frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 25 August 2025
DOI 10.3389/feduc.2025.1616879

Construct comparability and the limits of *post hoc* modeling: insights from International Baccalaureate multi-language assessments

Louise Badham[1,2]*, Michelle Meadows[2] and Jo-Anne Baird[2]

[1]International Baccalaureate, Assessment Research & Design, Cardiff, United Kingdom, [2]Oxford University Centre for Educational Assessment, Oxford University, Oxford, United Kingdom

OPEN ACCESS

EDITED BY
Corey Palermo,
Measurement Incorporated, United States

REVIEWED BY
Alejandro Veas,
University of Murcia, Spain
Frank Padellaro,
Dover Measurement Group, United States

*CORRESPONDENCE
Louise Badham
✉ louise.badham@education.ox.ac.uk

RECEIVED 23 April 2025
ACCEPTED 31 July 2025
PUBLISHED 25 August 2025

CITATION
Badham L, Meadows M and Baird J-A (2025) Construct comparability and the limits of *post hoc* modeling: insights from International Baccalaureate multi-language assessments. *Front. Educ.* 10:1616879.
doi: 10.3389/feduc.2025.1616879

Construct comparability was investigated across different subjects in the International Baccalaureate (IB) Diploma Programme (DP). A Rasch Partial Credit Model (PCM) was applied to historical assessment data to generate statistical measures of the relative "difficulty" of IB subjects and languages. Specifically, analysis centered on different language versions of literature assessments, where exams differ in content, but are designed to assess the same target constructs. Rasch analyses were conducted sequentially in three subsets of data. Three different conceptualizations of the linking construct were compared, with the aim of narrowing the definition to increase the validity of the comparisons. These ranged from different DP subjects being linked by "general academic ability," to linking English, Spanish and Chinese language versions of literature with the more relevant construct of "literary analysis." Ultimately, the Rasch analyses produced three different rank orders of "difficulty" for the assessments, illustrating the limitations of *post hoc* construct comparability investigations. Whilst literary analysis is the most theoretically defensible linking construct in this context, the approach relies on bilingual students taking different language versions of the assessments and therefore has limited operational applicability. There are also conceptual limitations, as bilingual examinees are not representative of all students in DP cohorts. Further research is recommended into how cohort characteristics can impact performance, as well as how constructs are defined for use across linguistic and cultural subgroups. Such investigations are crucial to avoid construct bias being introduced in the earliest stages of assessment design.

KEYWORDS

assessment, comparability, International Baccalaureate, Rasch model, test adaptation

1 Introduction

International awarding bodies like the International Baccalaureate (IB) face challenges in ensuring assessments are suitable for linguistically and culturally diverse student cohorts. International assessment guidelines such as the American Educational Research Association (AERA)'s *Standards for Educational and Psychological Testing*, suggest that for assessments to be "fair" in linguistically and culturally diverse contexts, validity evidence must be gathered to demonstrate that "different language versions measure comparable or similar constructs" (American Educational Research Association [AERA], 2014, p. 69). Additionally, final grades need to be comparable across languages so that meaningful comparative inferences can be made about results (Ercikan and Lyons-Thomas, 2013). Construct comparability across assessments in multiple languages (hereafter referred to as multilingual assessments) is a central concern for the IB, which delivers high-stakes assessments in multiple languages in over 150 countries.

This article explores the utility of Rasch analysis for investigating construct comparability across assessments in different IB Diploma Programme (DP) subjects and languages. In particular, it focuses on different language versions of DP literature assessments. The IB offers DP literature in approximately 75 different languages each year, and develops exam content separately in each language according to a common test specification. Marking and grade awarding processes are also separate in each language, with alignment supported by examiner training and cross-language standardization processes. DP scores are primarily used for university entrance purposes (IBO, 2019a), so grades across languages are expected to represent equivalent levels of attainment. By focusing on the DP Studies in language and literature, this study targets an underexplored discipline in multilingual assessment comparability. Furthermore, the investigation focuses on assessments with a lesser used model of multilingual test development, where content is produced separately in each language according to a common test specification, rather than being translated from a source language.

Statistical analyses of assessments that are translated (adapted) into different languages can miss systematic bias (El Masri et al., 2016). This study aims to demonstrate that similar challenges exist for multilingual assessments developed in parallel, where exam content differs across languages that share a common curriculum and assessment model. In both approaches, systematic bias may be introduced if constructs are "limited by designers' understandings of the network of knowledge, skills, and dispositions that inform these constructs" (Randall et al., 2022, p. 173). If the target construct is narrowly defined in the earliest stages of conceptualization, this can reverberate through all stages of the assessment lifecycle, systematically disadvantaging certain groups of students. The present study illustrates the limitations of *post hoc* analyses, which cannot easily detect such issues of bias.

Ultimately, it will be argued that a theoretically driven approach to selecting constructs used to link assessments and compare their levels of difficulty is essential. As such, this paper extends previous literature on the theory of linking constructs (e.g., Newton, 2005, 2010). This article also contributes to the field empirically, by comparing statistically, the difficulty of high-stakes literature assessments offered in different languages by the IB. Indeed, research on multilingual tests developed in parallel is scarce. Much of the academic literature on sources of bias focuses on the difficulty of items for different groups of students rather than tests, but here we add to the methodological literature on analysis of test level difficulty. Specifically, there is an extension to previous literature (e.g., He et al., 2018) on inclusion and representativeness of data (tests and learners) fitting the Rasch model and how this influences interpretation. Since assessment comparability underpins many of the uses of assessment data, a firm understanding of the ways in which our theories, analytical techniques and data affect our interpretation and potentially subsequent policies is crucial.

2 Background

2.1 Cross-lingual and cross-cultural assessment comparability

Cross-lingual assessment comparability is difficult to achieve due to challenges inherent in producing different language versions of exams. These include linguistic issues such as errors introduced when items are translated, or linguistic variation making content more cognitively demanding in one language than another (El Masri et al., 2016; Ercikan and Lyons-Thomas, 2013; Ercikan and Por, 2020). Such challenges may be compounded by different social or cultural dynamics in translation review teams leading to different conclusions and approaches to error detection in translated assessments (Zhao and Solano-Flores, 2023). Other aspects may introduce bias against linguistic, sociocultural and racial groups, including culturally-specific content causing misunderstandings (Lerner, 2021), or supposedly "neutral" language in item writing marginalizing subgroups by perpetuating colonialist narratives (Randall, 2021). There are also sociolinguistic considerations such as dialectical differences and regional variation of language (Solano-Flores and Li, 2009), and linguistic complexity of items creating construct-irrelevant barriers for students with certain language backgrounds (Oliveri, 2019). Such challenges make the aim of ensuring comparability across exams in different languages and cultures an extraordinarily complex endeavor. Inherent differences across languages mean that strict equating is impossible in cross-lingual assessment, therefore "weaker" linking approaches are more appropriate in these contexts (Sireci et al., 2016). Such links are frequently supported by statistical techniques to evaluate comparability across different language versions with the aim of minimizing bias, supporting comparability, and enhancing validity in multi-language assessments (Badham et al., 2025; Sireci et al., 2016).

Bias can be introduced at different stages of the assessment lifecycle, including during the selection of translation methods, the process of adapting assessments into different languages, or using statistical techniques to compare scores across languages (Ercikan and Lyons-Thomas, 2013; Oliveri et al., 2015). It is therefore important to investigate potential issues of bias, which may otherwise invalidate cross-lingual assessment comparisons (Sireci et al., 2016). Of the possible sources of bias, *construct bias* arguably poses the greatest threat to validity in multilingual assessments (Hambleton, 2005; van de Vijver and Poortinga, 2004). Clear and inclusive definitions of target constructs in assessments are essential to avoid narrow constructs being "baked into" assessment programs and introducing systematic bias (Randall et al., 2022).

2.2 The parallel development problem: limitations of traditional methods

Research designs for linking assessments across languages can broadly be classified into three categories: (a) *separate monolingual group designs*, where each language version is taken separately by each language group; (b) *bilingual group designs*, where bilingual examinees take both language versions of the exam; and (c) *matched monolingual group designs*, where the different groups of examinees are matched according to external indicators (e.g., demographic variables) (Sireci, 1997). Separate monolingual group designs are most typical (Badham et al., 2025), with the most common technique for investigating cross-lingual assessment comparability being differential item functioning (DIF) (Zumbo, 2003).

Techniques such as DIF are used to investigate whether differences across adapted assessments make items more cognitively demanding in one language compared to another (e.g., Grisay et al., 2009; Oliveri et al., 2013; Elosua and López-jaúregui, 2007). These analyses can be used to identify items that perform differently across subgroups (e.g., language of the test taken) by comparing how each group performed on individual questions compared with their overall score (Keng and Marion, 2020). Subsequent investigations can then explore the reasons for differences, such as via expert reviews. In computer-based testing, response data such as eye-tracking and computer log files also provide further opportunities to investigate differences across languages, such as whether response time is longer in one language compared to another (Hlosta et al., 2024). Such investigations are important for identifying more "difficult" items, so valid inferences can be made about the comparability of scores across languages (Ercikan and Oliveri, 2016).

However, analyses carried out at item-level *post hoc* can miss systematic issues of bias that permeate assessments (El Masri et al., 2016). If items perform differently across languages, it may be caused by translation error or semiotic difference. However, if the issue were at a construct rather than item-level, it would not be detected in item-level analyses, as the assessment would be saturated with bias. This can be particularly challenging in assessments targeting more complex, skills-based constructs. Oliveri et al. (2019) highlighted the complexity of assessing collaborative problem-solving skills in Arabic and English, where implications of tone formality vary due to Arabic being a diglossic[1] language. Marking criteria generated in a source language and translated into others can systematically favor one group but may go undetected in DIF analyses.

Methods for examining comparability across language versions also vary depending on how multi-language exams are developed (Badham et al., 2025). Multi-language test development approaches include *adaptation* where content is developed in one source language and directly translated into another, and *parallel development* where content is developed independently in each language according to a common blueprint (Ercikan and Lyons-Thomas, 2013; Oliveri et al., 2015). A high-level overview of the adaptation and parallel development approaches used in the IB is illustrated in Figure 1.

Adaptation is by far the most common approach to multilingual test development and is used for most international assessments (Ercikan and Por, 2020). However, parallel development may be more appropriate if the target construct cannot validly be assessed by adapting content across languages (Ercikan and Lyons-Thomas, 2013). Different test versions are generally recommended if constructs are not considered generalizable across different populations of examinees (Hernández et al., 2020). Whilst many IB multi-language assessments such as maths and the sciences use adaptation, parallel development is used in subjects like literature, due to literary content being inherently different across languages. However, a limitation of the parallel approach is that it is harder to investigate comparability across different language versions, and links established across the different versions will naturally be weaker (Badham et al., 2025).

In a parallel approach, rather than constructs designed to be comparable through common, translated items, how constructs are defined and represented on different language versions the assessment is intended to be comparable (ibid). The IB provides "paper setting instructions"; common guidance to exam authors across languages. These provide details relating to assessment content (reflecting the curriculum for each subject), number of marks, and format of exams (IBO, 2019a). Exam development is managed by IB assessment staff who support and train exam authors across languages and are ultimately responsible for the exam content. They lead "assessment editing meetings" where authors from different language versions often share and feedback on draft exam content (typically using English translations to access one another's languages). This parallel approach to multilingual exam development necessitates alternative approaches to investigating cross-lingual construct comparability in assessments. Yet, there remains a scarcity of empirical research investigating the comparability of multilingual assessments following a parallel development model.

2.3 Exploring inter-subject comparability with Rasch

When exams are developed, marked and graded separately in each language, the different language versions are in many ways more analogous to different subjects, rather than different language variants of a common subject. For example, English literature and Spanish literature may be considered separate subjects, compared to chemistry in French versus chemistry in German being language variants of the same subject. As such, methods and approaches used to explore inter-subject comparability may be more appropriate to explore in parallel development contexts, rather than traditional, cross-lingual assessment methods (such as DIF) which are intended for exams adapted from one language into another. Rasch analysis has often been used in inter-subject comparability studies, and therefore may be a suitable method for investigating comparability in multilingual, parallel test development.

Rasch models, originally developed by mathematician Georg Rasch (1960), comprise a collection of probabilistic models used to predict outcomes based on response patterns in tests and surveys. Rasch models operate by placing person abilities and item difficulties on the same measurement scale, based on patterns of responses from respondents to a set of items. They are used for many different purposes, including linking different forms of assessments that involve the same trait or construct (Boone et al., 2014). In the psychometrics tradition, Rasch analysis assumes a unidimensional underlying latent trait—that is, items are designed to measure a single, common attribute. However, the IB operates within a *curriculum-based assessment* paradigm, which differs from psychometrics approaches in that: "rather than attempting to assess an underlying trait, the attribute of interest is performance on assessments, which is assumed to be caused by the knowledge and skills…gained through studying the curriculum" (Baird and Opposs, 2018, p. 12). This is closely aligned to the assessment approach for A-Level and GCSE exams in the UK, where it has been claimed that Rasch analyses may be used to indicate "the correspondence between the grades achieved in a subject and the underlying construct of general academic capacity for achievement" (Coe et al., 2008, p. 78). Rasch modeling has consequently been used to investigate inter-subject comparability—where exam content is necessarily different across disciplines—based on the assumed latent trait of "general academic ability" (e.g., Coe, 2008; He et al., 2018; He and Black, 2019; Ofqual, 2022). Rasch models have also been used to examine inter-subject comparability in other

[1] Where two languages or varieties of language are used by one language community, with one being more formal or prestigious, and the other more informal.

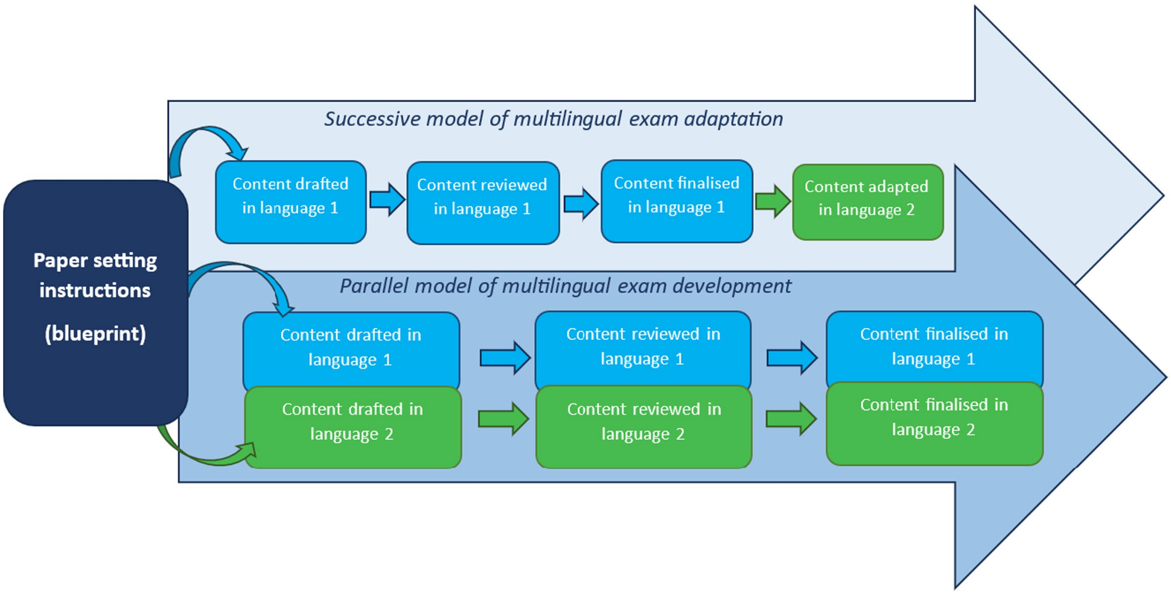

Successive model of multilingual exam adaptation
Content drafted in language 1
Content reviewed in language 1
Content finalised in language 1
Content adapted in language 2
Paper setting instructions (blueprint)
Parallel model of multilingual exam development
Content drafted in language 1
Content reviewed in language 1
Content finalised in language 1
Content drafted in language 2
Content reviewed in language 2
Content finalised in language 2

FIGURE 1
Overview of IB parallel & successive exam development.

assessment systems, such as university entrance exams in Spain (Veas et al., 2020), as well as in Tasmania and New South Wales in Australia (Lamprianou, 2009).

The use of these techniques has been controversial, with much debate about what is meant by "inter-subject comparability" (Coe, 2008; Newton, 2012), as well as questions raised about the "practicality and the transparency of using such a complex model for so high stakes purposes" (Lamprianou, 2009, p. 211). Nevertheless, Rasch has been a useful tool for exploratory analyses investigating comparability issues. The outcomes of Rasch inter-subject comparability investigations can have significant implications, including national educational policymaking decisions. In the UK, for example, extensive comparability investigations—including Rasch analyses to generate statistical estimates of the relative difficulty of A-level subjects—resulted in a more lenient grade awarding approach in some subjects (Ofqual, 2018). As such, Rasch has become an established method for exploring inter-subject comparability in the UK context, to the extent that it can impact assessment policy, practices and processes. Therefore, it is useful to explore whether this approach can be used to investigate construct comparability in multilingual assessments where content is different in each language.

The first step would be to decide the most appropriate linking construct across the assessments. In most inter-subject comparability studies, the latent trait is necessarily conceptualized at a high level as "general academic ability." Therefore, a limitation of these studies is that the common trait (i.e., the linking construct) cannot reflect the construct as it is defined in a particular subject (He et al., 2018). However, it may be possible to define the trait (or construct) more precisely when applied across cognate subjects, or different languages in the same subject—such as the construct of "literary analysis" in literature. This study applied Rasch analysis across languages in DP literature to investigate construct comparability across languages, and explore the impact of the linking construct on findings.

2.4 Aims

The overarching aim of this study was to explore construct comparability across different language versions of DP assessments developed in parallel, focusing on theoretical approaches for selecting and defining constructs to link multilingual assessments. The study was guided by the following research questions:

1. To what extent can Rasch analyses provide evidence of construct comparability in parallel-developed multilingual assessments in the IB Diploma Programme?
2. How do different IB data subsets and varying conceptualizations of the linking construct affect comparability inferences across subjects and language versions?
3. What are the strengths and limitations of relying on *post hoc* statistical analyses to evaluate comparability across different language versions of parallel-developed assessments?

This study was based on the hypothesis that inherent differences in how complex constructs such as literary analysis manifest across culturally and linguistically diverse contexts limit the validity of relying solely on *post hoc* statistical analyses to establish comparability. It is therefore expected that such analyses, while informative, must be interpreted alongside other sources of evidence to support fair and valid assessment outcomes for diverse student populations.

3 Methods

A Rasch Partial Credit Model (PCM) was applied to historical IB assessment data to determine measures of relative difficulty across language variants of DP literature. The original Rasch model was designed for dichotomously scored items, and the PCM was developed for application to polytomous items—i.e., items that have two or more ordered response categories. The PCM may be expressed mathematically as:

$$\frac{P_{ijx}}{P_{ijx-1}+P_{ijx}}=\frac{\exp\left(\theta_j-\delta_{ix}\right)}{1+\exp\left(\theta_j-\delta_{ix}\right)},x=1,2,\ldots,m_i$$

Where "P_{ijx} is the probability of the person j scoring x on item i, P_{ijx-1} is the probability of person j scoring $x-1$, θ_j is the ability of the person j, and δ_{ix} is an item parameter governing the probability of scoring x rather than $x-1$ on item i" (Masters and Wright, 1996, p. 102), and m is the maximum available score for item i. In inter-subject comparability studies, the Rasch PCM has been employed to explore the relative difficulty of subjects, with each subject treated as a polytomous item, and corresponding grades as ordered response categories. DP subjects are graded on a scale of 1 (low achievement) to 7 (high achievement). Therefore, subjects were treated as polytomous items with seven ordered response categories, to allow the Rasch analyses to be performed. For each stage of the analysis, a Rasch PCM with Joint Maximum Likelihood Estimation (JMLE) was fitted to the data using the *"TAM"* package in R statistical software (Robitzsch et al., 2022).

Model fit was judged based on outfit and infit statistics for items and persons. Whilst infit and outfit statistics are useful in identifying patterns in item and person functioning within Rasch models, they do not provide a complete picture of overall model fit. Nonetheless, they are "convenient quantitative measures of fit discrepancy" (Wright and Linacre, 1994, p. 370), and are widely used to examine whether empirical data demonstrate sufficient fit to Rasch models. A fit statistic of >1 indicates more variation between the observed and predicted response patterns than expected, which contributes toward *misfit*. The specific interpretation of what is deemed to be misfitting depends on the measurement context (Bond and Fox, 2007). For item fit, Coe's approach of requiring an "infit and outfit below 1.7 and at least one grade category with outfit of 1.5 or less" (Coe, 2008, p. 618) was applied, due to the similarity of context and purpose of investigation. However, it should also be noted that fit thresholds are somewhat arbitrary, and different fit thresholds could also be theoretically justified (e.g., He et al. (2018) set the threshold as 2.0[2] when comparing GCSE standards in England). Consequently, a different cut-off would result in different subjects being included in the model (ibid). Person fit was not specified by Coe, so this cut-off was set at 2.0, which is common to ensure the measurement system is not distorted (Linacre, 2002), and has been used in other Rasch inter-subject comparability studies (e.g., He et al., 2018). Whilst these statistics are a useful indicator of how well the data fit the model, they can lack accuracy (Karabatsos, 2000). This is acknowledged as a limitation, but was deemed sufficient for the pragmatic, exploratory purposes of the current study.

Unidimensionality is a central assumption of the Rasch model. Other than that explained by the attribute (or construct) of interest, it is assumed that there should be no systematic pattern of residuals as the remaining variation should be random (Andrich and Marais, 2019). Unidimensionality was investigated via a Principal Component Analysis of Residuals (PCAR), which examines which group of items measure a common trait (ibid). Another assumption of the Rasch model is local independence of items—that performance on one item should not directly impact performance on another. Local dependence of items in Rasch PCMs is often tested using Yen's Q_3 statistic by examining residuals on pairs of items (Christensen et al., 2017). However, "items" here comprise whole subjects,[3] and would not be expected to be dependent on one another as items may be within the same test (e.g., one question helping examinees respond correctly to a subsequent question). Moreover, local dependence does not typically impact the ordering of measures (Baghaei, 2008), which was the primary interest of the current study.

3.1 Subject selection

Only subjects with over 1,000 students were included in the analysis (except some of the Chinese subjects under investigation). This followed approaches in previous Rasch studies exploring A-level subject difficulty (He et al., 2018; He and Black, 2020; Ofqual, 2022). Rasch models assume some common frame of reference between tests to allow comparability to be explored. This commonality may comprise a common set subjects taken by multiple students, to allow comparisons of how different students perform on the same subjects. Pragmatically, as students usually choose different subject options in educational settings, this often involves subjects with a high degree of overlap in inter-subject comparability studies (i.e., many students taking popular subject combinations, such as maths, sciences or English). Alternatively, the commonality may consist of common students taking all the same subjects. The Rasch analysis in the present study was conducted in three stages: two based on common subjects and one on common persons.

Each DP subject can be taken at standard (SL) or higher level (HL). Most DP students choose three SL subjects and three HL subjects. The levels differ slightly in scope, and HL includes more teaching hours and students are expected to demonstrate broader knowledge and skills (often through an additional assessment task), so students' HL subjects tend to include disciplines they intend to study at university level. However, HL and SL are measured according to the same grade descriptors and subjects at both levels contribute equally to their final DP score (IBO, 2019a), so the levels do not necessarily imply a pre-requisite in terms of academic achievement. Thus, both higher and standard levels for each subject were included in the analysis to allow for a greater evidence base to explore comparability in the IB context.

The number of subjects in the model was reduced in each stage of analysis, as the definition of the underlying linking construct was increasingly narrowed and refined. Three subsets of DP assessment data were analyzed (Table 1). The first included subjects across the DP,

[2] Indeed, an item fit threshold of 2.0 was considered for the present study to be consistent with He et al. However, all subjects with fit statistics between 1.7 and 2.0 had a theoretical justification for being excluded from the model (as detailed in the results), so the 1.7 threshold was deemed most appropriate for the context of this study.

[3] As such, we refer to "subject" here to more accurately describe the objects of investigation, but they may be considered "items" for the purposes of the model.

TABLE 1 Rasch analysis stages & datasets.

Level of assessment selection	Commonality	No. subjects in final model (# students)	Focus	Assumed linking construct
1. Programme	Items (subjects)	41 (81,252)	Popular subjects across DP	General academic ability
2. Assessment objectives	Items (subjects)	18 (83,422)	Popular subjects, aligned by assessment objectives	Contextualized & evidenced argumentation
3. Subject	Persons (students)	12 (3,200)	Students taking literary subjects in two different languages	Literary analysis

the second was limited to subjects that targeted similar skills, and the third included only literary subjects that were of primary interest for the study.

The final stage of analysis focused solely on different language versions (English, Spanish and Chinese) of DP Studies in language and literature subjects, which shared the same overall assessment model (i.e., blueprint). For the May 2019 exam session—the focus of the present study—these subjects comprised the following assessments:

- **Paper 1 (20%/25%[4])**: written exam requiring one essay-based response. Students write a textual analysis of previously unseen source texts.
- **Paper 2 (25%)**: written exam comprising one essay based on literary texts studied in class. Students choose from a selection of essay-based questions.
- **Coursework (20%/25%)**: written task based on texts studied in class.
- **Oral exam (30%)**: oral commentary followed by questions from the teacher on a text studied in class (15%), and an individual oral presentation/activity (15%).

Despite small variations between levels (e.g., HL students could discuss three texts in Paper 2, vs. only two at SL), and subjects (e.g., literary vs. non-literary stimuli in Paper 1), the assessment models were broadly the same across levels and subjects. These assessments collectively contribute to the students' overall subject grades.

All assessments comprised extended responses (either written or oral) and were marked using the same criteria across all languages. IB subject grades are determined through a "grade awarding" process, whereby students' marks on individual assessments are transformed into an overall grade that represents their level of achievement in the subject (Badham, 2025). This involves both quantitative analysis (e.g., statistical data on performance trends over time) and qualitative evidence (e.g., expert review of student work against grade descriptors) to identify the "'turning point' on the original mark scale where the quality of the work shifts from one grade to the next" (ibid, p. 21). This is an "attainment referencing" approach to grade awarding. Students' grades are based on their performance in an individual subject whilst taking into account contextual factors such as easier or harder papers, or more or less stringent marking (Newton, 2011). For example, lower grade boundaries would be set to account for any severity in the marking. Consequently, rater stringency on individual assessments should be accounted for during the award of grades. Nonetheless, theoretically, our findings could be the product of differences in the marking or grading standard, or other factors such as construct differences across languages.

IB grade award focuses on three key "judgemental grade boundaries": the 2/3, 3/4, and 6/7 boundaries (IBO, 2019a), with remaining boundaries calculated through interpolation. As each boundary involves integrating statistical and judgemental evidence, it is theoretically possible for an awarding standard in a subject to be harsher at one boundary (e.g., grade 7) and more lenient in another (e.g., grade 4). Grade 7 is the focus here since it represents the highest level of attainment (difficulty) and is often crucial for university admissions in specific subjects. Other Rasch studies investigating individual grades in inter-subject comparisons have cautioned that interpretations need to be "specific to particular ability ranges" (Coe, 2008, p. 627).

4 Results

The Rasch model was applied to assessment data from the DP May 2019 exam session. As the analysis required commonality of subjects, only full Diploma students (who typically take exams across six different DP subjects) were included. The language versions of the DP Studies in language and literature subjects with the largest numbers of students (English, Spanish, and Chinese) were prioritized (Table 2).

4.1 DP dataset (1): "general academic ability"

4.1.1 Analytical strategy

Following an approach used in previous inter-subject comparability studies (e.g., Coe, 2008; He et al., 2018; Ofqual, 2022), the first stage of analysis involved comparing subjects across all academic groups based on the theoretical linking construct of "general academic ability." This is based on what Coe (2007, 2008) termed "general academic ability," "achievement" or "aptitude," which assumes the construct of interest is more a "generalisable capacity for learning" (Coe, 2007, p. 351), rather than a construct as defined in a particular subject. As such, this may be considered a *generalized approach* to construct comparability. The Rasch model was therefore applied to

[4] Paper 1 in literature contributed 20% to the overall subject grade, and Paper 1 in language & literature contributed 25% (and vice versa for the coursework). For further details, see supplementary materials.

TABLE 2 Student numbers for Chinese, English and Spanish DP Studies in language & literature subjects, May 2019.

Language/subject	Level	# Students
Chinese language & literature	Higher	678
	Standard	1,459
Chinese literature	Higher	538
	Standard	1,469
English language & literature	Higher	28,272
	Standard	17,414
English literature	Higher	39,740
	Standard	6,641
Spanish language & literature	Higher	2,756
	Standard	1,462
Spanish literature	Higher	6,307
	Standard	706

common (i.e., most popular) subject combinations from across the DP, including subjects as diverse as literature, chemistry, psychology and economics. IB students typically select six DP subjects, each from a different academic group[5], to receive a full Diploma. Popular subjects (e.g., history and maths) served as a proxy for common items.

Overall, 60 subjects[6] based on results from 84,818 DP students were first modeled, then reviewed to judge goodness of fit. These students represented schools from 141 different countries, 53% of whom came from the five most common countries: the United States ($n = 29{,}350$), Canada ($n = 4{,}359$), the UK ($n = 3{,}931$), India ($n = 3{,}666$), and China ($n = 3{,}556$). Approximately 56% of the student population were reported as female ($n = 47{,}552$), 44% as male ($n = 37{,}242$), and 24 did not specify a gender. DP language acquisition subjects regularly misfitted, which may be explained by "misplacement" of students in these courses. Language acquisition subjects are designed for language learners and are hierarchical in nature: language *ab initio* intended for beginners, language B standard level for intermediate, and higher level for advanced learners. Appropriate placement of students in these courses is important to ensure they are suitably challenged (IBO, 2021). Notwithstanding, students are sometimes placed in subjects not best suited to their linguistic abilities (e.g., native speakers placed in language B higher level). Thus, the alignment between student ability and subject difficulty can be distorted.

Music and visual arts also misfitted, which was consistent with fit issues found with creative subjects in previous Rasch inter-subject comparability studies (e.g., Coe, 2008) and is perhaps unsurprising given the difference in skills assessed in arts-based subjects. Finally, Chinese literature standard level had good fit at an overall subject level, but a significant fit issue at grade 4, likely due to the very small proportion of students in the lower grades (with only 0.96% receiving grades 1–3). Consequently, language acquisition subjects ($n = 14$), music, visual arts and Chinese literature standard level were removed from the model.

4.1.2 Results

The model was rerun with the remaining 41 DP subjects. Overall, 96% of students had both infit and outfit statistics <2.0. To maximize the strength of the model and achieve the most accurate results, students with infit and outfit statistics over 2.0 ($n = 3{,}566$) were removed. The Expected *a Posteriori* (EAP) reliability of the final model was high, at 0.91. All subjects showed good fit, with outfits and infits below 1.7, and grade thresholds with outfits less than 1.51. Item fit statistics are shown in Table 3.

Rasch-Thurstone thresholds were calculated to estimate the "difficulty" of individual grades in relation to the assumed construct of "general academic ability." Results for grades 2–7 are in Table 4, and shown graphically in a Wright Map (Figure 2). Subjects are ordered by "difficulty" according to the Rasch-Thurstone measures at grade 7. However, if the difficulty of a different grade were used, the order of the subjects would change (Coe, 2008). Students and subject grade thresholds are located along the same logit scale on the Wright Map, with higher logits indicating greater student "ability" and more "difficult" grades (Bond and Fox, 2007). The lowest grades achieved (usually grade 1, but occasionally 2 or 3) corresponded to logits below −10, indicating low ability estimates, and very small proportions of students at the lowest end of the scale. For example, in English literature SL, the grade distribution was: 0% at grade 1, 0.8% at grade 2, 5.8% at grade 3, 22.8% at grade 4, 38.1% at grade 5, 26.4% at grade 6, and 6.2% at grade 7.

Standard level consistently appears "harder" at grade 7 than higher level in literary subjects. This may be due to students selecting higher level in their strongest subjects, or others taking standard level in their second language for recognition of linguistic proficiency[7].

4.2 DP dataset (2): "contextualized and evidenced argumentation"

4.2.1 Analytical strategy

With the aim of refining the definition of the linking construct, the second stage of analysis took a *part-construct* approach to construct comparability, by focusing on assessments "designed to assess related constructs, rather than the same construct" (Newton, 2010, p. 48). Here, the intention was to make the common frame of reference between the subjects under investigation more theoretically relevant. As such, subjects that assess the same skills are compared, which may be more theoretically justifiable than comparing all disciplines under the broad construct definition of "general academic ability." Subjects were therefore narrowed to those with similar assessment objectives (AOs) to literature (see Appendix). Based on the mapping exercise of AOs, four additional subjects were selected to be included with literature subjects: history, global politics, theatre and film. These subjects assess contextual awareness and the selection of

[5] The six DP groups: Studies in Language & Literature, Language Acquisition, Individuals & Societies, Sciences, Mathematics, and The Arts.

[6] With higher level and standard level options classed as separate subjects.

[7] DP literature and language & literature are benchmarked against the Common European Framework of Reference (CEFR) (Ecctis, 2023) and often used as recognition of linguistic proficiency for university admissions.

TABLE 3 Item fit statistics for DP-wide Rasch analysis, ordered by outfit.

Subject	Level	Students	Outfit	Infit
Chemistry	Higher	14,646	0.78	0.78
Biology	Higher	25,149	0.80	0.80
Economics	Higher	14,809	0.82	0.81
Chemistry	Standard	10,975	0.82	0.82
Physics	Higher	12,030	0.82	0.83
Biology	Standard	15,311	0.86	0.86
Geography	Higher	4,980	0.86	0.86
Business management	Higher	12,209	0.88	0.87
Computer science	Higher	1,828	0.89	0.89
Environmental systems & societies	Standard	9,063	0.91	0.92
History	Higher	32,411	0.92	0.92
Sports exercise & health science	Standard	1,129	0.92	0.92
Global politics	Higher	1,946	0.92	0.92
Physics	Standard	8,847	0.93	0.93
Psychology	Higher	9,287	0.94	0.94
Business management	Standard	3,644	0.95	0.95
Economics	Standard	5,685	0.96	0.96
Information technology in a global society	Higher	1,239	0.98	0.97
Mathematics	Higher	14,013	0.99	0.98
Computer science	Standard	1,506	0.99	0.99
Geography	Standard	2,363	0.99	0.99
Design technology	Higher	1,074	1.00	1.00
Philosophy	Higher	1,548	1.03	1.03
History	Standard	5,350	1.06	1.06
Mathematics	Standard	37,251	1.08	1.08
Psychology	Standard	4,098	1.10	1.10
English language & literature	Higher	19,006	1.10	1.10
English literature	Higher	27,322	1.10	1.10
Spanish literature	Higher	5,466	1.10	1.10
Film	Higher	1,896	1.12	1.11
Maths studies	Standard	23,663	1.12	1.12
English language & literature	Standard	12,274	1.15	1.14
English literature	Standard	4,987	1.19	1.19
Chinese literature	Higher	457	1.22	1.20
Chinese language & literature	Higher	559	1.22	1.21
Philosophy	Standard	1,192	1.25	1.24
Theatre	Higher	2,052	1.25	1.24
Spanish language & literature	Higher	2,266	1.28	1.27
Chinese language & literature	Standard	1,266	1.42	1.42
Spanish literature	Standard	488	1.46	1.45
Spanish language & literature	Standard	852	1.49	1.46

relevant evidence in building an argument. Argumentation is a fundamental skill in literature, to the extent that "argumentation" may be considered a form of knowledge in the discipline (Elliott, 2021). As such, the linking construct for this analysis was assumed to be "contextualized and evidenced argumentation."

4.2.2 Results

Overall, 18 subjects were included in the second iteration of the Rasch model, based on the results of 78,567 students. These students represented schools from 141 different countries, with the most common again being the United States ($n = 29{,}244$), Canada ($n = 4{,}298$), the UK ($n = 3{,}592$), India ($n = 3{,}567$) and China ($n = 3{,}355$), which collectively made up 56% of the student population in dataset 2. Approximately 56% of the student population were reported as female ($n = 44{,}174$), 44% as male ($n = 34{,}369$), and 24 did not specify. Over 98% of students had infit and outfit statistics <2.0, and those over 2.0 ($n = 1{,}425$) were removed. All subjects fitted the model well, with outfits and infits below 1.7 (Table 5).

The EAP reliability was lower with this dataset at 0.63. Whilst higher reliability would be required if comparability metrics were to be operationalized in a high-stakes assessment context, it was sufficient for the exploratory nature of this investigation.

Rasch-Thurstone thresholds were calculated to estimate the "difficulty" of the subjects based on the theoretical construct of "contextualized and evidenced argumentation." Results for grades 2–7 are presented in Table 6, and shown in a Wright Map in Figure 3.

Strikingly, Chinese subjects appeared "harder" at grade 7 than Spanish in dataset 1, but "easier" in dataset 2. Subjects removed from dataset 2 were primarily maths or science-related, suggesting that students taking Chinese subjects performed better in maths and sciences than the other subjects retained in dataset 2. These early findings demonstrated challenges in using "general academic ability" as the linking construct.

4.3 DP dataset (3): "literary analysis"

4.3.1 Analytical strategy

The final stage followed a *full-construct* comparability approach, which "can apply only to situations in which alternative forms of an examination are developed to assess the same construct" (Newton, 2010, p. 48). This analysis applied only to English, Spanish and

TABLE 4 Dataset 1 subject threshold (grade) difficulties, ordered by grade 7.

Subject*	Level	Grade**					
		2	3	4	5	6	7
English language & literature	Standard	−Inf	−8.30	−5.21	−1.57	1.35	5.10
Information technology in a global society	Higher	−6.91	−4.70	−2.65	−0.38	1.93	5.08
History	Standard	−Inf	−4.96	−3.11	−0.24	2.54	4.94
English literature	Standard	−Inf	−7.09	−3.74	−0.70	2.04	4.89
Chinese language & literature	Standard	−Inf	−Inf	−4.87	−2.76	0.52	4.43
Computer science	Standard	−5.47	−2.06	−0.40	1.13	2.37	4.34
Chemistry	Higher	−4.81	−2.02	−0.22	0.99	2.38	4.33
Computer science	Higher	−4.70	−2.23	−0.70	0.78	2.51	4.27
History	Higher	−8.23	−5.17	−3.12	−0.44	1.93	4.21
English literature	Higher	−9.05	−6.26	−4.06	−1.31	1.47	4.20
Philosophy	Higher	−7.13	−4.74	−2.78	−0.90	1.02	4.13
Chinese literature	Higher	−Inf	−Inf	−Inf	−3.11	0.83	4.11
Chemistry	Standard	−4.99	−2.11	−0.47	0.80	1.99	4.04
Design technology	Higher	−6.91	−4.42	−2.21	−0.07	1.72	4.00
Mathematics	Higher	−4.56	−1.98	−0.48	0.94	2.39	3.86
Biology	Higher	−6.03	−3.37	−1.53	0.21	1.77	3.86
Psychology	Standard	−7.27	−3.99	−2.64	−0.87	1.14	3.85
Mathematics	Standard	−4.99	−2.34	−0.76	0.57	1.95	3.80
Economics	Standard	−6.21	−3.96	−1.89	−0.33	1.49	3.72
Geography	Standard	−7.75	−5.20	−3.06	−1.08	1.37	3.61
Chinese language & literature	Higher	−Inf	−Inf	−Inf	−3.63	−0.39	3.60
Philosophy	Standard	−Inf	−6.03	−3.00	−1.05	0.97	3.60
Biology	Standard	−7.07	−3.92	−1.87	−0.07	1.44	3.59
Global politics	Higher	−Inf	−6.69	−4.19	−1.65	0.94	3.59
Philosophy	Higher	−7.41	−5.65	−3.16	−0.93	1.05	3.57
English language & literature	Higher	−8.64	−7.19	−4.72	−1.97	0.53	3.48
Economics	Higher	−6.00	−3.97	−2.26	−0.63	1.31	3.48
Physics	Higher	−5.41	−2.68	−0.53	0.77	2.06	3.43
Spanish language & literature	Standard	−Inf	−5.58	−3.60	−1.29	0.78	3.40
Physics	Standard	−5.51	−2.63	−0.48	0.91	2.12	3.31
Film	Higher	−7.35	−4.06	−2.52	−0.62	1.27	3.24
Environmental systems & societies	Standard	−6.23	−4.23	−1.86	−0.22	1.56	3.20
Spanish literature	Standard	−Inf	−4.54	−2.82	−0.86	0.99	3.19
Geography	Higher	−7.59	−5.65	−3.02	−1.26	1.09	3.06
Business management	Higher	−8.65	−5.24	−3.37	−1.25	0.51	3.01
Spanish literature	Higher	−9.90	−6.35	−3.92	−1.69	0.48	3.00
Sports exercise & health science	Standard	−7.57	−4.41	−1.94	−0.53	1.26	2.98
Spanish language & literature	Higher	−Inf	−6.60	−4.60	−2.33	0.51	2.96
Maths studies	Standard	−6.01	−3.83	−2.44	−0.98	0.84	2.76
Business management	Standard	−7.41	−5.36	−3.90	−1.82	0.48	2.75
Theatre	Higher	−6.77	−4.82	−3.04	−1.24	0.36	2.64

*English subjects highlighted in blue, Chinese in orange & Spanish in green.

**Inf: indicates an effectively infinite threshold due to insufficient data.

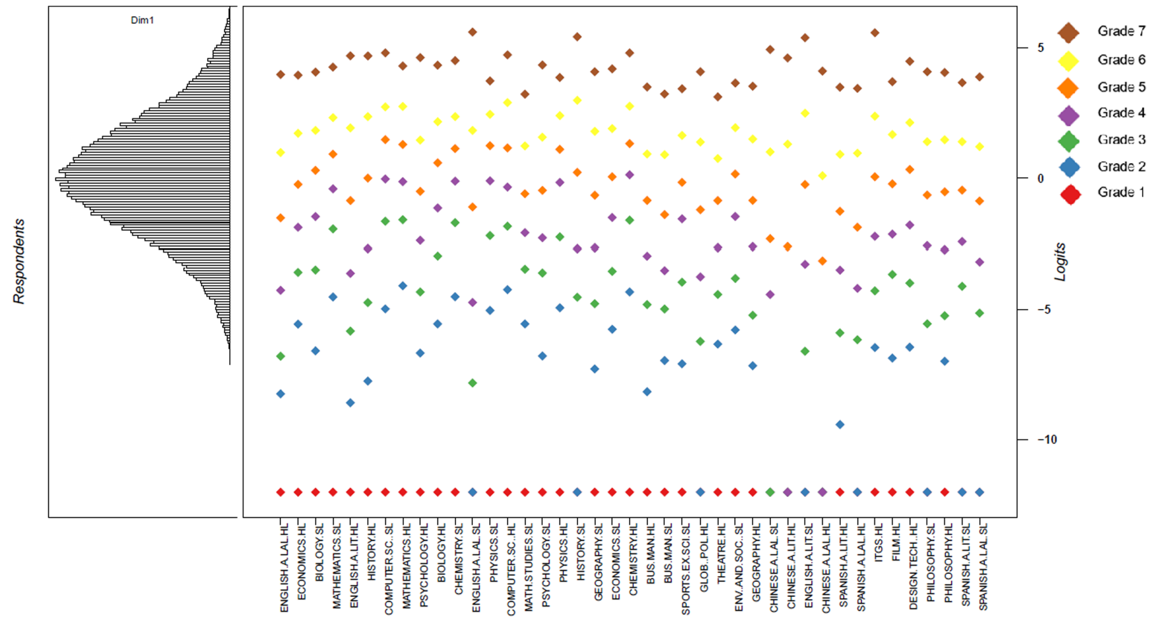


FIGURE 2
Wright Map for dataset 1.

Chinese versions of DP Studies in language and literature subjects. DP students can take certain subjects in two different languages to be eligible for a "Bilingual Diploma" (IBO, 2024), which is typically taken by multilingual students with prior academic experience in at least two languages (Rivera et al., 2014). The performance of students across pairs of languages (e.g., English and Spanish literature) was used to facilitate the analysis with "literary analysis" assumed as the linking construct. DP literature and language & literature subjects shared common assessment objectives and grade descriptors, so could reasonably be assumed to share the linking construct of literary analysis. Bilingual Diploma subject combinations most typically involve a combination of English and Spanish subjects (e.g., 1,248 students selected both English language & literature and DP Spanish language & literature). The second most common language combination was English and Chinese (e.g., 547 students selected English language & literature and Chinese language & literature). However, Chinese and Spanish literature subjects are rarely taken together.

4.3.2 Results

Only data for students (n = 3,377) taking two English, Spanish or Chinese subjects were included in the model, which represented schools from 67 countries. The main countries represented differed from datasets 1 and 2, with 60% of the dataset 3 population coming collectively from Mexico (n = 497), Colombia (n = 482), the United States (n = 404), Hong Kong (n = 373), and China (n = 282). Approximately 55% of the student population were reported as female (n = 1,874), 44% as male (n = 1,501), and 2 did not specify. The proportion of misfitting students was slightly higher than in the previous datasets, with 5% of students having infit and outfit statistics >2.0. Again, those over 2.0 (n = 177) were removed. All 12 subjects had outfits and infits below 1.7 (Table 7).

EAP reliability was initially 0.64, but increased to 0.72 when misfits were removed from the model. Results for grades 2–7[8] are presented in Table 8, and illustrated in a Wright Map in Figure 4.

4.4 Comparative overview of results across datasets

As noted previously, the EAP reliability estimates varied across the three sets of analyses: 0.91 (dataset 1), 0.63 (dataset 2), and 0.72 (dataset 3). The lower reliability in dataset 2 ("contextualized and evidenced argumentation") compared to dataset 1 ("general academic ability") may be partly explained by less information available per student in the model, given the reduced number of subjects. The decrease in reliability may also indicate greater homogeneity among examinees (i.e., students in dataset 2 having more similar ability levels or subject matter knowledge), as the subjects were aligned by assessment objectives so were more closely related in terms of content and skills. Conversely, however, the increase in reliability from dataset 2 to dataset 3 may reflect the fact that the remaining language versions of assessments have a very tight construct focus and are very aligned in what they measure.

Different rank orders for English, Spanish and Chinese subjects also emerged from the Rasch analyses (Figure 5). Some changes were small, such as English language & literature SL which remained constant across all three datasets, repeatedly appearing as one of the "hardest" subjects. However, as previously noted, it is striking that some Chinese and Spanish subjects switch places in the first two datasets. Other notable changes include English literature higher level

[8] Grade 1 is excluded, due to the sparsity of students at the lowest grade.

TABLE 5 Item fit statistics for AO-aligned subjects, ordered by outfit.

Subject	Level	Students	Outfit	Infit
History	Higher	33,401	0.94	0.94
English literature	Higher	28,141	0.96	0.96
Philosophy	Higher	1,605	0.97	0.97
English language & literature	Higher	19,488	0.97	0.97
Global politics	Higher	1,974	0.98	0.97
History	Standard	5,505	0.99	0.99
Spanish literature	Higher	5,657	0.99	0.99
Chinese literature	Higher	480	1.01	1.00
English language & literature	Standard	12,484	1.01	1.01
Philosophy	Standard	1,231	1.01	1.01
English literature	Standard	5,115	1.01	1.01
Film	Higher	1,949	1.03	1.03
Chinese language & literature	Higher	591	1.06	1.06
Spanish language & literature	Standard	869	1.09	1.08
Spanish language & literature	Higher	2,342	1.10	1.10
Theatre	Higher	2,122	1.12	1.12
Spanish literature	Standard	500	1.13	1.12
Chinese language & literature	Standard	1,317	1.22	1.20

appearing “hardest” in dataset 2, but third “easiest” in dataset 3. The fluctuations across the datasets suggest that the model revealed comparability issues in the data.

However, it should be noted that an inspection of the 95% confidence intervals (CIs) for item difficulty parameters revealed several overlapping CIs across subjects (Table 9). Each subject had a 95% confidence interval associated with its item difficulty parameter, with the brackets indicating subjects where parameters had overlapping confidence intervals and were not significantly different from each other. For example, Chinese literature HL in dataset 3 had a wide CI [3.95, 5.15] and overlapped with all other subjects. English literature HL and English language & literature SL in dataset 2, however, are significantly more difficult than English literature SL.

TABLE 6 Dataset 2 subject threshold (grade) difficulties, ordered by grade 7.

Subject*	Level	Grade**					
		2	3	4	5	6	7
History	Standard	−9.08	−4.45	−2.81	−0.19	2.44	4.73
History	Higher	−7.68	−4.56	−2.56	0.06	2.39	4.59
English literature	Higher	−8.66	−5.68	−3.57	−0.96	1.76	4.40
English language & literature	Standard	−Inf	−8.94	−6.00	−2.42	0.51	4.28
English literature	Standard	−Inf	−8.00	−4.74	−1.61	1.12	4.00
Philosophy	Higher	−6.22	−5.06	−2.74	−0.58	1.39	3.83
Film	Higher	−7.09	−3.58	−2.10	−0.20	1.73	3.76
Spanish language & literature	Standard	−Inf	−5.25	−3.47	−1.21	0.95	3.75
Global politics	Higher	−Inf	−6.26	−3.85	−1.39	1.10	3.73
English language & literature	Higher	−8.31	−6.72	−4.29	−1.66	0.80	3.72
Spanish literature	Higher	−9.59	−5.77	−3.30	−1.04	1.16	3.64
Spanish language & literature	Higher	−Inf	−5.77	−3.96	−1.87	0.95	3.49
Philosophy	Standard	−Inf	−5.56	−2.67	−0.84	1.04	3.48
Chinese literature	Higher	−Inf	−Inf	−Inf	−3.56	0.12	3.32
Spanish literature	Standard	−Inf	−4.57	−2.80	−0.87	0.99	3.26
Chinese language & literature	Standard	−Inf	−Inf	−5.42	−3.82	−0.87	3.14
Theatre	Higher	−6.07	−4.29	−2.52	−0.74	0.82	3.04
Chinese language & literature	Higher	−Inf	−Inf	−Inf	−4.07	−0.98	2.92

*English subjects highlighted in blue, Chinese in orange & Spanish in green.
**Inf: indicates an effectively infinite threshold due to insufficient data.

Looking across datasets, there are changes in the subject difficulty rankings that are unlikely to be due to random fluctuation. For example, looking at the relative differences between English literature HL and Spanish language & literature SL, English appears “harder” than Spanish in datasets 1–2, but “easier” in dataset 3. The different results across datasets illustrate potential implications of *post hoc* modeling in examining construct comparability across different language versions of assessments, where different linking constructs may produce contradictory results.

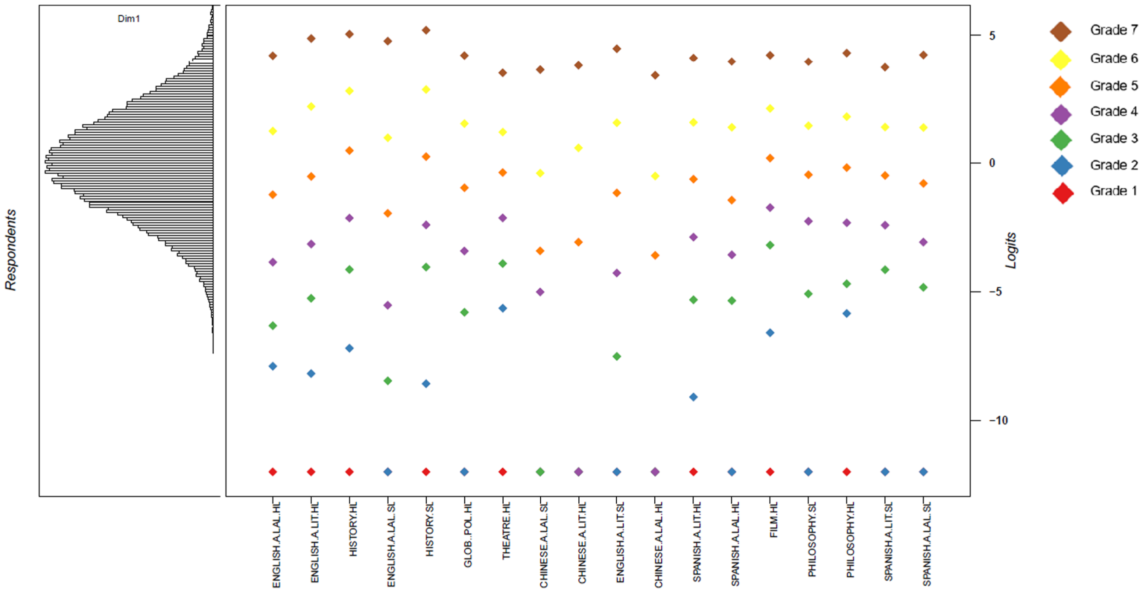


FIGURE 3
Wright Map for dataset 2.

TABLE 7 Item fit statistics for dataset 3, ordered by outfit.

Subject	Level	# Students	Outfit	Infit
Spanish literature	Higher	718	0.94	0.94
Chinese literature	Higher	58	0.95	0.94
English literature	Standard	207	0.95	0.94
Spanish language & literature	Standard	582	0.96	0.96
English language & literature	Standard	811	0.98	0.98
English language & literature	Higher	1,636	0.99	0.99
English literature	Higher	547	0.99	0.99
Spanish language & literature	Higher	812	1.00	1.00
Chinese literature	Standard	172	1.01	1.00
Spanish literature	Standard	240	1.02	1.02
Chinese language & literature	Standard	480	1.03	1.02
Chinese language & literature	Higher	140	1.08	1.06

4.5 Unidimensionality

Finally, unidimensionality was investigated for each dataset using PCAR[9] and examining scree plots. First, Cattell's (1966) scree test was employed, whereby only the components above the clearest inflexion point in the scree plot are retained. Figure 6 shows the scree plots for each dataset.

An examination of the scree plots suggested that the unidimensionality assumption was violated in each dataset. For dataset 1—where subjects across all DP subject groups were included—the curve dropped drastically between components 1 and 2 (with the eigenvalue dropping by 5.25), suggesting an argument for unidimensionality could be potentially made. The PCAR showed that the first component explained the largest proportion of the total variance, with additional components explaining progressively less variance suggesting that the data may predominantly be explained by a single latent trait. However, the first component only explained 20.80% of the total variance, and a small curve downwards is also evident at the 8th component on the scree plot. The PCAR also showed that eight additional components have eigenvalues over 1.40, and an eigenvalue threshold of 1.40 is a common cut-off, below which is likely to be random noise (Raîche, 2005). The cumulative proportion of variance explained by components 1–9 was also 65.03%. This evidence combined suggests

[9] Missing data posed some challenges, as no students would have taken all subjects included in the model. As data was missing by design rather than at random (i.e., due to optionality in the DP where students can choose different combinations of subjects), this was dealt with via pairwise deletion.

TABLE 8 Dataset 3 subject threshold (grade) difficulties, ordered by grade 7.

Subject*	Level	Grade**					
		2	3	4	5	6	7
Chinese literature	Standard	−Inf	−Inf	−6.13	−2.22	0.96	5.43
Spanish language & literature	Standard	−Inf	−5.32	−3.50	−0.91	1.49	4.99
Spanish language & literature	Higher	−Inf	−5.27	−3.54	−1.51	1.28	4.59
English language & literature	Higher	−Inf	−6.83	−4.19	−1.01	1.55	4.47
English language & literature	Standard	−Inf	−Inf	−6.71	−1.97	0.53	4.35
Spanish literature	Higher	−Inf	−Inf	−4.49	−1.83	0.79	3.97
Chinese language & literature	Higher	−Inf	−Inf	−Inf	−Inf	−0.69	3.96
Spanish literature	Standard	−Inf	−5.88	−2.82	−0.56	1.24	3.82
Chinese language & literature	Standard	−Inf	−Inf	−4.92	−3.50	−0.34	3.80
English literature	Higher	−Inf	−6.55	−4.61	−1.89	0.89	3.64
English literature	Standard	−Inf	−Inf	−Inf	−2.69	0.23	3.41
Chinese literature	Higher	−Inf	−Inf	−Inf	−3.01	0.17	3.14

*English subjects highlighted in blue, Chinese in orange & Spanish in green.
**Inf: indicates an effectively infinite threshold due to insufficient data.

multidimensionality, thus violating the Rasch assumption of unidimensionality.

For dataset 2, the Cattell scree test suggested that at least two components be retained, with the elbow appearing between components 2 and 3—although it could arguably be as many as 8, with a drop also seen between components 8 and 9. The PCAR also indicated that the unidimensionality assumption was violated, with components 1–2 only accounting for 37.13% of the total variance, whereas the cumulative proportion of variance explained for the first 8 components was 87.62%. This combined evidence strongly suggested multidimensionality. The evidence of multidimensionality was stronger in dataset 2 than in dataset 1, which is somewhat surprising given that subjects here were selected to assess common skills. However, this is likely due to there being less information about each student included in the model, due to fewer connections and overlap across subjects (data from 6 subjects for each student was included in dataset 1, and only 2–3 subjects per student in dataset 2).

Finally, for dataset 3 a case could theoretically be made for unidimensionality, with the sharpest drop seen between components 1 and 2, and 29.69% of the total variance explained by the first component. However, the curve noticeably flattened out from component 5 (where the cumulative percentage of variance explained was 70.19%), suggesting that the unidimensionality assumption was again violated. The unidimensionality assumption—a requirement of the Rasch model—is seemingly violated in each dataset. This suggests that the DP data is not ideally suited to Rasch analyses, which is perhaps an inevitable challenge in using Rasch for inter-subject comparability purposes (Ofqual, 2022), and is acknowledged as a limitation of the study.

Whilst a distinguishing feature of Rasch models is the stipulation for the attribute of interest to be unidimensional, this is a challenging requirement to be applied strictly in analysis of operational data. In reality, even a single, short-item response will inevitably reflect more than one underlying trait—for example, involving reading the question, thinking about the answer, and writing the response (Bond and Fox, 2007). Thus, the violation of the unidimensionality assumption here is not unexpected due to the complex nature of the constructs under investigation. Indeed, extended responses—as are typical in literature assessments—will violate the unidimensionality assumption, as essays are inherently multifaceted, encompassing aspects such as the quality of ideas, organization, and development of the argument (Sato, 2022). Moreover, in curriculum-based assessment, the underlying construct may be more reasonably be considered a "composite variable" that reflects the curriculum of interest, rather than a single latent trait (Baird and Opposs, 2018, p. 13). Therefore, lower correlations between different constructs (e.g., literary and linguistic competence) are not considered an issue as long as the assessments represent the specified curriculum (ibid). A limitation of using Rasch to examine comparability across subjects in curriculum-based assessment is the violation of unidimensionality, as it is not realistic to assume the different exams will all only measure a single ability in these contexts (Ofqual, 2022). This limitation should therefore be acknowledged when Rach modeling is used operationally, particularly when results of these analyses are used to inform policymaking or decisions related to assessment implementation.

5 Discussion

This study used Rasch modeling to generate estimates of the relative "difficulty" of different IB subjects—focusing specifically on English, Spanish and Chinese versions of Studies in language and literature subjects—in three subsets of IB assessment data. The dataset was increasingly narrowed, with the aim of gradually sharpening the definition of the linking construct. First, 41 subjects were modeled based on the linking construct of "general academic ability," then 18 subjects in relation to "contextualized and evidenced argumentation," and finally 12 subjects based on the construct of "literary analysis." In doing so, it demonstrated that statistical techniques can be used—to an extent—to evaluate comparability in parallel-developed multilingual assessments. As such, it extends previous research on comparability in parallel multi-language assessments, which has

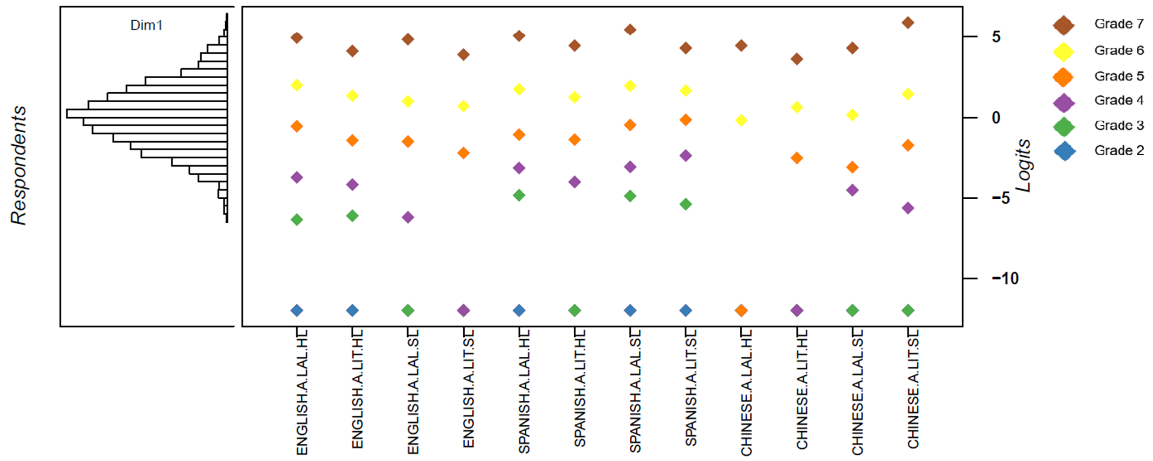


FIGURE 4
Wright Map for dataset 3.

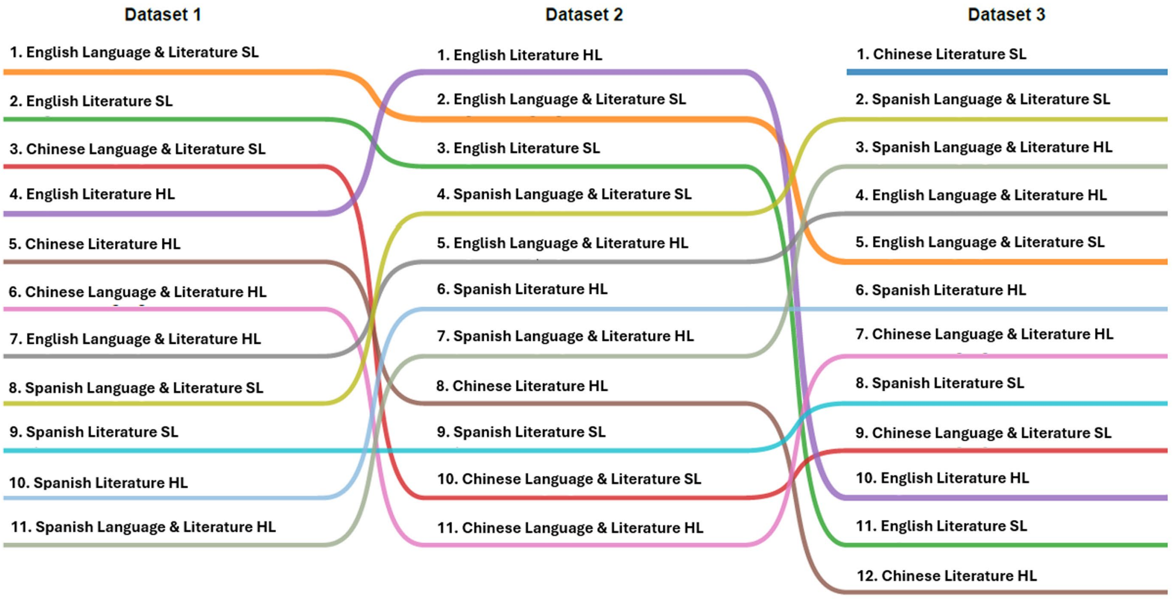

Dataset 1
1. English Language & Literature SL
2. English Literature SL
3. Chinese Language & Literature SL
4. English Literature HL
5. Chinese Literature HL
6. Chinese Language & Literature HL
7. English Language & Literature HL
8. Spanish Language & Literature SL
9. Spanish Literature SL
10. Spanish Literature HL
11. Spanish Language & Literature HL
Dataset 2
1. English Literature HL
2. English Language & Literature SL
3. English Literature SL
4. Spanish Language & Literature SL
5. English Language & Literature HL
6. Spanish Literature HL
7. Spanish Language & Literature HL
8. Chinese Literature HL
9. Spanish Literature SL
10. Chinese Language & Literature SL
11. Chinese Language & Literature HL
Dataset 3
1. Chinese Literature SL
2. Spanish Language & Literature SL
3. Spanish Language & Literature HL
4. English Language & Literature HL
5. English Language & Literature SL
6. Spanish Literature HL
7. Chinese Language & Literature HL
8. Spanish Literature SL
9. Chinese Language & Literature SL
10. English Literature HL
11. English Literature SL
12. Chinese Literature HL

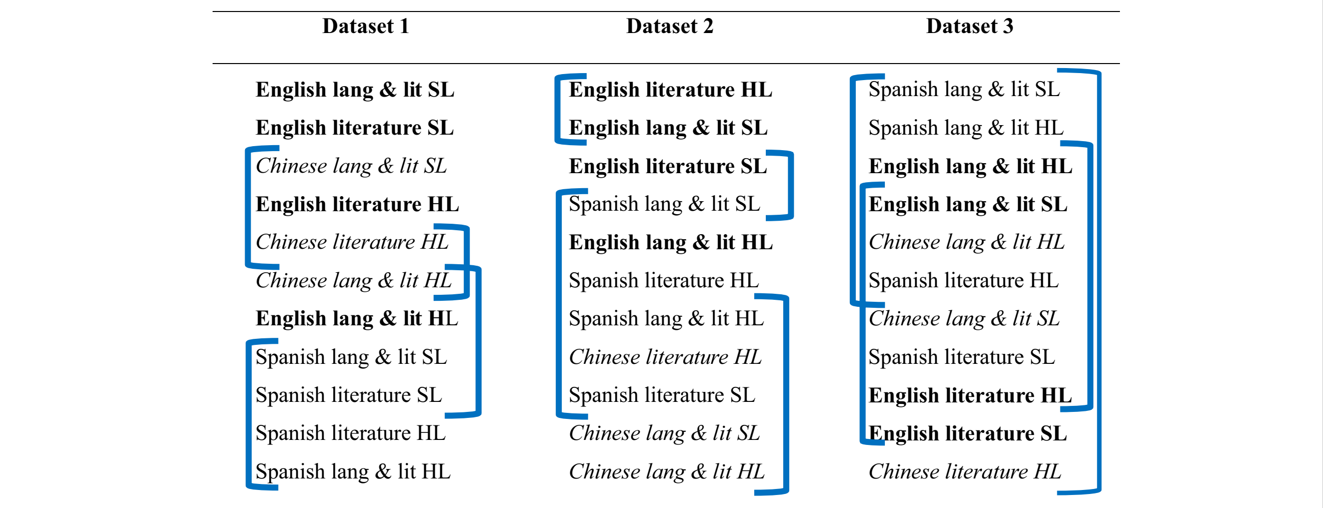
FIGURE 5
Overview of changes in rank orders across datasets.

largely been limited to qualitative procedures that align standards across different language versions using expert judgment (Badham et al., 2025; Davis et al., 2008).

The study also highlights the importance of having a meaningful rationale for data selection, and the need for a clearly defined, theoretically driven linking construct. Yet, whilst each linking construct in the present study had a defensible rationale and theoretical justification based on the academic literature, they each had limitations. Dataset 1 used the generalized linking construct of "general academic ability" to examine comparability of student performance across diverse academic subjects. In the absence of shared curriculum and assessment content, this has been a common approach in inter-subject comparability studies (e.g., Coe, 2008; He et al., 2018; Ofqual, 2022; Veas et al., 2020). This has the advantage of being able to model larger datasets which, as seen in dataset 1, tends to result in higher reliability. However, the approach is limited as it cannot account for the specific knowledge and skills as defined in each individual subject (He et al., 2018). Challenges with generalized linking constructs such as "general academic ability" may also account for certain subject types, including language acquisition and creative subjects misfitting in the first iteration of the model. Therefore, it is questionable whether it is valid to use the same measure to compare such different fields, for example, as maths and literature.

The second dataset aimed to mitigate the limitations of the generalized approach by taking a part-construct approach to comparability. It focused on subjects that targeted similar skills in their assessment objectives, defined here as "contextualized and evidenced argumentation." As such, the linking construct increased

TABLE 9 Item difficulty parameter rank orders across datasets*.



Dataset 1	Dataset 2	Dataset 3
English lang & lit SL	**English literature HL**	Spanish lang & lit SL
English literature SL	**English lang & lit SL**	Spanish lang & lit HL
Chinese lang & lit SL	**English literature SL**	**English lang & lit HL**
English literature HL	Spanish lang & lit SL	**English lang & lit SL**
Chinese literature HL	**English lang & lit HL**	*Chinese lang & lit HL*
Chinese lang & lit HL	Spanish literature HL	Spanish literature HL
English lang & lit HL	Spanish lang & lit HL	*Chinese lang & lit SL*
Spanish lang & lit SL	*Chinese literature HL*	Spanish literature SL
Spanish literature SL	Spanish literature SL	**English literature HL**
Spanish literature HL	*Chinese lang & lit SL*	**English literature SL**
Spanish lang & lit HL	*Chinese lang & lit HL*	*Chinese literature HL*

*Brackets indicate subjects where item difficulty parameters have overlapping confidence intervals.

in relevance by accounting for disciplinary-specific knowledge and skills. A limitation in this second approach was seen in the reduced reliability (dropping from 0.91 in dataset 1 to 0.63 in dataset 2). This may partly be explained by the reduced overlap across subjects in dataset 2, which contains less information about student performance (with 6 subjects per student in dataset 1, versus 2–3 subjects in dataset 2). It may also indicate increased homogeneity in dataset 2, where students of similar abilities across cognate subjects demonstrate similar knowledge and skills—making it harder for the model to distinguish between different ability levels. The differences in the rank orders of subject "difficulty" between the first two datasets illustrate the challenges in making such comparisons based on different linking constructs.

The final analysis aimed to explore a full-construct approach to comparability, by comparing different versions of assessments that aimed to assess the same target construct (Newton, 2010). This comprised different language versions of subjects that all assess literary analysis. The final stage of the analysis was based on Bilingual Diploma students who had taken different language versions of the same subjects, which produced yet another rank order. This final analysis allowed the most precise, and disciplinary-specific definition of the linking construct. However, whilst the linking construct of "literary analysis" seems most theoretically appropriate for representing the construct of interest and comparing languages within the same discipline, it presents other challenges. On a practical level, by relying on Bilingual Diploma students, the size of the dataset was limited. More importantly, other factors such as motivation may impact these results, as only students with a keen interest in literature would likely take two versions of the subject. Similarly, bilingual students naturally develop discourse strategies that may be advantageous in literature studies (García, 2020), so may perform differently from their peers. As such, Bilingual Diploma students are not likely to be representative of all DP students. Moreover, bilingual students represented different cultural contexts, with datasets 1 and 2 primarily constituting Anglophone countries (US, Canada, UK) and dataset 3 being predominantly countries where English is not the dominant language (Mexico, Colombia, Hong Kong). Therefore, there are also likely to be cultural variations in how constructs are manifested across the subsets of data.

This leads to the question of which dataset and linking construct can be justified. Stone and Stenner (2014) argued that a doubly prescriptive model is required, in which not only the measurement requirements of the Rasch model are necessary, but also a theoretically justified framing of the data. Without such a frame of reference, items or people can be deleted from the data to better ensure model fit, but this could alter the construct and conclusions from the analysis. The analyses presented here are an application of their argument. Subjects or people could be stripped out of the analyses to better fit the data to the Rasch model, but the question of what should be included or removed is separate and can only be addressed by theoretical considerations.

Despite limitations, a full-construct approach is most theoretically defensible for evaluating comparability across different language versions of the same subject. Adopting a more precise definition of the linking construct—that reflects the curriculum on which all language versions are based—has the potential to improve fairness and validity in multilingual assessment contexts. Broader conceptualizations of the linking construct (such as "general academic ability" in dataset 1) risk introducing construct-irrelevant bias in cross-lingual assessment comparisons. This may include language proficiency being a confounding factor, when students take assessments in a language other than their first language (as many DP students do for non-literary subjects), which can impact their performance (Elosua, 2016). Similarly, sociolinguistic differences may have an impact when different subject areas are compared. For example, Leung and Revina suggested an alignment between Chinese language and mathematics, as linguistic features such as "visual–spatial properties of the Chinese characters" can enhance "visual-motor integration skills [that] may contribute to more efficient learning of mathematics" (Leung and Revina, 2023, p. 1472). Thus, comparisons between maths and Chinese language may yield different results than similar comparisons between maths and Spanish, for example. Adopting a full-construct approach to examining comparability has the potential to enhance IB processes for DP multilingual assessments developed in parallel, by providing more valid comparability evidence that targets the specific construct of interest.

However, differences in "difficulty" estimates for the same subjects across the three datasets also illustrate the challenges of

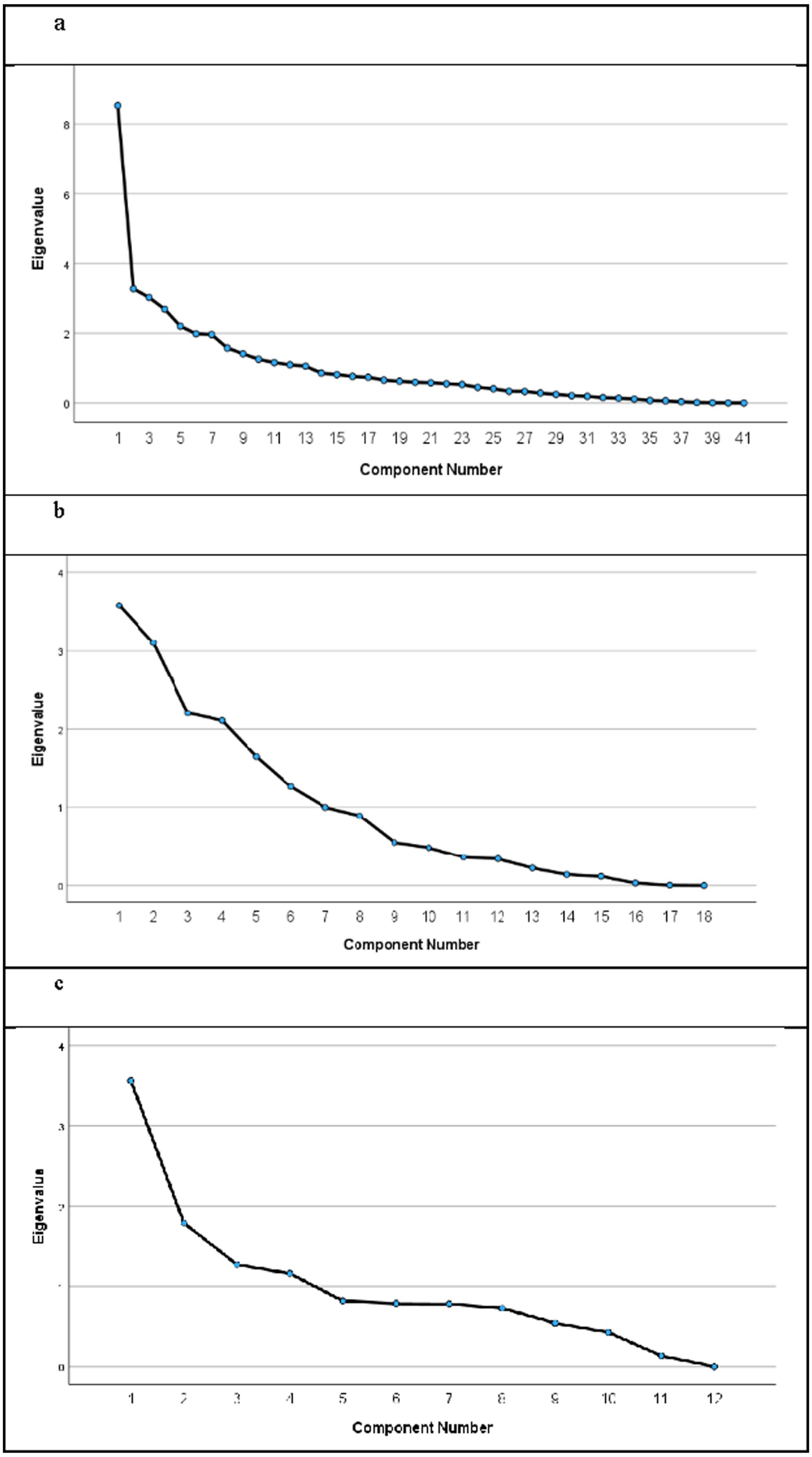


FIGURE 6
Scree plots across the three datasets. **(a)** Dataset 1: "general academic ability". **(b)** Dataset 2: "evidenced and contextualized argumentation". **(c)** Dataset 3: "literary analysis".

using such approaches to investigate comparability *post hoc*. This aligns with our initial hypothesis that inherent linguistic and cultural differences in complex, multifaceted constructs limit the extent to which *post hoc* statistical analyses can be used to evaluate comparability in multi-language assessments that have been developed in parallel. These results are important, as *post hoc* analyses such as Rach modeling are used to generate statistical estimates of relative "difficulty" of different subjects (or different language versions of subjects) and inform assessment policy and practices (e.g., Ofqual, 2018). Yet, these analyses typically take place too late in the assessment lifecycle, as decisions have already been made on the basis of scores (Solano-Flores, 2019). In high-stakes summative assessments like the DP, these can have potentially far-reaching consequences, including impacting students' admissions into higher education or future career prospects.

Nevertheless, such analyses can be informative and have the potential to support assessment processes—when combined with other evidence. For instance, comparability outcomes may be considered alongside expert judgment on student work (i.e., examiner judgment on levels of attainment in relation to grade descriptors), qualitative evidence on the comparability of target constructs, and statistical evidence including relative student performance in each subject from one year to the next. These different sources of evidence can be integrated in grade awarding processes to determine the most appropriate grade boundaries for a subject. For example, if a particular language version of a subject consistently appears as "harder" than its counterparts, grade boundaries may be lowered to account for increased difficulty, if supported by other evidence (e.g., relating to the relative performance of first language and non-first language users and whether this impacts the perceived difficulty of a subject).

6 Conclusion

Different conclusions could be drawn based on the three sets of analyses presented in this study: each based on a different assumed definition of the linking construct. Such differences are not surprising, given that each analysis was based on a different subset of data. Similar differences in comparability outcomes would likely arise with the use of different statistical techniques (although the nature and extent of these differences may vary depending on the specific method employed). Moreover, there are limitations applying the Rasch model in this context, such as a violation of the unidimensionality assumption. Nevertheless, the differences in rank orders across the three datasets do underscore the challenges and risks of taking action based on *post hoc* comparability analyses. As the *QuantCrit* movement emerging from Critical Race Theory reminds us, "numbers are not neutral," so data and methods should be interrogated as they can, if left unchecked, contribute to social justice challenges (Gillborn et al., 2018, p. 158). In the present study, results from the DP-wide analysis in the first dataset could lead to an interpretation that Spanish is leniently graded compared to Chinese, whereas the assessment objectives-based analysis in dataset 2 might suggest the reverse. Therefore, taking actions such as adjusting marking or grading standards based on one of these interpretations could potentially introduce bias against a linguistic group, if not validated by additional evidence.

Such evidence might relate to how constructs manifest in different linguistic and cultural contexts, since constructs are not necessarily universal across languages and cultures (Hambleton, 2002). In multilingual and multicultural assessments, construct bias may be introduced if different constructs are measured across different linguistic and cultural groups (van de Vijver and Poortinga, 2004). Such challenges are magnified in multilingual assessments, where different academic approaches may be associated with different languages, such as a tendency to prioritize contextual evidence in Spanish literature compared to a preference for formal and technical features in English literature (Badham and Furlong, 2023; Galache-Ramos, 2017). These different approaches are exemplified in published samples of DP literature student work. An English literature sample essay on Dickenson's poem "A Bird Came Down the Walk," for instance, focuses primarily on literary devices such as "personification of the bird," "syntax," and "juxtaposition of ideas" (IBO, 2019b). In contrast, a Spanish literature sample on Sábato's "The Tunnel," emphasizes elements such as historical context ("to understand the psychology and personality of this character, it is first essential to immerse oneself in the context in which the novel was written"[10]) and the impact of the author's philosophical beliefs ["Sábato was an existentialist, which greatly influenced the creation of many of his works"[11] (IBO, 2019c)].

Fundamental differences in the underlying construct across language versions raises further questions about construct comparability investigations. Previous cross-lingual assessment comparability studies have highlighted the limitations of DIF analyses in identifying systematic issues of bias when assessments have been adapted from one language into another (e.g., El Masri et al., 2016). Yet despite their limitations, such comparability investigations are helpful in identifying potential issues for further investigation (e.g., if an item is performing differently in one language, it may indicate a translation issue). Similarly, findings from this study suggest Rasch analyses have considerable limitations in comparing parallel language versions of an assessment. However, they may help identify potential issues for further investigation—such as linguistic or cultural differences in the manifestation of the target construct—as long as there is sufficient data, and a theoretically justified linking construct.

Previous Rasch inter-subject comparability studies have emphasized the need to balance analyses against other evidence before adjusting standards to accommodate different levels of "difficulty" (Ofqual, 2022). The limitations of *post hoc* analyses conducted in this study were clear and also underline the need for additional evidence, for example relating to cohorts' linguistic profiles, and the definition and manifestation of target constructs across cultural and linguistic groups. In addition to a meaningful rationale for data selection and a clearly defined, theoretically driven linking construct, this study also emphasizes the importance of balancing comparability metrics with other empirical evidence about performance across different language groups. This includes validity evidence relating to the comparability (or similarity) of target constructs across different linguistic and cultural

[10] *'Para entender la psicología y la personalidad de este personaje es esencial en primer lugar sumergirse en el contexto en que se escribió esta novela'.*

[11] *'Sábato era un existencialista alta, y esto influyó altamente en la creación de muchas de sus obras'.*

groups (American Educational Research Association [AERA], 2014). Additionally, it may include comparing assessment performance between different subpopulations, such as linguistic minority and linguistic majority groups in a particular assessment (Elosua, 2016; Faulkner-Bond and Sireci, 2015). Such evidence is essential to avoid systemic bias being introduced against linguistic or cultural subgroups.

7 Limitations and recommendations for further research

As outlined previously, dataset 3 offered the most theoretically justifiable approach to linking multilingual assessments but was limited by an unrepresentative student population (e.g., students being motivated to study more than one literature subject, and multilingual skills potentially enhancing literary competencies). To address representation issues with Bilingual Diploma students, one useful research avenue may be to apply a matched monolingual groups design. For example, given the diversity of student cohorts in IB schools, different groups of examinees could be matched by individual school (e.g., English and Spanish literature students from the same school). This design could help control for regional effects as well as school type (e.g., private vs. state school). As IB schools frequently offer literature subjects in multiple languages, this would likely also have the benefit of providing a larger dataset for analysis.

Another constraint was that the PCAR analysis suggested a violation of the unidimensionality assumption, which may be an inherent limitation of using Rasch modeling for inter-subject comparability purposes (Ofqual, 2022), indicating this type of data is not best suited to the model. Rank order differences across datasets illustrate challenges of *post hoc* comparisons in the absence of other evidence relating to subject and language "difficulty." The unidimensionality violation may make these results and interpretations misleading in an operational context. Yet, in the psychometrics literature, there is a debate about whether the model or the data should be prioritized. Andrich (2004) argued that the Rasch model should be used for measurement purposes. The current analyses seek to explore differences in constructs and standards, hence the choice of model. Other models could be explored, such as a multidimensional Rasch model (e.g., Briggs and Wilson, 2003), which may address the unidimensionality issue. Whilst the focus here was to explore the use of a single measurement of subject and language difficulty based on subject grades, further research using multidimensional models may be helpful in exploring specific dimensions that differ in multi-language assessments (e.g., essay structure versus analysis of literary devices). However, the purpose here was not to pursue ideal model fit, but to use the model pragmatically as a statistical tool to demonstrate that when the linking construct changes, so too do the comparability outcomes.

In a multilingual context, the complex nature of the disciplinary construct creates challenges, as literary and linguistic competences are inherently intertwined in the assessment of literature (Moosavinia and Razavipour, 2017). This is problematical because it conflicts with the Rasch unidimensionality requirement, but also because the linguistic competency element implies construct differences across languages. This highlights the needs to balance statistical evaluations of comparability such as Rasch alongside other forms of evidence relating to the nature of the target constructs, and the perceived "difficulty" across subjects and languages. Rasch models are used to explore inter-subject comparability in curriculum-based assessments, impacting both theoretical and operational discussions, including policy decision-making (e.g., Coe, 2008; He et al., 2018; Ofqual, 2018). Thus, these investigations are important to inform assessment practitioners of the potential limitations and challenges of Rasch analyses in these contexts.

This study has illustrated that whilst statistical estimates of comparability for parallel-developed multilingual assessments can be produced, there are challenges in interpreting the findings. *Post hoc* statistical analyses such as the Rasch model have the potential to identify if assessments in one language appear "harder" than another. However, different approaches can generate different—often statistically significant—conclusions about the relative difficulties of subjects. Thus, questions remain about which subjects are appropriate to compare, how, and under what conditions. Most importantly, however, such analyses do not diagnose the reasons for apparent differences in "difficulty," nor can they detect systematic bias. Further investigation is needed into underlying factors that impact assessment results across languages and cultures. These include the influence of cultural norms and linguistic features on performance, such as the prioritization of different skills across cultures, or how character-based versus Roman alphabet languages may impact learning in certain subjects. Additional research is also needed into the impact of cohorts' linguistic characteristics, such as bilingual versus monolingual student performance. Finally, it is recommended that further research be conducted to investigate processes of defining and articulating constructs at the very start of the assessment design process. This may include more detailed investigations into samples of student work in dataset 3. For example, bilingual subject matter experts could carry out comparative qualitative analyses of samples to examine with the construct of interest manifests in different ways across languages. Gathering comparability evidence at the start of the assessment lifecycle may also help avoid construct bias and reduce the risk of systematically disadvantaging linguistic and cultural subgroups in multilingual assessments.

Data availability statement

The datasets presented in this article are not readily available because the data comprises students' assessment results from an International Baccalaureate (IB) exam session. Anonymized data may be available upon request to the IB. Requests to access the datasets should be directed to assessment.research@ibo.org.

Author contributions

LB: Visualization, Project administration, Formal analysis, Methodology, Data curation, Writing – review & editing, Investigation, Writing – original draft, Resources, Conceptualization. MM: Supervision, Methodology, Writing – review & editing. J-AB: Supervision, Writing – review & editing, Methodology.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. In addition to providing time and access to data/resources, the International Baccalaureate (IB) funded the open-access fee for this publication.

Acknowledgments

This research was carried out under the auspices of a Doctor of Philosophy (DPhil) at the University of Oxford in the UK, and was supported by the International Baccalaureate. The authors would like to thank Professor David Andrich for his invaluable comments on an earlier version of this manuscript, as well as Dr. Christopher Wheadon for his helpful insights and guidance. Initial results were shared as a conference presentation, but are published here for the first time.

Conflict of interest

LB is an employee of the International Baccalaureate (IB), whose assessments are the subject of this research. The research was conducted as part of the author's doctoral studies. The IB supported the research through access to data/resources and allocating work time for research activities.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2025.1616879/full#supplementary-material

References

American Educational Research Association [AERA] (2014). Standards for educational and psychological testing. Washington DC.

Andrich, D. (2004). Controversy and the Rasch model: a characteristic of incompatible paradigms? *Med. Care* 42, I7–I16. doi: 10.1097/01.mlr.0000103528.48582.7c

Andrich, D., and Marais, I. (2019). "Violations of the assumption of Independence I-multidimensionality and response dependence" in A course in Rasch measurement theory: measuring in the educational, social and health sciences. eds. D. Andrich and I. Marais (Singapore: Springer Nature Singapore), 173–185.

Badham, L. (2025). Statistically guided grading judgements: contextualisation or contamination? *Oxf. Rev. Educ.* 51, 17–35. doi: 10.1080/03054985.2023.2290640

Badham, L., and Furlong, A. (2023). Summative assessments in a multilingual context: what comparative judgment reveals about comparability across different languages in literature. *Int. J. Test.* 23, 111–134. doi: 10.1080/15305058.2022.2149536

Badham, L., Oliveri, M. E., and Sireci, S. G. (2025). Navigating multi-language assessments: best practices for test development, linking, and evaluation. *Psicothema* 37, 1–11. doi: 10.70478/psicothema.2025.37.19

Baghaei, P. (2008). Local dependency and Rasch measures. *Rasch Meas. Trans.* 21, 1105–1106.

Baird, J.-A., and Opposs, D. (2018). "The standard setting project: assessment paradigms" in Examination standards. how measures and meanings differ around the world (London: UCL Institute of Education Press), 2–25.

Bond, T. G., and Fox, C. M. (2007). Applying the Rasch model: fundamental measurement in the human sciences. *2nd* Edn. New York: Lawrence Erlbaum Associates Publishers.

Boone, W. J., Staver, J. R., and Yale, M. S. (2014). Rasch analysis in the human sciences. Dordrecht: Springer Netherlands.

Briggs, D. C., and Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *J. Appl. Meas.* 4, 97–100.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivar. Behav. Res.* 1, 245–276. doi: 10.1207/s15327906mbr0102_10

Christensen, K. B., Makransky, G., and Horton, M. (2017). Critical values for Yen's Q_3: identification of local dependence in the Rasch model using residual correlations. *Appl. Psychol. Meas.* 41, 178–194. doi: 10.1177/0146621616677520

Coe, R. (2007). "Common examinee methods" in Techniques for monitoring the comparability of examination standards. eds. P. Newton, J. A. Baird, H. Goldstein, H. Patrick and P. Tymms (London: QCA), 331–371.

Coe, R. (2008). Comparability of GCSE examinations in different subjects: an application of the Rasch model. *Oxf. Rev. Educ.* 34, 609–636. doi: 10.1080/03054980801970312

Coe, R., Searle, J., Barmby, P., Jones, K., and Higgins, S. (2008). Relative difficulty of examinations in different subjects. Report for SCORE (Science Community Supporting Education). Durham: CEM Centre, Durham University.

Davis, S. L., Buckendahl, C. W., and Plake, P. S. (2008). When adaptation is not an option: An application of multilingual standard setting. *J. Educ. Meas.* 45, 287–304. doi: 10.1111/j.1745-3984.2008.00065.x

Ecctis. (2023). The international baccalaureate diploma Programme: referencing selected IB DP English, French, German, and Spanish subjects to the common European framework of reference for languages (CEFR). Ecctis. Available online at: https://ibo.org/globalassets/new-structure/university-admission/pdfs/ecctis_ib-dp-cefr_final-report_august_2023.pdf (Accessed April 15, 2025).

El Masri, Y. H., Baird, J.-A., and Graesser, A. (2016). Language effects in international testing: the case of PISA 2006 science items. *Assess. Educ. Princ. Policy Pract.* 23, 427–455. doi: 10.1080/0969594X.2016.1218323

Elliott, V. (2021). Knowledge in English: canon, curriculum and cultural literacy. London: Routledge.

Elosua, P. (2016). Minority language revitalization and educational assessment: do language-related factors impact performance? *J. Socioling.* 20, 212–228. doi: 10.1111/josl.12176

Elosua, P., and López-jaúregui, A. (2007). Potential sources of differential item functioning in the adaptation of tests. *Int. J. Test.* 7, 39–52. doi: 10.1080/15305050709336857

Ercikan, K., and Lyons-Thomas, J. (2013). "Adapting tests for use in other languages and cultures" in APA handbook of testing and assessment in psychology, Vol. 3: Testing and assessment in school psychology and education (Washington, DC: American Psychological Association), 545–569.

Ercikan, K., and Oliveri, M. E. (2016). In search of validity evidence in support of the interpretation and use of assessments of complex constructs: discussion of research on assessing 21st century skills. *Appl. Meas. Educ.* 29, 310–318. doi: 10.1080/08957347.2016.1209210

Ercikan, K., and Por, H.-H. (2020). "Comparability in multilingual and multicultural assessment contexts" in Comparability of large-scale educational assessments: issues and recommendations. ed. National Academy of Education (Washington, DC: National Academy of Education), 205–225.

Faulkner-Bond, M., and Sireci, S. G. (2015). Validity issues in assessing linguistic minorities. *Int. J. Test.* 15, 114–135. doi: 10.1080/15305058.2014.974763

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 diploma programme [Master's thesis]. Bath: University of Bath.

García, O. (2020). Diploma programme: supporting all students' first or best language. [internal report]: IBO.

Gillborn, D., Warmington, P., and Demack, S. (2018). QuantCrit: education, policy, 'big data' and principles for a critical race theory of statistics. *Race Ethn. Educ.* 21, 158–179. doi: 10.1080/13613324.2017.1377417

Grisay, A., Gonzalez, E., and Monseur, C. (2009). Equivalence of item difficulties across national versions of the PIRLS and PISA reading assessments. *IERI* 2, 63–83.

Hambleton, R. K. (2002). "Adapting achievement tests into multiple languages for international assessments" in Methodological advances in cross-national surveys of educational achievement. eds. A. Gamoran and A. C. Porter (Washington, DC: National Academies Press), 58–79.

Hambleton, R. K. (2005). "Issues, designs, and technical guidelines for adapting tests into multiple languages" in Adapting educational and psychological tests for cross-cultural assessment (New York: Psychology Press), 3–38.

He, Q., and Black, B. (2019). Statistical evidence pertaining to the claim of grading severity in GCSE French, German and Spanish and the impact of statistical alignment of standards on outcomes. Ofqual. Available online at: https://core.ac.uk/download/pdf/237701048.pdf

He, Q., and Black, B. (2020). Impact of calculated grades, centre assessment grades and final grades on inter-subject comparability in GCSEs and a levels in 2020. Coventry: Ofqual. (Accessed April 15, 2025).

He, Q., Stockford, I., and Meadows, M. (2018). Inter-subject comparability of examination standards in GCSE and GCE in England. *Oxf. Rev. Educ.* 44, 494–513. doi: 10.1080/03054985.2018.1430562

Hernández, A., Hidalgo, M., Hambleton, R., and Gómez-Benito, J. (2020). International test commission guidelines for test adaptation: a criterion checklist. *Psicothema* 3, 390–398. doi: 10.7334/psicothema2019.306

Hlosta, M., Herzing, J. M. E., Seiler, S., Nath, S., Zai, F. K., Bergamin, P., et al. (2024). "Analysis of process data to advance computer-based assessments in multilingual contexts". In Assessment analytics in education. Advances in analytics for learning and teaching. eds. M. Sahin and D. Ifenthaler (Springer).

IBO (2019a). Assessment principles and practices – quality assessments in a digital age. Cardiff: IBO.

IBO (2019b). Language A assessed student work. Cardiff: first assessment 2021.

IBO (2019c). Lengua A: Ejemplos de trabajos evaluados de alumnos. Cardiff: Primera evaluación en 2021.

IBO (2021). DP language courses: overview and placement guidance. Cardiff: IBO.

IBO (2024). Diploma programme assessment procedures 2024. Cardiff: IBO.

Karabatsos, G. (2000). A critique of Rasch residual fit statistics. *J. Appl. Meas.* 1, 152–176

Keng, L., and Marion, S. (2020). "Comparability of aggregated group scores on the "same test"" in Comparability of large-scale educational assessments: Issues and recommendations. ed. National Academy of Education (Washington, DC: National Academy of Education), 49–74.

Lamprianou, I. (2009). Comparability of examination standards between subjects: an international perspective. *Oxf. Rev. Educ.* 35, 205–226. doi: 10.1080/03054980802649360

Lerner, M. (2021). How can you create culturally fair content for tests and surveys? (test translation and localisation): cApStAn. Cultural Suitability and Sensitivity Review. Available online at: https://www.capstan.be/cultural-suitability-and-sensitivity-review/

Leung, F. K. S., and Revina, S. (2023). "The influence of culture on students' mathematics achievement in East Asia" in International handbook on education development in the Asia-Pacific. eds. W. O. Lee, P. Brown, A. L. Goodwin and A. Green (Singapore: Springer Nature), 1463–1479.

Linacre, J. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Meas. Trans.* 16:878.

Masters, G., and Wright, B. (1996). "The partial credit model" in Handbook of modern item response theory. eds. W. van der Linder and R. K. Hambleton (New York: Springer).

Moosavinia, S. R., and Razavipour, K. (2017). Instructors' perceptions of the construct-relevance of language in the assessment of literature. Learning and assessment: making the connections. ALTE 6th international conference, Bologna, Italy.

Newton, P. E. (2005). Examination standards and the limits of linking. *Assess. Educ. Principles Policy Pract.* 12, 105–123. doi: 10.1080/09695940500143795

Newton, P. E. (2010). Thinking about linking. *Meas. Interdiscip. Res. Perspect.* 8, 38–56. doi: 10.1080/15366361003749068

Newton, P. (2011). A level pass rates and the enduring myth of norm-referencing. *Res. Matters*, 20–26. doi: 10.17863/CAM.100441

Newton, P. E. (2012). Making sense of decades of debate on inter-subject comparability in England. *Assess. Educ. Principles Policy Pract.* 19, 251–273. doi: 10.1080/0969594X.2011.563357

Ofqual (2018). Policy decision: inter-subject comparability in A level sciences and modern foreign languages. Available online at: https://assets.publishing.service.gov.uk/media/5bf433ff40f0b60783ad9374/ISC_Decision_Document_20.11.18.pdf (Accessed April 15, 2025).

Ofqual (2022). An investigation of inter-subject comparability in GCSEs and A levels in summer 2021. Available online at: https://www.gov.uk/government/publications/an-investigation-of-inter-subject-comparability-in-gcses-and-a-levels-in-summer-2021/an-investigation-of-inter-subject-comparability-in-gcses-and-a-levels-in-summer-2021 (Accessed April 15, 2025).

Oliveri, M. E. (2019). Considerations for designing accessible educational scenario-based assessments for multiple populations: a focus on linguistic complexity. *Front. Educ.* 4:88. doi: 10.3389/feduc.2019.00088

Oliveri, M. E., Ercikan, K., and Simon, M. (2015). A framework for developing comparable multilingual assessments for minority populations: why context matters. *Int. J. Test.* 15, 94–113. doi: 10.1080/15305058.2014.986271

Oliveri, M. E., Ercikan, K., and Zumbo, B. D. (2013). Analysis of sources of latent class differential item functioning in international assessments. *Int. J. Test.* 3, 272–293. doi: 10.1080/15305058.2012.738266

Oliveri, M. E., Lawless, R., and Mislevy, R. J. (2019). Using evidence-centered design to support the development of culturally and linguistically sensitive collaborative problem-solving assessments. *Int. J. Test.* 19, 270–300. doi: 10.1080/15305058.2018.1543308

Raîche, G. (2005). Critical eigenvalue sizes (variances) in standardized residual principal components analysis (PCA). *Rasch Meas. Trans.* 19:1012.

Randall, J. (2021). "Color-neutral" is not a thing: redefining construct definition and representation through a justice-oriented critical antiracist Lens. *Educ. Meas. Issues Pract.* 40, 82–90. doi: 10.1111/emip.12429

Randall, J., Slomp, D., Poe, M., and Oliveri, M. E. (2022). Disrupting white supremacy in assessment: toward a justice-oriented, antiracist validity framework. *Educ. Assess.* 27, 170–178. doi: 10.1080/10627197.2022.2042682

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Chicago: Institute of Education Research.

Rivera, C., Tressler, T. R., McCreadie, J., and Ballantyne, K. (2014). IB diploma programme study: factors influencing students to earn a bilingual diploma. The George Washington University Center for Equity and Excellence in Education. Available online at: https://ibo.org/globalassets/new-structure/programmes/dp/pdfs/bilingual-diploma-final-report.pdf (Accessed April 15, 2025).

Robitzsch, A., Kiefer, T., and Wu, M. (2022). TAM: test analysis modules. R package version.

Sato, T. (2022). Assessing critical thinking through L2 argumentative essays: an investigation of relevant and salient criteria from raters' perspectives. *Lang. Test. Asia* 12:9. doi: 10.1186/s40468-022-00159-4

Sireci, S. G. (1997). Problems and issues in linking assessments across languages. *Educ. Meas. Issues Pract.* 16, 12–19. doi: 10.1111/j.1745-3992.1997.tb00581.x

Sireci, S. G., Rios, J. A., and Powers, S. (2016). "Comparing scores from tests administered in different languages" in Fairness in educational assessment and measurement. eds. S. G. Sireci, J. A. Rios and S. Powers (New York: Routledge), 181–202.

Solano-Flores, G. (2019). Examining cultural responsiveness in large-scale assessment: the matrix of evidence for validity argumentation. *Front. Educ.* 4:43. doi: 10.3389/feduc.2019.00043

Solano-Flores, G., and Li, M. (2009). Language variation and score variation in the testing of English language learners, native Spanish speakers. *Educ. Assess.* 14, 180–194. doi: 10.1080/10627190903422880

Stone, M., and Stenner, J. (2014). Frames of reference. *Rasch Meas. Trans. Rasch Meas. SIG Am. Educ. Res. Assoc.* 28, 1479–1482.

van de Vijver, F. J. R., and Poortinga, Y. (2004). "Conceptual and methodological issues in adapting tests" in Adapting educational and psychological tests for cross-cultural assessment (New Jersey: Psychology Press).

Veas, A., Navas, L., Pozo-Rico, T., and Miñano, P. (2020). University entrance examinations in Spain: using the construct comparability approach to analyze standards quality. *Front. Psychol.* 11:127. doi: 10.3389/fpsyg.2020.00127

Wright, B. D., and Linacre, J. (1994). Reasonable mean-square fit values. *Rasch Meas. Trans.* 8:370.

Zhao, X., and Solano-Flores, G. (2023). Test translation review: a study on discussion processes and translation error detection in consensus-based review panels. *Front. Educ.* 8:1303617. doi: 10.3389/feduc.2023.1303617

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Lang. Test.* 20, 136–147. doi: 10.1191/0265532203lt248oa