targeted by the survey questions and was higher than 0.20 (in absolute value) for three variables: matches and patterns, accuracy of wording and reliance on the English version of the mark scheme (Table 18).

Table 18.

Distance from neutral category at survey item level in the Spanish/English comparison of chemistry items.

Survey question	number of categories	neutral category	mean rating	distance from neutral category
<i>info</i>	5	3	3.01	0.01
<i>matpat</i>	4	4	3.71	-0.29
<i>sreg</i>	3	2	2.00	0.00
<i>freg</i>	3	2	2.00	0.00
<i>word</i>	4	4	3.46	-0.54
<i>gram</i>	3	1	1.06	0.06
<i>nclaus</i>	3	2	2.00	0.00
<i>lclaus</i>	5	3	3.12	0.12
<i>layfor</i>	4	4	3.95	-0.05
<i>house</i>	4	4	3.87	-0.13
<i>dok</i>	5	3	3.04	0.04
<i>mseng</i>	4	1	1.28	0.28

Matches and patterns

Matches and patterns were judged to be maintained during the translation from English to Spanish in most Chemistry HL items.

Wording

The choice of wording in the Spanish versions conveys the meaning of the information in the source version mostly to entirely accurately. An example of an item where Spanish translation was judged as *inaccurate* is shown in Figure 5.

Figure 5. Comparability of the English and Spanish versions of item 2e of Paper 2, Chemistry HL in terms of accuracy of wording.

- | | |
|--|-----|
| <p>(e) The experiment gave an error in the rate because the pressure gauge was inaccurate. Outline whether repeating the experiment, using the same apparatus, and averaging the results would reduce the error.</p> | [1] |
| <p>(e) El experimento dio un error en la velocidad porque el manómetro era impreciso. Resume si repetir el experimento, usando el mismo aparato, y promediar los resultados reduciría el error.</p> | [1] |

The term *inaccurate* in English was translated into *impreciso* in Spanish. Accuracy and precision refer to different attributes of measurement where *accuracy* represents the difference between a measurement taken of a dimension and the real value of the dimension, while *precision* describes the variation recorded in the measurement of the same dimension repeatedly with the same equipment. For science students, a question about accuracy will lead to a different answer than a question about precision. It is clear from the rest of the question in English and the mark scheme of the item (Figure 6) that the response related to *precision* and repetition of measurement rather than its *accuracy*. The DIF manifested in this item is of a large magnitude and in favour of students completing the test in Spanish (C-) suggesting that the inaccurate term in English could have introduced an ambiguity that Spanish-speaking students did not have to deal with.

Figure 6. Mark scheme of item 2e of Paper 2, Chemistry HL.

2.	e	no AND it is a systematic error/not a random error OR no AND «a similar magnitude» error would occur every time ✓	French and Spanish: Accept "yes AND precision can be improved by repeating trials".	1
----	---	--	---	---

Reliance on the English version of the mark scheme

The ambiguity between the terms *accuracy* and *precision* in the item in English can carry on into the interpretation of the mark scheme. A note in the mark scheme (Figure 6) suggests that answers in Spanish referring to precision and repetition of measurement are awarded full credit but it is unclear whether such a response is only acceptable for examinees taking the item in French and Spanish or whether this response is also acceptable for students sitting the item in English.

Physics SL French

The French versions of the selected 36 Physics SL items were highly comparable to their respective English versions. The distance between expert reviewers' rating and the neutral category was less than 0.10 in absolute value for most variables targeted by the survey questions (Table 19) and the distance exceeded 0.20 (in absolute value) for three variables: matches and patterns, wording and length of clauses.

Table 19.

Distance from neutral category at survey item level in the French/English comparison of Physics items

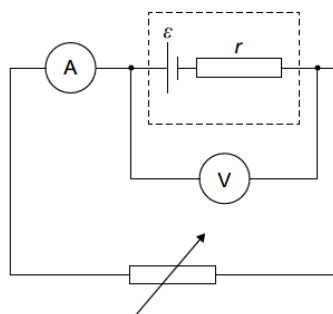
Survey question	number of categories	neutral category	mean rating	distance from neutral category
<i>info</i>	5	3	3.10	0.10
<i>matpat</i>	4	4	3.74	-0.26
<i>sreg</i>	3	2	1.93	-0.07
<i>freg</i>	3	2	2.00	0.00
<i>word</i>	4	4	3.64	-0.36
<i>gram</i>	3	1	1.03	0.03
<i>nclaus</i>	3	2	2.00	0.00
<i>lclaus</i>	5	3	3.25	0.25
<i>layfor</i>	4	4	3.83	-0.17
<i>house</i>	4	4	3.94	-0.06
<i>dok</i>	5	3	2.99	-0.01
<i>mseng</i>	4	1	1.03	0.03

It was not possible based on the survey data to understand why some expert reviewers judged that some translated items did not maintain the matches and patterns in the source version. In any case, the deviation from the neutral category does not exceed 0.30, so overall, matches and patterns are nearly exactly the same across the English and French versions of items.

Based on the data in Table 19, the choice of wording in the French version conveys the meaning of the information in the source version mostly accurately. The example in Figure 7 is of an item judged to have a wording accuracy between inaccurate and somewhat accurate. Examining the item closely suggests that this could be due to the use of the past tense in English, “*The current I and the terminal potential difference V are measured*”, which was translated into an active voice in French referring to the student as doing the action of measurement. The use of passive voice can increase the cognitive demand of text. Another possible source of lower accuracy in the wording goes back to the command terms and the translation of *Outline* into *Résumez* as discussed earlier (Biology SL French section). In this case, it is unclear how such inconsistencies could lead to significant differences in students’ performance. Indeed, this item has a small DIF magnitude favouring French speaking students (A-) and the survey data do not provide a compelling case to explain the DIF on translation grounds.

Figure 7. Comparability of the English and French versions of item 1a of Paper 3, Physics SL.

1. A student investigates the electromotive force (emf) ε and internal resistance r of a cell.



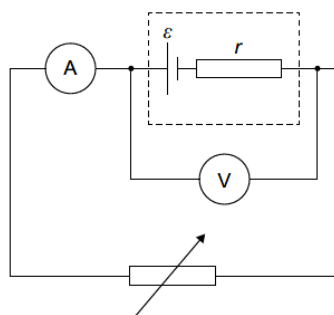
The current I and the terminal potential difference V are measured

For this circuit $V = \varepsilon - Ir$.

[Part missing for space considerations]

- (a) The student has plotted error bars for the potential difference. Outline why no error bars are shown for the current.

1. Un élève effectue une recherche sur la force électromotrice (f.é.m.) ε et sur la résistance interne r d'une pile.



Il mesure le courant I et la différence de potentiel aux bornes V .

Pour ce circuit, $V = \varepsilon - Ir$.

[Part missing for space considerations]

- (a) L'élève a tracé des barres d'erreur pour la différence de potentiel. Résumez pourquoi aucune barre d'erreur n'est montrée pour le courant.

[1]

The Spanish versions of the selected 35 Physics items were highly comparable to their respective English versions. The distance between expert reviewers' rating and the neutral category was less than 0.10 in absolute value for most variables targeted by the survey questions (Table 20) and the distance exceeded 0.20 (in absolute value) for five variables: matches and patterns, wording, length of clauses, layout and format and reliance on the English version of the mark scheme.

Table 20.
Distance from neutral category at survey item level in the Spanish/English comparison of Physics items.

Survey question	number of categories	neutral category	mean rating	distance from neutral category
<i>info</i>	5	3	3.09	0.09
<i>matpat</i>	4	4	3.59	-0.41
<i>sreg</i>	3	2	1.99	-0.01
<i>freg</i>	3	2	1.99	-0.01
<i>word</i>	4	4	3.67	-0.33
<i>gram</i>	3	1	1.04	0.04
<i>nclaus</i>	3	2	2.01	0.01
<i>lclaus</i>	5	3	3.47	0.47
<i>layfor</i>	4	4	3.77	-0.23
<i>house</i>	4	4	3.90	-0.10
<i>dok</i>	5	3	3.07	0.07
<i>mseng</i>	4	1	1.27	0.27

Matches and patterns

Matches and patterns are nearly exactly the same across English and Spanish versions of items. As outlined in the previous section, a closer look at the items judged to have less comparable versions of matches and patterns was hard to interpret based on the survey data.

Wording

Similar to other subject-language combinations, the Spanish versions of the Physics SL items convey the meaning of the information in the source version entirely accurately except for some translated versions judged as having *somewhat accurate* wording.

Layout and formatting

The layout and formatting of the Spanish versions of the items were judged to be *mostly* to *exactly the same* to the respective English versions for most items except for three items for which the translations were judged to be *mostly the same*. Figure 8 provides an illustration of such an item. The difference in layout refers to the equation in English is embedded in the sentence while it is not in

Spanish and follows a colon. Once again, it is difficult to conclude that such difference in layout could have resulted in a difference in performance across language groups. Indeed, the DIF is once again small, advantaging the English-speaking students (A+).

Figure 8. Comparability of the English and Spanish versions of item 26 of Paper 1, Physics SL.

<p>26. Three conservation laws in nuclear reactions are</p> <ul style="list-style-type: none"> I. conservation of charge II. conservation of baryon number III. conservation of lepton number. <div style="border: 1px solid red; padding: 5px; margin: 10px 0;"> <p>The reaction</p> $n \rightarrow \pi^- + e^+ + \bar{\nu}_e$ <p>is proposed.</p> </div> <p>Which conservation laws are violated in the proposed reaction?</p> <ul style="list-style-type: none"> A. I and II only B. I and III only C. II and III only D. I, II and III 	<p>26. Tres leyes de conservación para las reacciones nucleares son:</p> <ul style="list-style-type: none"> I. la conservación de la carga II. la conservación del número bariónico III. la conservación del número leptónico. <div style="border: 1px solid red; padding: 5px; margin: 10px 0;"> <p>Se propone la reacción:</p> $n \rightarrow \pi^- + e^+ + \bar{\nu}_e$ </div> <p>¿Que leyes de conservación son violadas por la reacción propuesta?</p> <ul style="list-style-type: none"> A. Solo I y II B. Solo I y III C. Solo II y III D. I, II y III
--	---

Relying on the English version of the mark scheme

Relying on the Spanish version of the mark scheme is, overall, not problematic, however, the mark scheme of item Figure 9 was judged as *somewhat problematic* to *problematic to a large extent*. It is not possible based on the data provided by the survey and a close examination of the item to understand what was judged as particularly problematic for this item.

Figure 9. English version of the mark scheme of item 11biii of Paper 3, Physics SL.

11.	b	iii	<p>image formed by 10 cm lens is greater than 10 cm/further to the right of the first lens ✓</p> <p>so second lens must also move to the right OR lens separation increases ✓</p>	<p>Award [1 max] for bald "separation increases".</p>	2
-----	---	-----	--	--	---

Discussion

The findings from the expert and qualitative review of items in Phase 2 of this research are again very positive for the current translation model adopted by the IB, as the majority of items were judged to be highly comparable between the French and Spanish target versions and the English source version for the 2019 Physics SL, Chemistry HL and Biology SL examinations. Some inconsistencies appear in specific items but these inconsistencies, overall, tend to be minor and, as per much of the above descriptive discussion, not systematic with respect to the magnitude or direction of the DIF identified in Phase 1. Specifically, the Chemistry HL papers had many more items categorised as medium to large DIF but the translated versions of these items were, based on the judgement of the expert reviewers, more comparable to the English version on the variables targeted by our survey. Thus, with regard to

the first research question of this phase, generally there does not appear to be a systematic relationship between the differential difficulty of these items across the languages and the linguistic and translation differences between the source and translated items as judged by expert reviewers. Nonetheless, this relationship has only been qualitatively evaluated in this phase of the research and will be more robustly investigated in Phase 3 through the inclusion of the survey variables in the explanatory model.

Of the expert review criteria that did show some deviation between the source and target versions of the items, *matches and patterns* (matpat) and *accuracy of wording* (word) showed the most consistent and largest degree of deviation. Therefore, the performance of these two criteria in the explanatory modelling of Phase 3 is of particular interest. However, even for these criteria, the deviations remain small and unlikely to explain on their own the differences in difficulty of the items across the language versions. This is consistent with previous research examining translation impact, which has shown that when high quality translation is adopted, such as in PISA, language effects are only observed in some items and are often erratic (e.g., Huang et al., 2016; El Masri & Andrich, 2020).

Regarding research question two for this phase, given the high levels of inter-rater agreement for the majority of the criteria across the three subjects and two target languages, it appears 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 results regarding the reliability of the survey provide confidence for the use of these variables in the Phase 3 modelling. Nonetheless, some criteria showed consistently lower reliability than the typical 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 reliability of their judgement by, for example, ensuring their wording is clear to reviewers (e.g., the clarity of 'clause') and that judges have a standardised understanding of their meaning.

Moving forward into Phase 3 of the research not only provides an opportunity to more systematically evaluate the relationship between these expert review variables and the differential difficulty of items across language versions, it also introduces the review of items using NLP technology. Various NLP variables, which address a wide range of features and layers of text, have been shown to relate to textual complexity and many of these features would be very difficult for even expert human reviewers to discern. Thus, the NLP analysis of the items may reveal further linguistic differences between their language versions than this expert review, which may in turn help explain their differential demand. These NLP features are elaborated in the next sections.

Phase 3: Building a model to explain DIF across languages

In this phase, we combined the findings from the first two phases along with further linguistic analysis of the items to build an explanatory model of the DIF observed in Phase 1. This modelling was performed using machine learning techniques that are adept at identifying the most important predictor features for an outcome in ways that overcome the limitations of conventional statistical approaches. Phase 3 addressed the following research questions:

- Do linguistic and translation differences between source and target language versions of questions explain differences in their difficulty across the languages?
- What specific linguistic and other features are most associated with differences in difficulty across the source and target language versions of questions?

Carrying on from Phase 2, this model was built for the Physics SL, Chemistry HL and Biology SL subjects only. Two models were developed: one that includes the expert review variables from Phase 2 and so only applied to the subset of items from that phase, and, a second model that only includes the NLP-based variables, which was applied to all items with DIF estimates from Phase 1 for these three subjects. Moreover, as machine learning techniques are most effective with large amounts of data, we included the 2018 DP examination items for these three subjects to ensure more robust cross-validation of the findings. The NLP and machine learning approaches used in this phase of the research are further elaborated in the next sections.

Method

Quantifying differences in text complexity using NLP

To quantify the differences in text complexity across the different language versions of the items, we used an open-source, multilingual framework for analysing text complexity known as *ReaderBench*. This is a text processing framework for automatically assessing the complexity of text using analysis techniques from NLP (Dascalu et al., 2017a). The *ReaderBench* textual complexity analysis software framework has been developed on the basis of theoretical frameworks of linguistic complexity across languages and has been validated in multiple experiments (Gutu et al., 2016; Gutu-Robu et al., 2018; Dascalu et al., 2018).

How ReaderBench analyses text

ReaderBench makes use of an array of linguistic resources and text processing applications in order to process multiple textual analyses and provide a wide array of indices (over 300) for understanding textual complexity in multiple languages (Dascalu et al., 2017a). The analysis framework for *ReaderBench* is available in multiple languages including English, French and Spanish. The linguistic resources used to process the text comprise lexical ontologies, semantic models, corpora, and

lexicons. These resources are essentially compositions and/or large collections of real-usage written text, dictionaries, word lists and language processing models that inform the computational analysis of linguistic information. Linguistic resources such as these can be language-specific and/or multilingual (or a combination of both). In other words, some resources may be a single monolingual collection of texts from a specific language, while others may be a bilingual word list that provides information on word meanings in each language as well as on the connection between the words and their meanings across multiple languages.

In order to analyse the textual complexity of any text, the framework performs an initial pre-processing step using Standard Core NLP. In this step NLP software is applied to the input text to perform processes such as tokenising (break text into linguistic units), tagging (identifying and labelling specific features of text) and dependency parsing (identifying structural relationships within sentences) (Gutu et al., 2016; Gutu-Robu et al., 2018). Once these pre-processes are completed semantic models (such as Word2vec and Latent Semantic Analysis) and ontologies (such as WordNet and WOLF) are applied in order to determine similarities between units of texts (Dascalu et al., 2017b). These models and ontologies use large-scale linguistic information specific to each language from various linguistic corpora to ensure that the processes are computed reliably for each language (Dascalu et al., 2017a). Corpus selection is an important consideration for analyses such as the linguistic features of the text being analysed (in this case the text from questions in the DP Science examinations) is compared to linguistic features from real-life usage (in this case the language-specific corpora). This means that aspects such as the context and language variety of the source text might impact on the findings from the analysis.

The corpora used by *ReaderBench* for the languages being investigated in this study were considered to be sufficiently broad across dialects to be appropriate for application in this context. The English corpus is the TASA (Touchstone Applied Science Associates, Inc.) corpus, which consists of 37,651 different documents covering a broad variety of different topics (such as Literature, Arts, Science, Economics and Social Studies). It is a very widely used corpus across Europe and the United States. The French corpus is Le Monde, which is a newspaper corpus of over 500,000 articles covering a broad range of topics. The Spanish corpus derived from the project for the study of the literate norm of the main cities of Ibero-America and the Iberian Peninsula (PILEI project) and has broad coverage of varieties of Spanish from both Europe and Latin America.

One of the foremost benefits of using dedicated language corpora such as these is that it is possible to describe textual complexity in a language-specific way, thereby enabling comparisons of textual features across three language versions of the same text. In other words, the software uses linguistic features from real-life usage to analyse the text under investigation. This means that the picture of linguistic complexity presented from the analysis considers the relative demands of each language through real-usage data, enabling a more reliable picture of textual complexity to be used for the comparison of different language versions of the same text taking into account each language's intrinsic complexity.

What ReaderBench does

ReaderBench automatically classifies input text according to a variety of metrics separated into five categories; 1) surface, 2) syntax and morphology, 3) word, 4) cohesion, and 5) discourse (Dascalu, et al., 2018):

- The surface analysis considers the basic features of the text that include indices such as: word length, sentence length, paragraph length, and commas per sentence. These indices provide information predominantly related to the length of linguistic features in the text. Longer units of meaning may present a higher cognitive load to readers, potentially impacting the complexity of the text at a surface level. An additional surface-level indicator computed by *ReaderBench* is entropy measures for characters and words. One of the foremost areas which may present challenge for readers in any language is new or unfamiliar linguistic information in the text. This information could be in the form of words, letters, sentences or even punctuation. The more expected or predictable a sentence is for a reader, the easier that sentence is to understand. Entropy values are a valuable indicator of the predictability of linguistic units (e.g., characters, words, or strings of words) and therefore provide insight into the complexity associated with linguistic units at a surface level.
- The syntactic analysis looks at aspects of the text related to the structure of the text. Syntax refers to the linguistic rules and principles that affect how words are structured within a sentence or phrase to create meaning. Analysis at a syntactic level, therefore, looks at various aspects within sentence structures that may make them more complex for a reader. In order to do this, the software identifies and tags (labels) various parts of speech (i.e., nouns, prepositions, verbs) within the text. By tagging parts of speech, it is possible to obtain information regarding the level of complexity presented by the specific words within each sentence. The syntactic analysis presents information on all content words both in terms of average numbers of words as well as the numbers of unique words in each form. Unique words refer to words that are less frequently seen in real-language use and may therefore present increased complexity to a reader. The syntactic

analysis also focuses on specific parts of speech that are known to indicate more elaborate and complex text structure such as prepositions, adjectives, adverbs, and pronouns. Additionally, the software ‘sections’ the text according to syntactic levels known as a parse tree. By parsing sentences into syntactic levels, it is possible to identify aspects of the text structure which might present a higher level of complexity to readers. A single sentence with a large number of dependent phrases is likely to present a higher degree of complexity to a reader – sentences that are identified as having deeper and more complex dependency trees therefore give an indication of the syntactic complexity in the text.

- The word¹⁰ analysis considers features of the text that affect complexity at an individual word level. These indices include aspects related to the word *form* such as the number of syllables per word, as well as aspects that relate to the word *meaning* such as the specificity of words in the text. By considering both word form and word meaning as indicators of complexity, the word-level analysis does not only consider the words at a structural level, but also at a semantic level. The indices provided regarding word form relate to aspects within words that may make them more complex to a reader, these include the number of syllables per word as well as the number and length of affixes (prefixes and suffixes) in the words being analysed. Longer words in terms of syllable count, as well as words with longer affixes are considered more complex to decode and therefore add complexity to text at a word level. Indices related to word meaning relate to the number and density of meanings that can be associated with a particular word. These are reflected in characteristics such as word specificity, which gives an indication of the how rare a word might be. High frequency words (i.e., words that are commonly used in a language) present less challenge for readers and can decrease the complexity of a text, whereas encountering even one rare word in a sentence can impact the extent to which the entire sentence is understood by a reader (Graesser et al., 2011). Other indices of word complexity at a meaning level include aspects related to word polysemy and hypernym tree depth (semantic depth of words). These relate to the number of potential meanings a single word or word form may have, and therefore impact the relative potential that readers may have of misinterpreting the meaning of a sentence or text based on the word/s used.
- The final two categories – cohesion and discourse – describe aspects of the text such as cohesion between paragraphs, overall document flow, distribution of voices through the document, and structural connectives between paragraphs. These aspects relate to longer input text such as essays, articles, or short stories and as the items being analysed in this study are shorter (most items have only one or two sentences), only the first three categories were applicable to the

¹⁰ All word indices consider the lemmas of content words (nouns, verbs, adjectives, and adverbs)

analysis of textual complexity in this study. The items are therefore analysed in terms of surface complexity, syntactic complexity, and word complexity.

Steps conducted for the ReaderBench analysis

The analysis of the textual complexity of the items was carried out in the following steps:

1. All text that was included in items and/or relevant to answering an item was extracted from the Biology SL, Chemistry HL and Physics SL papers 1, 2 and 3 for each language (for 2019). Text was classified according to whether it formed part of the question for each item or whether it was included multiple choice answer options, equations/formula, text in tables, text in graphics, or text in the answer box for each item. Text that was not relevant to items (such as descriptions of options, mark allocation, cover page information) was not extracted.
2. Item text for items in each paper for all three languages was then collated into a new document to be checked for errors/inconsistencies in extractions. In this step the researcher confirmed that item codes were consistent across all three languages, that the text classified as 'item text' was the same across language versions for each item (i.e., no other text was included as item text), and that no item-text was missing in any language version. During this step, the item-text was converted to plain text and any extra line breaks were removed to ensure consistency when processing the text.
3. Item-text was then processed individually for each item in each language to obtain textual complexity indices using 'Demo client for *ReaderBench* 's Textual Complexity service' (English, French and Spanish). See *ReaderBench* Tutorial in Appendix 9 for instructions of how to use this service.
4. The raw output from the *ReaderBench* service was then transferred to an excel spreadsheet to be converted from a comma-separated string into columns for each index.
5. These steps were then repeated for the 2018 versions of the Biology SL, Chemistry HL and Physics SL papers 1, 2 and 3 for each language.

Features selected from the ReaderBench analysis

As described above, *ReaderBench* provides over 300 different indices and so a smaller subset of feature metrics was selected that we believed would be pertinent to explaining any differences in difficulty across language versions of Science DP examination items, as well as indices that could be meaningfully interpreted to inform real-life translation processes. The selected features are summarised in Tables 21 and 22 below.

Table 21.

Index name and description of the selected features from the *ReaderBench* analysis.

Index name	Description
AvgSentNoUnqWd	Average number of unique content words per sentence
AvgSentWdEntropy	Average word entropy per sentence in the document
ChNgramEntropy_2	Average entropy using distributions of ngrams by character
AvgSentUnqPOSMMain_adj	Average number of unique adjectives per sentence
AvgSentUnqPOSMMain_noun	Average number of unique nouns per sentence
AvgSentUnqPOSMMain_verb	Average number of unique verbs per sentence
AvgSentUnqPOSMMain_adv	Average number of unique adverbs per sentence
AvgSentUnqPOSMMain_pron	Average number of unique pronouns per sentence
AvgSentNoWd	Average number of words per sentence
AvgSentNoPunct	Average number of punctuation marks per sentence
AvgSentPOSMMain_adj	Average number of adjectives per sentence
AvgSentPOSMMain_pron	Average number of pronouns per sentence
AvgSentPOSMMain_adv	Average number of adverbs per sentence
AvgSentParseTreeDpth	Average parsing tree depth per sentence
AvgSentDep_det	Average parsing tree depth per sentence at specified level
AvgSentDep_nmod	Average parsing tree depth per sentence at specified level
AvgSentDep_case	Average parsing tree depth per sentence at specified level
AvgSentDep_amod	Average parsing tree depth per sentence at specified level
AvgSentDep_obj	Average parsing tree depth per sentence at specified level
AvgSentDep_aux	Average parsing tree depth per sentence at specified level
AvgSentDep_compound	Average parsing tree depth per sentence at specified level
AvgSentDep_nsubj	Average parsing tree depth per sentence at specified level
AvgSentDep_advcl	Average parsing tree depth per sentence at specified level
AvgSentDep_advmod	Average parsing tree depth per sentence at specified level
AvgSentDep_mark	Average parsing tree depth per sentence at specified level
AvgSentDep_acl	Average parsing tree depth per sentence at specified level
AvgSentDep_dep	Average parsing tree depth per sentence at specified level
AvgSentDep_ccomp	Average parsing tree depth per sentence at specified level
AvgSentDep_cop	Average parsing tree depth per sentence at specified level
AvgWordWdLen	Average word length (characters)
AvgWordWdDiffLemma	Average distance between lemma and word stems
AvgWordMaxDepthHypernymTree	Max word depth in hypernym tree
AvgWordAvgDepthHypernymTree	Average word depth in hypernym tree
AvgWordPathsHypernymTree	Number of paths to hypernym tree root
AvgWordWdPolysemy	Average word polysemy count (only content words)

Table 22.

Index name of the selected features and their relationship with text complexity.

Index name	Relationship with complexity
AvgSentNoUnqWd	More unique words per sentence gives an indication of the complexity of comprehending the sentence for readers
AvgSentWdEntropy	Word entropy gives an indication of the predictability of words. When words are less predictable complexity increases.
ChNgramEntropy_2	Ngram entropy gives an indication of the predictability of the relationship between items (in this case characters) in sequence. When sequences of items are less predictable, complexity increases.
AvgSentUnqPOSMMain_adj	Higher occurrence of unique adjectives reflects increased sentence complexity
AvgSentUnqPOSMMain_noun	Higher occurrence of unique nouns reflects increased sentence complexity
AvgSentUnqPOSMMain_verb	Higher occurrence of unique verbs reflects increased sentence complexity
AvgSentUnqPOSMMain_adv	Higher occurrence of unique adverbs reflects increased sentence complexity
AvgSentUnqPOSMMain_pron	Higher occurrence of unique pronouns reflects increased sentence complexity
AvgSentNoWd	More words per sentence could indicate higher complexity
AvgSentNoPunct	Higher number of punctuation marks can reflect increased complexity
AvgSentPOSMMain_adj	Higher occurrence of adjectives can reflect increased sentence complexity
AvgSentPOSMMain_pron	Higher occurrence of pronouns can reflect increased sentence complexity
AvgSentPOSMMain_adv	Higher occurrence of adverbs can reflect increased sentence complexity
AvgSentParseTreeDpth	The depth of a parse tree can reflect increased syntactic complexity and therefore make text more complex to comprehend
All <i>AvgSentDep</i> Indices	Increased depth of a parse tree can reflect increased syntactic complexity and therefore make text more complex to comprehend
AvgWordWdLen	Longer words can indicate more complex text to decode and understand
AvgWordWdDiffLemma	These distances give an indication of the complexity of word forms in the text
AvgWordMaxDepthHypernymTree	The distances in hypernym trees provide a representation of the number of meanings associated with the content words, therefore providing an indication of semantic complexity in the text
AvgWordAvgDepthHypernymTree	
AvgWordPathsHypernymTree	
AvgWordWdPolysemy	The average number of possible meanings for a word can increase the complexity of understanding when reading

Modelling DIF using a machine learning approach

The DIF estimates from Phase 1 were modelled in terms of the expert review variables from Phase 2 and NLP-based variables from the current phase, as well as several other relevant variables using a machine learning approach, which similar to Phase 1, involved the fitting of multiple models to evaluate the best fitting model. This approach extends the work of El Masri et al. (2017) by using a modelling approach that does not solely rely on stepwise regression, as this method is known to have a number of biases including being less effective when there is a large number of potential explanatory variables (Smith, 2018). Given the multitude of variables being included in the explanatory model of DIF across the different language versions of the items in this phase of research, this is a significant shortfall. Consequently, we also applied regression approaches that have emerged from machine learning, including Elastic Net regression (Zou & Hastie, 2005) and Random Forest regression (Grömping, 2009), and all three approaches were implemented using the R package 'caret' (Kuhn, 2008).

These two approaches were selected because they overcome a number of limitations of standard regression, including tolerance of highly correlated explanatory variables, which is the case across many of the variables included in our modelling, and because they provide unbiased estimates for models with a very large number of explanatory variables. This multi-pronged approach to modelling should ultimately lead to the most optimal identification of the set of features to explain the DIF across the three language versions.

Elastic Net regression is the more similar to stepwise regression in that it models the outcome variable in terms of linear relationships with the predictor variables and therefore provides regression coefficients that can be interpreted in the same manner as standard regression. However, unlike stepwise regression, which removes predictors based on their lack of statistical significance, Elastic Net regression includes additional parameters¹¹ that are used to 'penalise' predictor variables with the smallest coefficient estimates, shrinking and/or setting them to zero and thereby optimising the model for the most predictive variables.

Random Forest regression is the more distinct approach as it is able to model linear and non-linear relationships between the predictor variables and the outcome variable, as well as complex interactions between the predictor variables, and thus it is a much more flexible modelling approach. For this reason, it is increasingly used for purposes similar to the current phase of research where researchers want to use features of items and other variables to predict aspects of item responses (e.g., Han, He, & von Davier, 2019). This approach creates a number of decision trees and randomly

¹¹ In machine learning parlance, these are referred to as 'hyperparameters'.

assigns and evaluates several predictor variables at the various nodes of each decision tree. After optimising each of the individual tree, the predictions are averaged across each of the trees to attain a single model prediction for each case. Similar to the penalty parameters in Elastic Net Regression, the additional parameters that need to be 'tuned' for Random Forest regression include the number of decision trees and the number of predictor variables that are randomly trialled for each node of the trees. Despite being the more flexible approach, a major downside of Random Forest regression is that it is more of a 'black-box' in terms of interpreting the relationships between the individual predictor variables and the outcome variable. Nonetheless, it provides a ranking of importance of the predictor variables in terms of their contribution to the predictive success of the model.

Running Machine Learning models

As alluded to in the above discussion, machine learning models typically include parameters that affect the overall success of the model and that need to be 'tuned' from the data. Moreover, a major concern in a machine learning approach is to obtain generalisable model estimates and to avoid 'overfitting' your model to your specific dataset. A common approach to achieve the best tuning for these parameters and to avoid the issue of overfitting is to employ k-fold cross validation methods.

K-fold cross validation involves randomly splitting your dataset into k folds where you use $k - 1$ folds as the training data for the model and then test the model on the fold that is left out of the training, and this process is then repeated k times. In this way, you are only ever testing the model on data that was not used to train the model. The number of folds to use is dependent on your sample size, but a minimum of five folds is argued to give the best results. Moreover, this process can be extended so that you repeat the k-fold validation process several times to then average the model fit statistics across all the training and testing repetitions to get the optimal tuning of the model parameters and unbiased estimates of the model performance. However, as these approaches involve splitting your data into training and testing sets, the more data you have at your disposal, the less biased and more generalisable the model estimates will be.

Evaluating model performance in the machine learning context involves fit statistics that are also common in standard regression. These statistics include the Root Mean Square Error (RMSE) statistic, which quantifies the amount of model prediction error and thus smaller values indicate better model performance, as well as the R^2 statistic, which represents the percentage of variance in the outcome variable that is explained by the model and thus higher values indicate better model performance. Both statistics were used to compare relative performance across the three modelling approaches employed in this phase of the research.

Steps in building the explanatory model

Building and evaluating the explanatory model for DIF across the language versions of the items using these machine learning approaches was a multi-step process:

1. The datasets were collated whereby the outcome variable was the DIF estimate for each of the item-target language combinations (i.e., some items had both a French and a Spanish DIF estimate) evaluated in Phase 2 of the research for the Physics SL, Chemistry HL and Biology SL subjects. As the DIF estimate is a standardized statistic and therefore generally interpretable, the explanatory modelling was simultaneously run across all three subjects. The predictor variables include the subject area and paper the item was from, as well as the target language the DIF estimate was for (i.e., French or Spanish). All the expert review variables from Phase 2 were also included as predictors in the model. Finally, the difference scores for the selected NLP feature metrics between the source language and the relevant target language were included as predictors, i.e., the DIF is being predicted in terms of the differences in the complexity metrics across the source and target languages.
2. Some data pre-processing was required, including dropping predictor variables that had no or very close to zero variance, dropping variables that had more than 50% missing data, median imputing missing values for the other predictors with missing values, scaling and centering the continuous variables, and creating dummy variables for the categorical variables.
3. The stepwise regression, Elastic Net regression and Random Forest regression models were then applied to the data using 5-fold cross validation that was repeated ten times and the best model was selected in terms of each model's average RMSE and R^2 values across the cross-validation repetitions. Moreover, each model was implemented in a way that variables that added no predictive utility to the model were omitted from the final model.
4. The best fitting model of the three was then evaluated in terms of its overall performance (how much variance it explained in the DIF outcome variable) and the most important predictor variables, i.e., those that help explain the most variance in the DIF estimates, as represented by higher %IncMSE¹² values for the predictor.
5. These steps were then repeated for all items with DIF estimates from the Physics SL, Chemistry HL and Biology SL examinations from both 2019 and 2018¹³. The expert review variables were not included as they were only available for a subset of the 2019 items. This additional modelling was conducted because the much larger dataset enables a more robust evaluation of how well

¹² The increase in mean square error of predictions as a result of the variable being permuted (values randomly shuffled).

¹³ The DIF estimates for the 2018 items were calculated in an identical manner to the process described in Phase 1 of the research.

differences in these NLP based complexity features explain DIF across the source and target language versions of items. Because the calculation of these NLP metrics is semi-automated, unlike the expert review, it was practical to obtain these measures for all 2018 and 2019 items for the three subjects. Due to the use of both 2018 and 2019 items, calendar year was added as a predictor variable to the models. Moreover, given the substantially larger dataset, 10-fold cross validation with 10 repeats was conducted in fitting these models.

Results

The modelling results for the subset of items from Physics SL, Chemistry HL and Biology SL that were expert reviewed in Phase 2 are presented first, followed by the findings for the combined 2018 and 2019 data for these three subjects where the expert review variables are omitted as predictor variables.

Explanatory models including expert review variables

The performance of each of the three explanatory models is presented in Table 23 below. As can be seen, the Random Forest regression model showed better fit to the data in terms of a lower average RMSE and a higher average R^2 across the cross-validation samples. The Random Forest model explains 11% of the variance in the DIF estimates for the subset of items from Phase 2 of the research.

Table 23.

Average model fit statistics for the three explanatory models.

Model	Mean RMSE	Mean R^2
Stepwise	16.21	0.04
Elastic Net	14.28	0.07
Random Forest	13.98	0.11

The final set of variables included in the Random Forest model and their relative importance for the overall model performance are presented in Table 24. The table contains variables from the expert review, the NLP complexity indices from the textual analysis as well as categorical variables referring to the paper type, subject, language and type of response.

Table 24.

Final set of predictors in the Random Forest model in order of their relative importance.

Predictor	%IncMSE	Type
msres.1	100	Category
Subject.BIOLOGY	98.8311	Category
paper.3	92.1341	Category
Subject.CHEMISTRY	87.5666	Category
AvgSentDep_nmod	70.6289	NLP
AvgSentDep_aux	68.7264	NLP
AvgSentDep_amod	68.2468	NLP
paper.1	67.027	Category
msres.2	62.9543	Category
AvgSentUnqPOSMain_adv	62.5955	NLP
paper.2	60.7759	Category
AvgSentDep_obj	57.8925	NLP
AvgSentDep_det	57.4497	NLP
AvgSentDep_obl	49.6793	NLP
msres.3	49.0757	Category
msres.8	43.6646	Category
AvgSentWdEntropy	42.6421	NLP
AvgSentPOSMain_adv	42.3447	NLP
Subject.PHYSICS	40.6069	Category
AvgSentDep_advmod	37.7631	NLP
AvgSentUnqPOSMain_verb	36.5267	NLP
msres.7	35.6784	Category
word	34.9645	Expert review
AvgSentDep_mark	31.278	NLP
AvgSentNoWd	30.9989	NLP
AvgSentUnqPOSMain_adj	29.3965	NLP
Language.FRENCH	29.3638	Category
info	28.7371	Expert review
AvgSentUnqPOSMain_noun	26.7859	NLP
gram	26.2506	Expert review
AvgSentDep_case	25.9988	NLP
AvgSentDep_compound	22.8662	NLP

Predictor	%IncMSE	Type
msres.6	22.8553	Category
AvgSentNoPunct	22.7424	NLP
lclaus	20.1404	Expert review
house	19.4081	Expert review
Language.SPANISH	19.2351	Category
AvgSentPOSMain_adj	19.0348	NLP
AvgSentDep_advcl	17.7665	NLP
AvgSentDep_dep	17.6057	NLP
AvgSentUnqPOSMain_pron	17.3691	NLP
AvgSentNoUnqWd	17.0712	NLP
AvgSentDep_ccomp	16.8267	NLP
dok	16.2238	Expert review
freg	15.5085	Expert review
AvgSentDep_acl	14.9848	NLP
layfor	14.1284	Expert review
matpat	13.7277	Expert review
AvgSentPOSMain_pron	13.5389	NLP
AvgSentParseTreeDpth	12.629	NLP
AvgSentDep_appos	10.0148	NLP
AvgSentDep_cop	9.8841	NLP
ChNgramEntropy_2	8.90994	NLP
AvgSentDep_nsubj	8.15007	NLP
msres.4	2.49503	Category
msres.5	0	Category

Interestingly, many of the expert review variables are not among the top important predictors from the final model. Moreover, aside from the categorical variables (subject, response type, etc.) the top 10 most important predictor variables are differences in NLP complexity metrics across the source and target language versions of the items. These findings highlight the importance of running the next set of models with the larger dataset from both 2018 and 2019 to more robustly evaluate how well differences in these NLP variables explain DIF, as well as the relative importance of the different NLP variables.

Explanatory models including both 2018 and 2019 data

As a reminder, this set of models does not include the expert review variables, as they were only available for the subset of 2019 items evaluated in Phase 2, and calendar year has been added as a predictor variable. The performance of each of the three explanatory models is presented in Table 25 below. In this set of models, the Random Forest regression model showed slightly better fit to the data in terms of a lower average RMSE and a higher average R^2 across the cross-validation samples. In this case, the Random Forest model explains, on average, 4% of the variance in the DIF estimates for the full set of items from 2018 and 2019 across the three subjects.

Table 25.
Average model fit statistics for the three explanatory models.

Model	Mean RMSE	Mean R^2
Stepwise	9.27	0.01
Elastic Net	9.26	0.02
Random Forest	9.14	0.04

The final set of variables included in the Random Forest model and their relative importance for the overall model performance are presented in Table 26 below. These variables and their implications for translation will now be discussed.

Table 26.
Final set of predictors in the Random Forest model in order of their relative importance.

Predictor	%IncMSE
AvgSentWdEntropy	100
AvgSentNoPunct	75.4398
paper.2	74.641
AvgSentDep_det	67.9017
AvgSentNoUnqWd	63.5303
paper.1	61.3895
paper.3	60.2188
AvgSentUnqPOSMain	59.9305
AvgSentDep_amod	58.6371
AvgSentNoWd	58.489
Subject.CHEMISTRY	57.4893

Predictor	%IncMSE
AvgSentUnqPOSMMain_adv	56.2341
AvgSentDep_case	55.7135
AvgSentPOSMMain_adv	45.6774
AvgSentDep_aux	45.2839
AvgSentDep_advmod	45.2572
AvgSentDep_nmod	43.6823
AvgSentParseTreeDpth	38.6841
AvgSentPOSMMain_adj	38.4879
AvgSentDep_compound	37.6742
AvgSentDep_ccomp	37.2257
AvgSentUnqPOSMMain_adj	36.9706
AvgSentDep_advcl	35.2148
AvgSentDep_nsubj	35.1915
AvgSentUnqPOSMMain_verb	34.1463
AvgSentDep_obj	29.3462
AvgSentUnqPOSMMain_pron	28.2804
AvgSentDep_appos	27.2369
Subject.BIOLOGY	27.088
Subject.PHYSICS	25.302
Year.2019	22.8649
AvgSentDep_cop	22.7911
AvgSentPOSMMain_pron	20.8503
Year.2018	20.2032
ChNgramEntropy_2	20.0998
AvgSentDep_mark	15.3187
AvgSentDep_nummod	12.4032
Language.SPANISH	7.33575
AvgSentDep_acl	6.93739
AvgSentDep_obl	4.92265
AvgSentDep_dep	1.45204
Language.FRENCH	0

Discussion

Overall, with respect to the first research question for this phase, there were mixed findings regarding how the linguistic and translation differences between source and target language versions of questions explain differences in their difficulty across the language versions. Firstly, most of the variables from Phase 2's expert review were not found to be substantial predictors of the language DIF, but this is consistent with the above discussed descriptive findings for these variables from Phase 2. This is likely attributable to the lack of variation in these expert review variables, i.e., most of the survey item responses that involved a comparison of aspects of language and translation across the versions of the items indicated they were the same or very similar despite the items having differing levels of language DIF.

In both cases, the top predictors included NLP text complexity metrics as well as variables that described categorical features of the assessments such as subject, response type and paper version were among the top predictors. Table 27 below provides a description of the categorical variables and the feature to which they refer.

Table 27.
Description of variables indicating categorical features.

Variable name	Description
msres.1	Response type: multiple choice
Subject.BIOLOGY	Subject: Biology SL
paper.3	Paper: 3
Subject.CHEMISTRY	Subject: Chemistry HL
paper.1	Paper: 1
msres.2	Response type: numerical solution without working
paper.2	Paper: 2
msres.3	Response type: numerical solution with working
msres.8	Response type: paragraph or more
Subject.PHYSICS	Subject: Physics SL
msres.7	Response type: couple of sentences
Language.FRENCH	Language: French
msres.6	Response type: phrase or sentence
Language.SPANISH	Language: Spanish
msres.4	Response type: visual representation only
msres.5	Response type: word

Although many of these categorical variables were found to be substantial predictors of language DIF, this is likely attributable to the relationship between the category and its relative DIF estimates. For

instance, the topmost important predictor, *msres.1*, refers to items with a multiple choice response type and Phase 1 of the analysis found lower instances of DIF in multiple choice items. The high predictive value of *msres.1* is therefore more likely to be due to the relationship between multiple choice items and DIF estimates than the linguistic features of multiple choice items. Similarly, the subject Biology had the lowest instance of moderate to high DIF observed across all items and it is one of the highest predictors of language DIF in the model. As the predictive properties of these variables is likely to be related to their relationship with DIF estimates in Phase 1, the remainder of this discussion section covers the predictor variables that relate more directly to linguistic features of the items from both the expert review and NLP text complexity metrics.

Differences in the NLP text complexity metrics as well as differences found during the expert review 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 extent. The performance of the Random Forest model, the best fitting model in both cases, was better for the smaller subset of items from Phase 2, accounting for 11% of the variance in the DIF outcome variable as opposed to only 4% for the larger dataset including all items and 2018 data. The better performance of the model in the former case, despite omitting all of the expert review variables from the final model, may be attributable to the model overfitting the much smaller dataset. However, the subset of items from Phase 2 only included items with statistically significant language DIF of differing magnitudes and so there was less of a restriction of range issue, i.e., most z-score DIF estimates for the full cohort of 2018 and 2019 items are small or close to zero. This may also explain the relatively poor performance of both linear models given the negative impact that restriction of range of the outcome variable has on such models.

Regarding the second research question for this phase of research, the rest of the discussion will elaborate the findings and implications regarding the NLP text complexity indices as well as the expert review variables that were found to be the most predictive of language DIF. In particular, we will elaborate in more detail how these indices are related to text complexity to facilitate understanding of how differences in these indices across the language versions of items may manifest in language DIF. These have been organised into three themes that describe the linguistic features related to text complexity: word choice, sentence length and structural complexity.

Word choice

One of the foremost areas that may present challenge for readers in any language is new or unfamiliar information in the text. The more expected or predictable a sentence is for a reader, the easier that sentence is to understand (Graesser et al., 2011). This information could be in the form of words, letters, sentences or even punctuation. At a word level, the more words in the text that are known

and familiar, the more likely it is that students will understand what they read. In order to understand the impact of knowledge of words on textual complexity, the frequency of words is considered to be a good indicator of whether students are likely to be familiar with the words they encounter in the text. If a word occurs frequently in a language generally, then the likelihood that students will know the word and its meaning increases (Chen, 2016). The relationship between words used in the text and the frequency of those words in the language on the whole is therefore a central consideration when assessing the complexity of a text (Crossley et al., 2008; Graesser et al., 2011).

Understanding word frequency usually involves using linguistic resources such as high frequency word lists from language corpora and comparing the words in the text being analysed with the words that appear as high frequency words in the word list (McCarthy & O’Keeffe, 2010). High frequency word lists are easily accessible in a number of languages and are widely used to understand the relative frequency of words in a text. *ReaderBench* applies a variety of linguistic resources to assess the relative frequencies of words in the input text and provides information on the number of ‘unique’ words in different categories. Unique words refer to words that are not considered high frequency words and are therefore more likely to be unfamiliar to readers. Table 28 provides a list and description of the textual complexity indices related to word frequency that differences across language versions of the items were found to be important predictors of the magnitude of DIF in the items across the languages.

Table 28.
Description of variables related to word choice.

Index name	Description	Relationship with text complexity
AvgSentNoUnqWd	Average number of unique content words per sentence	More unique words per sentence gives an indication of the complexity of comprehending the sentence for readers
AvgSentUnqPOSMain_adj	Average number of unique adjectives per sentence	Higher occurrence of unique words reflects increased sentence complexity
AvgSentUnqPOSMain_noun	Average number of unique nouns per sentence	Higher occurrence of unique words reflects increased sentence complexity
AvgSentUnqPOSMain_verb	Average number of unique verbs per sentence	Higher occurrence of unique words reflects increased sentence complexity
AvgSentUnqPOSMain_adv	Average number of unique adverbs per sentence	Higher occurrence of unique words reflects increased sentence complexity
AvgSentWdEntropy	Average word entropy per sentence in the document	Word entropy gives an indication of the predictability of words. When words are less predictable complexity increases.
word	Accuracy of wording	Word choice may introduce undue complexity to sentences.

As is evident in Tables 26 and 28, the number of unique words in specific parts of speech (nouns, verbs, adverbs, and adjectives), as well as the entropy values of words added to the relative complexity of

the different language versions of the items. Entropy values are a valuable indicator of predictability of linguistic information for readers (Clark, 2013). An entropy approach involves quantifying a specific aspect of text (i.e., a word) to describe it not just in terms of whether it is used ‘consistently’ or ‘inconsistently’, but also accounting for the probability that the linguistic unit might occur in a consistent or inconsistent way in real-life usage (Borgwaldt, Hellwig, & De Groot, 2005). In other words, entropy provides a metric for understanding the degree to which a specific word in a text might be considered familiar to readers.

Another variable related to word choice that was among the top important predictors from the expert review phase was *word* which refers to the accuracy of wording in the target versions. For this variable, experts were asked to evaluate the extent to which the word choice in the target version corresponded accurately with meaning conveyed by the words used in the English version. If word choice affected or construed the meaning conveyed in the target versions, this is likely to have added to the linguistic complexity of that item for students taking the examination.

It is important to note that it is not necessary to avoid unfamiliar words entirely when designing items, particularly in subjects such as science where subject-relevant terminology is required. However, when evaluating the performance of the same item across different languages, the relative frequencies of words in the item for each language should be considered as well as the impact that specific word choice may have in a subject-specific context such as science.

Sentence length

Reading and understanding text involves a variety of cognitive skills and strategies, and as such, the cognitive load associated with a particular item may present added challenge (Crossley et al., 2008). 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. In terms of textual features, the association between sentence length and complexity occurs at the surface level of text. The two indices for which differences across language versions were found to be important predictors of the magnitude of DIF related to sentence length are described in Table 29.

Table 29.
Description of textual complexity indices related to sentence length.

Index name	Description	Relationship with text complexity
AvgSentNoWd	Average number of words per sentence	More words per sentence could indicate higher complexity
Iclaus	Length of clauses	More words per clause could indicate higher complexity

As is evident from the table, two features regarding sentence length can be associated with textual complexity: the average number of words per sentence from the NLP variables and the length of clauses from the expert review. In the case of the average number of words per sentence, it is important to consider not just the length of the sentences in items, but the relative length of sentences in the same items across language versions. In other words, longer sentences might not present additional challenge if they are consistent across language versions. The expert review variable *lclaus* referred specifically to the differences between the length of clauses in the target versions and the English version. Sentences or clauses of different length could present different levels of complexity to students taking the test in different languages. Additionally, although shorter sentences are often associated with the readability of a text, the number of words in a sentence should be considered carefully in association with other linguistic features of the text (Graesser et al., 2014). At times, increasing the length of a sentence might actually reduce the textual complexity, for instance if words are added to clarify a concept or to describe an unfamiliar term.

Structural complexity

The structure of text is governed by language principles related to the grammar of a language. Investigating different features of a text in terms of the grammatical and syntactical features can give an indication of the structural complexity of the text. Textual analysis software assists with understanding the structural complexity of a text by looking at the types of words used (parts of speech), as well as the different syntactic levels (parse tree). The indices for which differences between language versions were found to be important predictors of the magnitude of DIF in the items across languages related to structural complexity are described in Table 30.

Table 30.
Description of textual complexity indices related to structural complexity.

Index name	Description	Relationship with text complexity
AvgSentPOSMain_adj	Average number of adjectives per sentence	Higher occurrence of adjectives can reflect increased sentence complexity
AvgSentPOSMain_adv	Average number of adverbs per sentence	Higher occurrence of adverbs can reflect increased sentence complexity
AvgSentDep_	Average parsing tree depth per sentence at various levels	The depth of a parse tree can reflect increased syntactic complexity and therefore make text more complex to comprehend
AveSentNoPunc	Average number of punctuation marks per sentence	Higher number of punctuation marks can reflect increased complexity
gram	Grammatical errors	Grammatical errors may affect text comprehensibility

As is evident from the table, the indices relevant to structural complexity relate to parts of speech used in the text, and the depth of the parse tree. In terms of parts of speech, there are certain parts of speech that, when used in a text, reveal increased complexity (Dascalu et al., 2018). Parts of speech such as adverbs and adjectives also serve as indicators of complexity as they relate to more elaborate

sentence structures. Sentences with adverbs for example, provide additional detail and are more likely to contain additional clauses.

The depth of the parse tree refers to the syntactic complexity of a sentence. The different structural elements of a sentence can be represented using a network diagram known as a parse tree (or syntax tree). This is processed by NLP software by parsing (separating) and labelling units of text within a sentence according to their grammatical role. Deeper parse trees represent sentences with more complex grammatical structures. As can be expected, sentences with multiple additional clauses and more complex grammatical structures can present increased challenge to readers. As with sentence length, although increased parse tree depth may make sentences appear more complex at a surface level, it may also add clarity for readers. It is therefore essential to consider the items carefully and pay specific attention to differences in the relative parse tree depth across language versions.

The number of punctuation marks per sentence, which was the second most important variable in the final model, provides insight into the complexity of sentences at a syntactic level. More commas in a sentence, for example, can give an indication of the number of clauses in the sentence, which can be associated with increased complexity (Kuboň et al., 2006). However, as with the number of words in a sentence, increased punctuation can aid in providing clarity for readers. Complex grammatical structures associated with long sentences should therefore be considered carefully by experts to assess whether they increase or decrease the textual complexity of the sentence in question, and particularly across different language versions of a sentence.

Translated texts that rely on one language source (in this case English), can result in sentences that have complex, awkward or even inaccurate grammatical structure. The expert reviewers evaluated grammatical complexity associated with grammatical errors in the target versions. While this variable was not among the highest predictors in the model, it was the third most important predictor of the expert review variables and gives an indication of the linguistic complexity associated with complex grammatical structures in translated versions.

Conclusion

The findings of this phase of research provide evidence that linguistic differences in translated versions of items from the source version can explain the differential difficulty of items across languages to some extent. Moreover, this relationship was more apparent for differences in the computer assessed NLP indices rather than for the human rated variables. This may be attributable to the greater sensitivity of the NLP indices to subtle and varied linguistic features that go beyond what is possible for humans to judge, and thus these indices also provide greater variability when quantifying differences across language versions of items. Nonetheless, even the best performing model for the

two datasets only accounted for 11% of the variance in the language DIF outcome variable and so it is clear that other substantial factors are contributing to the observed differences in difficulty for some of the translated versions of the items across the three languages. In the next section, we will draw upon the findings from the three phases of research to make high level and practical recommendations for IB DP Science examination design and translation processes.

Recommendations

Overall, the findings of this research study have been very positive for the IB's current translation processes with DP Science examinations, but nonetheless, each phase of the research has raised findings that suggest recommendations for future improvements. Before overviewing these recommendations from each phase of the research, we will summarise the current translation processes at the IB to provide context.

Translation processes at IB

The IB adopts a clear process for the translation of its assessments. The process includes several stages including a translation stage and a revision stage before the assessment is ready for production and press. The translation stage is led by qualified professional translators commissioned by IB to translate its assessments from English to one of the many target languages. Translators are required to have the target language as a first language. In addition, they need to have a degree or proven professional expertise in the subject domain (e.g., Biology, Chemistry, History, Mathematics, etc.) and at least 4 to 5 years of translation experience ideally working with educational material and in international organizations. Following translation, revisers who are native speakers of the target language and subject experts, such as IB teachers and university professors, verify the quality of translation of the assessments and the adherence of the translated version to IB guidelines. Revisers are not required to have any translation qualification.

IB provides the translators and revisers with guidelines for its translation procedures. Translators are required to ensure that the translated versions of the assessments are as accurate and as faithful as possible to the English source version in terms of question difficulty, language used, and formatting and style. Translators are provided with additional documents such as subject guides, subject-specific glossaries and lists of command terms to ensure that the translations include terminology that is congruent with the terminology adopted in IB curricula and textbooks. Translators are reminded to avoid region-specific vocabulary given the globalised nature of the IB community. Translators need to follow the layout, formatting and style adopted in the English version of the assessment and hence need to ensure that the terms in italics, bold, upper case, etc. and house style are mirrored in the translated version except when such features violate conventions in the target language.

Revisors are provided with similar documents and guidelines for reviewing the quality of translation. Their scrutiny includes verifying the translated version for accuracy and completeness. Revisers need to ensure that the target version is comparable to the English version in terms of language, style and layout. The guidelines require that revisors compare the accuracy of translation at a sentence level at first, and that they assess the quality of the target version in terms of fluency and grammar.

IB has put in place a clear process for translating its assessment and their selection of translators and revisors is largely in line with guidelines adopted more broadly for translating and adapting assessments into different languages (International Test Commission, 2017). Translators and revisors are bilingual and have the target language as a first language, they have a strong background in the subject translated, and while translators have professional qualifications and experience in translation, revisors are typically well versed in IB assessments. This model of selection helps prevent the unintentional introduction of elements that make the test easier or more difficult in the adapted versions. Nevertheless, there is room for improving the process in line with the literature in this area. Below we list a number of recommendations for improving the process based on the findings of the project and the research literature more broadly.

Recommendations from Phase 1

While Phase 1 was primarily concerned with identifying items that show differential difficulty between the source English version and the French and Spanish target versions, and not with providing explanations for any differences, the findings do raise recommendations for future improvement of processes.

- **Review multiple-choice items that show differential rates of guessing across language versions**

Given the apparent relationship between guessing and DIF across the language versions, it is recommended that IB investigate the multiple-choice items with high infit values (>1.3) (see Appendix 4) to evaluate why these items might be showing elevated levels of guessing, including evaluating the functioning of the distractors and whether there may be greater ‘guessability’ in one language versus others to inform future multiple-choice item design and translation.

- **Review items that show medium and large language DIF for the other three subjects and other years**

While the DIF analyses revealed a relatively small percentage of substantial, i.e., medium and high, DIF across the subject-level combinations, it is recommended that IB evaluates all items that were found to have this degree of language DIF in the three subject-levels that were not included in the second and third phases of this research study in a similar manner to these phases, as well as DP Science examinations from other calendar years where item-level responses are available. This will help IB to accrue more evidence regarding how much this DIF is attributable to linguistic and translation effects across the language versions.

Recommendations from Phase 2

While the findings of the Expert Review of DIF items was overwhelmingly positive for the current IB translation processes, it still raised a number of recommendations for improving these processes, which included the following themes:

- **Back-translation at the revision stage of the assessment**

Based on IB documentation, revisors are required to scrutinise the quality of the translated version of the assessment by comparing the accuracy of the translation to the source version before carrying out an additional check on the translated version in isolation of the source version. While such a model has merit, it may lead to some errors in the translation being missed. Evidence suggests that reviewers are incapable of flagging all problems in questions (e.g., Graesser et al., 2006) and this may be exacerbated when revisors are having to judge two language versions side by side where their judgement of quality can be affected by confirmation bias. Therefore, it is recommended that back-translation is used as part of this revision process.

- **Decentring the assessment**

The choice of translation design when it comes to multilingual educational assessments is an important consideration. Translators can adopt forward translation, back-translation, translation plus review, decentred translation or a combination of different translation models. Decentred translation models are beneficial as they are less dependent on a single source language. Having a single source language can give too much importance to the linguistic, syntactic and stylistic conventions of that language and thereby impact the target versions (Grisay, 2003). In decentred models, two source versions of an assessment are created in different languages a target version (e.g., Spanish) is then created from two language versions (e.g., English and French) of the assessment that act as source versions. This is sometimes done at item design level, where items are designed in two source languages before the full assessment is designed. Consequently, by not relying exclusively on a single source language, assessments can be less culture- and dialect-based.

Dialects are forms of a language used within specific regions or social groups. Dialects within a language can become a threat to the validity of translated tests. British English and American English use different vocabulary at times, such as pavement and sidewalk, lift and elevator, trousers and pants or aubergine and eggplant. For instance, Schittekatte et al. (2003) and Tewes (2003) highlighted the challenges faced in adapting the WISC-III intelligence test into Dutch and German respectively. Regional differences in dialects in those two languages posed major problems when constructing an unbiased version of the verbal items in the Netherlands and Belgium, or Germany, Austria, and Switzerland. While dialects were a feature of the IB science assessments that could not be explored

empirically for the reasons outlined earlier in this document, IB could think of producing a version that would apply across dialects of a language.

An important consideration with any translation design approach is to ensure that the quality assurance or review process is able to happen bidirectionally. In other words, if an issue is identified in the target version it might not be exclusive to the target version only and reviewers should be able to return to the source version to resolve the issue wherever relevant. This can be done through multistage review procedures, reconciliation processes or extensive cross-checking.

- **House style – reviewing command terms**

The issue of house style reflected in the translation of command terms only emerged in Biology SL. It is recommended, however, to ensure that the lists of command terms are translated into the target languages without introducing awkwardness in the language or nuanced difference in their meanings across languages.

- **Translation of mark scheme**

While there was only little evidence from the expert review suggesting that the absence of a translated version of the mark scheme for DP Science items is problematic, expert reviewers expressed that significant issues had arisen with past examinations because of the absence of translated mark schemes. This is unsurprising, as mark schemes can often have issues even without the added complexity of multiple languages. For example, mark schemes can be misinterpreted by markers, especially novice ones, and as a result, marking can vary significantly in terms of leniency/severity. Issues with marking reliability may become even more significant if markers have to interpret answers written in a different language, and so it is recommended that further research is done on a wider range of subject to evaluate whether the lack of translation of the mark schemes is having an impact on the validity of the assessment across multiple languages.

Recommendations from Phase 3

As a consequence of applying NLP techniques to evaluate text complexity in the third phase of the research study, it was shown that there were subtle linguistic differences across translated versions of items in many cases and that these differences were associated with the language DIF to some extent. Consequently, there are several recommendations that emerge from this phase of research, particularly with respect to the NLP indices, the difference of which across the language versions were found to be the most important features for explaining the DIF. These recommendations centre around paying greater attention to linguistic features that are known to be associated with text complexity and ensuring that these are comparable across source and translated versions of the DP science examination items.

- **Account for textual complexity associated with word choice**
- 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.
- High frequency word lists are available from a variety of linguistic resources. It would be beneficial to consult high frequency word lists when making translation or item design decisions regarding word choice, particularly accounting for the relative frequency of words chosen when translating text.
- Accounting for the relative frequency of words across languages can be aided by the use of NLP software such as *ReaderBench*. Pre-screening items using textual analysis software can give an indication of the number of unique words in each language and can aid in identifying whether there are items that may present additional challenge in a specific language version.
- **Account for textual complexity associated with sentence length**
- When considering sentence length, 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.
- An important consideration for the translation of items is the relative sentence length across languages. Accounting for this can be aided by the use of NLP software such as *ReaderBench*. Pre-screening items using textual analysis software can give an indication of the number of words across sentences in each language and can therefore assist in identifying whether there are items that may present an additional challenge in a specific language version.
- **Account for textual complexity associated with structural complexity**
- When developing items, care should be taken to use 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.
- NLP software such as *ReaderBench* can be used to pre-screen items to understand the relative frequency of certain parts of speech in different language versions and can therefore assist in identifying whether there are items that may present additional challenge in a specific language version.
- As far as possible when designing items, it is beneficial to 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 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.

Conclusion

The overarching conclusion from this research study is 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. This was supported by the small percentages of medium and large language DIF observed across the six subjects and by the expert review of the different language versions of the DIF items where almost all the items were judged as being very similar or the same across the languages for multiple translation and linguistic criteria. Therefore, it appears that the current IB translation processes involving forward translation and review and revision, drawing on both translation and IB curriculum and 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, even after controlling for related student factors (gender, sub-region and first language match), 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 language DIF.

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 is 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 even 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 may be meaningfully applied. Nonetheless, even in the context of DP science examinations and their translation, the NLP analysis has been revealing and could be easily integrated as a screening tool in the IB translation process, as well as inform the translators about features of text that affect its complexity both within and across languages and which they may otherwise not attend to. 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.

References

- Asil, M., & Brown, G. T. (2016). Comparing OECD PISA reading in English to other languages: Identifying potential sources of non-invariance. *International Journal of Testing*, 16(1), 71-93.
- Borgwaldt, S. R., Hellwig, F. M., & De Groot, A. M. B. (2005). Onset entropy matters—Letter-to-phoneme mappings in seven languages. *Reading and Writing* (Vol. 18, pp. 211-229). Springer.
- Chen, A. C.-H. (2016). A critical evaluation of text difficulty development in ELT textbook series: A corpus-based approach using variability neighbor clustering *System* (Vol. 58, pp. 64-81).
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3), 181-204.
- Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *Tesol Quarterly*, 42(3), 475–493.
- Dascalu, M., Crossley, S. A., McNamara, D. S., Dessus, P., & Trausan-Matu, S. (2018). Please *ReaderBench* this text: a multi-dimensional textual complexity assessment framework. In *Tutoring and intelligent tutoring systems* (pp. 251-271). Nova Science Publishers, Inc.
- Dascalu, M., Gutu, G., Ruseti, S., Paraschiv, I. C., Dessus, P., McNamara, D. S., Crossley, S., & Trausan-Matu, S. (2017a, September). *ReaderBench*: a multi-lingual framework for analyzing text complexity. In *European Conference on Technology Enhanced Learning* (pp. 495-499). Springer, Cham.
- Dascalu, M., Westera, W., Ruseti, S., Trausan-Matu, S., & Kurvers, H. (2017b, June). *ReaderBench* learns Dutch: building a comprehensive automated essay scoring system for Dutch language. In *International Conference on Artificial Intelligence in Education* (pp. 52-63). Springer, Cham.
- El Masri, Y., & Andrich, D. (2020). The trade-off between model fit, invariance, and validity: The case of PISA science assessments. *Applied Measurement in Education*, 33(2), 174-188.
- El Masri, Y. H., Baird, J. A., & Graesser, A. (2016). Language effects in international testing: The case of PISA 2006 science items. *Assessment in Education: Principles, Policy & Practice*, 23(4), 427-455.
- El Masri, Y. Y. H., Ferrara, S., Foltz, P. W. P., & Baird, J.-A. (2017). Predicting item difficulty of science national curriculum tests: the case of key stage 2 assessments. *Curriculum Journal*, 28(1), 59–82.
- Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing*, 5(1), 23-35.
- Ferrari, A., & Dept, S. (2020). *Pros and cons of back translation in assessments and surveys (webinar)*. cApStAn LQC.
- Galache Ramos, M. (2017). *Is my good better than yours? An exploration of examiners' interpretation of 'common terms' in criterion-referenced assessment in the International Baccalaureate Programme*. (Unpublished Master's Dissertation). University of Bath, Bath, United Kingdom.
- Graesser, A. C., Cai, Z., Louwerse, M. M., & Daniel, F. (2006). Question Understanding Aid (QUAID). A web facility that tests question comprehensibility. *Public Opinion Quarterly*, 70(1), 3–22.

- Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, 40(5), 223–234.
- Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H., & Pennebaker, J. (2014). Coh-Metrix Measures Text Characteristics at Multiple Levels of Language and Discourse. *The Elementary School Journal*, 115(2), 210–229.
- Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international assessment. *Language Testing*, 20(2), 225–240.
- Grisay, A., De Jong, J. H. A. L., Gebhardt, E., Berezner, A., & Halleux-Monseur, B. (2007). Translation equivalence across PISA countries. *Journal of Applied Measurement*, 8(3), 249.
- Grisay, A., Gonzalez, E., & Monseur, C. (2009). Equivalence of item difficulties across national versions of the PIRLS and PISA reading assessments. *IERI*, 2, 63–83.
- Grömping, U. (2009). Variable importance assessment in regression: linear regression versus random forest. *The American Statistician*, 63(4), 308–319.
- Gutu, G., Dascalu, M., Trausan-Matu, S., & Dessus, P. (2016, September). *ReaderBench* goes online: a comprehension-centered framework for educational purposes. In *13th International Conference on Human-Computer Interaction (RoCHI 2016)* (pp. 95–102). MATRIX ROM.
- Gutu-Robu, G., Sirbu, M. D., Paraschiv, I. C., Dascălu, M., Dessus, P., & Trausan-Matu, S. (2018). Liftoff–*ReaderBench* introduces new online functionalities. *Romanian Journal of Human-Computer Interaction*, 11(1), 76–91.
- Halliday, M. A. K., & Martin, J. R. (1993). *Writing Science: Literacy and Discursive Power*. London: The Falmer Press.
- Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting Educational and Psychological Tests for Cross-Cultural Assessment* (pp. 3–38). Psychology Press.
- Han, Z., He, Q., & von Davier, M. (2019). Predictive feature generation and selection using process data from PISA interactive problem-solving items: An application of random forests. *Frontiers in Psychology*, 10, 2461.
- Hauger, J. B., & Sireci, S. G. (2008). Detecting differential item functioning across examinees tested in their dominant language and examinees tested in a second language. *International Journal of Testing*, 8(3), 237–250.
- Huang, X., Wilson, M., & Wang, L. (2016). Exploring plausible causes of differential item functioning in the PISA science assessment: language, curriculum or culture. *Educational Psychology*, 36(2), 378–390.
- International Test Commission. (2017). *ITC Guidelines for Translating and Adapting Tests (Second edition)*. [www.InTestCom.org]
- Kiefer, T., Robitzsch, A., & Wu M. (2020). *TAM: Test Analysis Modules*. R package version 3.5-19, URL <https://CRAN.R-project.org/package=TAM>.

- Kreiner, S., & Christensen, K. B. (2014). Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika*, 79(2), 210-231.
- Kuboň, V., Lopatková, M., Plátek, M., & Pognan, P. (2006, September). Segmentation of complex sentences. In *International Conference on Text, Speech and Dialogue* (pp. 151-158). Springer, Berlin, Heidelberg.
- Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28(i05).
- Le, L. T. (2009). Investigating gender differential item functioning across countries and test languages for PISA science items. *International Journal of Testing*, 9(2), 122-133.
- McCarthy, M., & O'Keeffe, A. (2010). *The Routledge Handbook of Corpus Linguistics*. Taylor & Francis Group.
- McGrane, J. A., Butow, P. W., Sze, M., Eisenbruch, M., Goldstein, D. & King, M. T. (2014). Assessing the invariance of a culturally competent multi-lingual unmet needs survey for immigrant and Australian-born cancer patients: a Rasch analysis. *Quality of Life Research*, 23(10), 2819-2830.
- Oliveri, M. E., Olson, B. F., Ercikan, K., & Zumbo, B. D. (2012). Methodologies for investigating item- and test-level measurement equivalence in international large-scale assessments. *International Journal of Testing*, 12(3), 203-223.
- Paek, I., & Wilson, M. (2011). Formulating the Rasch differential item functioning model under the marginal maximum likelihood estimation context and its comparison with Mantel-Haenszel procedure in short test and small sample conditions. *Educational and Psychological Measurement*, 71(6), 1023-1046.
- Penfield, R., & Camilli, G. (2007). Test fairness and differential item functioning. *Handbook of statistics*, 26, 125-167.
- Robinson, M., Johnson, A. M., Walton, D. M., & MacDermid, J. C. (2019). A comparison of the polytomous Rasch analysis output of RUMM2030 and R (ltm/erM/TAM/lordif). *BMC medical research methodology*, 19(1), 36.
- Sandilands, D., Oliveri, M. E., Zumbo, B. D., & Ercikan, K. (2013). Investigating sources of differential item functioning in international large-scale assessments using a confirmatory approach. *International Journal of Testing*, 13(2), 152-174.
- Schittekatte, M., Resing, W., Kort, W., Vermeir, G., & Verhaeghe, P. (2003). The Netherlands and Flemish-speaking Belgium. In J. Georgas, L. G. Weiss, F. J. R. van de Vijver, & D. H. Saklofske (Eds.), *Culture and Children's Intelligence: Cross-Cultural Analysis of the WISC-III* (pp. 109–119). Academic Press.
- Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, 38(5), 553-573.
- Smith, G. (2018). Step away from stepwise. *Journal of Big Data*, 5(1), 1-12.
- Spencer, K. A. (2007). Predicting children's word-spelling difficulty for common English words from measures of orthographic transparency, phonemic and graphemic length and word frequency.

British Journal of Psychology, 98, 305-338.

- Tewes, U. (2003). Germany. In J. Georgas, L. G. Weiss, F. J. R. van de Vijver, & D. H. Saklofske (Eds.), *Culture and Children's Intelligence: Cross-Cultural Analysis of the WISC-III* (pp. 129–135). Academic Press.
- Wiliam, D. (1994). Creating matched National Curriculum assessments in English and Welsh: test translation and parallel development. *The Curriculum Journal*, 5(1), 17-29.
- Wu, A. D., & Ercikan, K. (2006). Using multiple-variable matching to identify cultural sources of differential item functioning. *International Journal of Testing*, 6(3), 287-300.
- Xie, Y., & Wilson, M. (2008). Investigating DIF and extensions using an LLTM approach and also an individual differences approach: an international testing context. *Psychology Science*, 50(3), 403.
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2), 301-320.
- Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language assessment quarterly*, 4(2), 223-233.