# Evaluating the Impact of Passing Conditions and Retakes on Precision in the IB's Diploma Programme

Adam Maghout

Student number: 7692617

*Course: Methodology and Statistics for the Behavioural, Biomedical and Social Sciences*

*Candidate journal: Assessment in Education: Principles, Policy & Practice*

*Supervisor: Anton Béguin*

*(International Baccalaureate)*

Word count: 2485

FETC approval: 24-2025

Due date: 12/05/2025

Submission date: 12/05/2025

# Evaluating the Impact of Passing Criteria and Retakes on Precision in the IB Diploma Programme

Adam Maghout[a]
[a]*Methodology and Statistics, Utrecht University, Utrecht, The Netherlands*

High-stakes examinations have an obvious impact on student wellbeing and academic outcomes. Yet, little research has been led into optimising complex passing criteria in assessments, defined as passing criteria that combine several conjunctive and compensative rules. Using exam scores from May and November 2023, this study applied simulation-based methods within a Classical Test Theory (CTT) framework to obtain the passing rate and precision of the current International Baccalaureate Diploma Programme (IBDP) passing criteria. Five alternative sets of passing criteria were also tested and compared using precision, specificity, sensitivity, positive predictive value (PPV) and negative predictive value (NPV). Furthermore, a retake was simulated under each decision rule, assuming a random true score change between test and retake. Removing Higher- and Standard-Level minima left the passing rate unchanged, whereas lowering the 24-point total or the cap on grade 3s markedly increased passes. All rules showed high precision, but simpler criteria scored highest at the expense of specificity; retakes further boosted sensitivity. Demographic analyses revealed that rule performance hinged on the proportion of true failing candidates. Moderate simplification of the IBDP passing criteria therefore appears defensible, although larger changes risk diluting the standards of the diploma. Future work should develop analytic methods for tuning composite passing rules.

**Keywords:** measurement precision; international baccalaureate; retakes; high-stakes decision; classical test theory

# 1 Introduction

High-stakes examinations exert considerable academic, financial and emotional pressure on candidates as their outcomes determine entry to further study, employment and other life-course opportunities (Marchant, 2004; Cho and Chan, 2020). In response, transnational programmes such as the *Programme for International Student Assessment* (PISA) have encouraged systems to align curricula and teaching standards, while the discipline of educational measurement supplies the technical safeguards required to keep tests fair across jurisdictions (Waldow, 2009). A central safeguard is the rejection of single-score pass/fail decision rules, as decisions based on a single observed exam mark are vulnerable to misclassification whenever transient factors, such as exam unreliability or sickness, depress performance. Standard 12.10 of the *Standards for Educational and Psychological Testing* codifies this principle by

recommending a *composite* approach in which the scores from several exams are combined to reach a decision (American Educational Research Association et al., 2014). The procedure used to reach this decision is referred to as a *decision rule*.

Reflecting this guidance, the International Baccalaureate *Diploma Programme* (DP) awards the diploma when a candidate satisfies an overall threshold of 24 points, accrued across six subjects and core components, and simultaneously meets five subsidiary conditions that limit the impact of a single assessment ("DP Passing Criteria", 2023). Despite its global reach, the statistical *precision* of the DP's decision rule has never been derived analytically (Hill and Saxton, 2014). Accordingly, the International Baccalaureate Organisation, which oversees the programme, has recently shown interest in doing so as part of a quality review exploring the simplification of DP passing criteria ahead of future exam sessions.

Beyond the DP context, systematic evidence on composite decision rules remains scarce. To the best of the author's knowledge, only two quantitative studies, one in Dutch secondary education and another in a Dutch university programme, have used simulation-based methods to gauge how decision rules influence diploma precision (van Rijn et al., 2012; Yocarini et al., 2018). No comparable analysis to date has been applied to any international secondary qualification. Moreover, although retake opportunities are designed to mitigate false negatives by giving proficient students a second chance to demonstrate competence (Coggeshall, 2021), their interaction with composite rules has never been modelled, leaving an important gap in both research and policy guidance.

The current study aimed to evaluate the impact of five proposed amendments of the DP passing criteria on passing rate and precision. Building on the work of van Rijn et al. (2012), simulation-based methods, conducted using reliability within a Classical Test Theory (CTT) framework, were extended to include a retake. This allowed the comparison of decision rule precision before and after retake for each proposed amendment. Additionally, figures related to the number of false positives and negatives were computed to determine their correlation with the simplification of passing criteria and the existence of a retake. Although calibrated to DP data, the approach was conceived as a transferable template for other high-stakes examinations that rely on composite decision rules. As the same methods were also applied to the current decision rule, this study aimed to provide the IBO with clear guidelines concerning the simplification of the DP passing criteria. It is hoped that these guidelines can be used as an example of what can be achieved by using simulation-based methods in educational measurement. Furthermore, the methods presented for retakes will enable future studies outside the Netherlands, offering a realistic yet straightforward way to evaluate precision before and after retake. Finally, this study aimed to identify key differences in how the amendments affected precision in different regions and for different genders. In addition to informing targeted policies, this allowed for the comparison of decision rules in groups with varying passing rates.

To achieve this, this article first introduces terminology related to decision rules and the current format of the DP, together with a formal definition of precision and other measures of

misclassification. A description of the data and proposed amendments to the decision rule is then provided, as well as the simulation design used to generate true and observed scores for test and retake. As results, decision rule precision under each amendment before and after retake are reported, along with sensitivity, specificity, positive predictive value and negative predictive value. Finally, the implications of the research are presented, as well as limitations and avenues for further research. Clear guidelines are given concerning the simplification of DP passing criteria.

# 2 Background

## 2.1 Decision rules

Composite decision rules require meeting multiple criteria for a favourable outcome. For example, requiring minimum grades in both mathematics and chemistry constitutes a composite decision rule, as the decision is reliant on two separate outcomes. These rules typically outperform simpler, single-criterion rules when the underlying tests are correlated, since they reflect overlapping constructs (Vermeulen-Kerstens et al., 2012). However, this advantage assumes that the measured abilities can compensate for each other. In practice, this assumption may be violated, particularly in cases where high ability in one subject cannot make up for low ability in another (Mehrens and Phillips, 1989). For instance, while strong reading comprehension in Spanish might offset weaker communication skills, no level of road sign recognition can realistically compensate for an inability to brake effectively during a driving test.

Decision rules are typically conjunctive or compensative. Conjunctive rules require sufficient scores on all tests. Conversely, compensative rules only require a sufficient score on average (van Rijn et al., 2012). This is illustrated in Figure 1. Lord (1962) notes that even if using a conjunctive rule would be ideal with perfectly reliable scores, this approach becomes flawed when applied to real-world test scores, which contain measurement error. Indeed, because observed scores are imperfect, decisions based on them will not match the optimal choices based on true scores. In practice, large-scale assessment rules often combine conjunctive and compensative aspects, which are referred to as complex decision rules (Douglas and Mislevy, 2010).

## 2.2 The Diploma Programme

DP students are required to select six subjects, three or four taken at Higher Level (HL) and the remainder at Standard Level (SL) (Tay, 2023). These must fall within each of the six IB subject groups, that are outlined in Table 1. Students may also decide to take an additional subject in groups 2, 3 and 4 instead of a subject in group 6. The exact options available vary depending on the delivering institution and the country (Maire, 2021). HL subjects explore material in greater depth, reflected in more contact hours and complex assessments. Subjects are graded from 1 to 7, with a 42-point maximum.

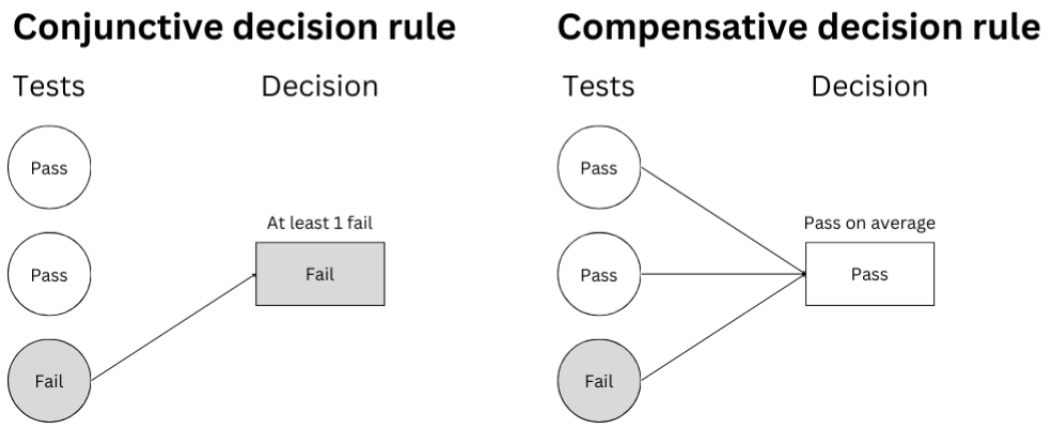


**Figure 1:** Conjunctive and compensative decision rules.

Completion of the DP core is also required, which includes three components: a project in Creativity, Activity, and Service (CAS), an Extended Essay (EE), and a Theory of Knowledge (TOK) component (Tay, 2023). Whilst CAS is assessed on a pass/fail basis, the EE and TOK are graded on a scale from A to E. Together, they award three additional points, bringing the maximum possible score for the DP to 45. Subject grades, including those for the core, are determined based on performance across multiple components. For instance, in biology HL, grades are derived from results on three examination papers and a practical assessment. These components are further subdivided into individual items, resulting in a three-level hierarchical structure for student grades (item score, component score, subject grade). An unattempted component yields an N. Students may also retake subjects to pass or to improve a previous grade, although, depending on the institution, they may not be required to retake all components within the subject ("Retaking Examinations", 2023). Currently, six requirements govern the DP ("DP Passing Criteria", 2023):

- No fewer than 24 points in total, of which at least 12 at HL and 9 at SL. If only two SL subjects are taken, this last requirement is lowered to 5.
- No N or E awarded on TOK or EE.
- No grade 1 awarded.
- No more than two grade 2s awarded.
- No more than three grade 3s or below awarded.
- CAS requirements are met.

Most requirements are compensatory; for example, the 24-point minimum lets students offset a poor grade with a strong one. However, some aspects are deliberately conjunctive, preventing prioritisation of certain subjects over others. Though simplified in 2014, further simplification of the passing criteria is under consideration to maintain standards of the diploma but reduce overlap between requirements.

**Table 1:** Subject Groups in the DP.

| | |
|---|---|
| **1** | Studies in Language and Literature |
| **2** | Language Acquisition |
| **3** | Individuals and Societies |
| **4** | Sciences |
| **5** | Mathematics |
| **6** | The Arts |

## 2.3 Measurement precision

In accordance with van Rijn et al. (2012), the notion of measurement precision used here stems from the Classical Test Theory (CTT) framework. For a single test, a participant's observed score $X$ is given by a true score $T$, and an error term $E$ related to person and test characteristics:

$$X = T + E \tag{1}$$

The true score $T$ represents the average score a participant would obtain over an infinite number of equivalent test administrations, assuming no learning or fatigue effects. The goal is to ensure $X$ closely approximates $T$, effectively minimising $E$. The extent to which this is achieved is quantified by the test's reliability, which is the main measure of precision in CTT. A higher reliability indicates that observed scores more consistently approximate true scores across individuals. Common reliability indicators include Cronbach's $\alpha$ (Cronbach, 1951) and McDonald's $\omega_t$ (McDonald, 1999), both measures of internal consistency. They are related to the systematic measurement error of a test $SE_{meas}$:

$$SE_{meas} = \sigma_X \sqrt{1 - R} \tag{2}$$

where $\sigma_X$ is the standard deviation of the test scores and $R$ is the reliability. Thus, if a decision rule is applied on the basis of a single test, the misclassification rate will depend on the reliability. Misclassification is defined here as students being assigned the wrong passing decision given their test scores, where the true passing decision is obtained by applying the decision rule to the unobserved true scores, which can be approximated using simulation-based methods but are never known in practice.

Although methods for creating reliability measures for composite scores exist (Gulliksen, 1950), these are ill-suited for the DP's complex passing criteria, notably due to the matrix-dependent conversion of core grades to awarded points. Thus, rather than using reliability only, the notion of measurement precision was extended to the complement of misclassification over all subjects. DP criteria were then precise if they yielded few false positives and negatives, as defined in Table 2a. An important distinction should be made between true

**Table 2:** Misclassification measures.

**(a)** Confusion matrix of observed (O) vs. true (T) outcomes.

| | | **T** | |
|---|---|---|---|
| | | **Pass** | **Fail** |
| **O** | **Pass** | *True Positive (TP)* | False Positive (FP) |
| | **Fail** | False Negative (FN) | *True Negative (TN)* |

**(b)** Additional misclassification measures.

$$\text{Precision} = \frac{TP + TN}{TP + FP + FN + TN}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Positive Predictive Value (PPV)} = \frac{TP}{TP + FP}$$

$$\text{Negative Predictive Value (NPV)} = \frac{TN}{TN + FN}$$

passes and true positives. A true pass refers to a case in which the underlying (true) passing status is a pass, whereas a true positive refers to a case where a true pass is also correctly observed as a pass. Both terms are not interchangeable, just as true fails and true negatives represent distinct concepts.

In addition to precision, the notions of sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV), as defined in Table 2b, were borrowed from van Rijn et al. (2012) to provide additional context to the precision measures per decision rule. In the present study, sensitivity represented the proportion of proficient students that were correctly given a pass. Likewise, specificity measured the proportion of non-proficient students that were correctly assigned a fail. These measures were used to examine changes in the rates of false positives and negatives as the passing criteria varied, as raw counts of these errors are affected by the number of truly proficient and non-proficient students and thus do not offer a standardised basis for comparison. Furthermore, PPV and NPV indicated the proportion of passing and failing students who were correctly classified, respectively.

# 3 Methods

## 3.1 Overview

This section first introduces the datasets that were used for the study, then presents the decision rules that were compared to the current DP decision rule. Afterwards, the methods for obtaining precision estimates from the observed data are detailed. To obtain the precision and passing rate of each rule from the data, simulations were set up in three steps, drawing inspiration from van Rijn et al. (2012):

**Initialisation**

1. Estimate the true component scores from the observed data.
2. Compute subject-level grades from the true component scores.
3. Apply the decision rule to obtain $T$, the true pass/fail outcome, and determine the initial passing rate based on the true grades.

**Simulation Before Retake**

For each simulation iteration ($n = 1, \ldots, 1\,000$):

4. Draw random error terms from a normal distribution for each component.
5. Add these errors to the true component scores to obtain simulated observed scores.
6. Compute subject-level grades from the observed component scores.
7. Apply the decision rule to the observed grades to obtain the pass/fail outcome $O_n$.
8. Compare $T$ and $O_n$ to construct a misclassification table and compute precision and other misclassification measures.

**Simulation After Retake**

If $O_n$ = Fail for a student:

9. Add a random increment from a normal distribution to the true component scores to simulate improvement in ability between the initial test and the retake. This yields the true component scores after retake.
10. Compute subject-level grades from the adjusted true component scores.
11. Draw a second set of random error terms from a normal distribution.
12. Add these errors to the adjusted true scores to simulate observed scores after retake.
13. If the observed score after retake is worse than before retake, replace it with the score before retake, to avoid score decrease. Otherwise, keep it as is.
14. Compute subject-level grades from the post-retake observed scores.
15. Apply the decision rule to both true and observed grades to obtain pass/fail outcomes $T_n^*$ and $O_n^*$ respectively.
16. Compare $T_n^*$ and $O_n^*$ to construct a post-retake misclassification table and derive precision estimates.

For students with $O_n$ = Pass, retake steps were not applied, as a retake was not simulated for students who already passed. Their scores and outcomes after retake were thus assumed to be identical to those prior to retake, that is $T = T_n^*$ and $O_n = O_n^*$. Thus, at each of the 1 000 iterations, students had four scores: their static pre-retake true score, their pre-retake observed score, their post-retake true score and their post-retake observed score. Consequently, they also had four passing decisions ($T$, $O_n$, $T_n^*$ and $O_n^*$) per iteration and per decision rule, which were all identical in cases where both the true and observed scores in May were already passes. A detailed account of the methods used is provided in the sections that follow.

## 3.2 Data

The analysis used three datasets containing component and item scores for examinees who sat DP components during the May and November 2023 exam sessions. May results established a baseline without retakes; November data covered students who retook subjects. Of the 179 772 students in the May data, only 73 690 took enough components to be assigned a passing decision. This disparity is because DP components can be taken over several sessions

**Table 3:** Gender and region distribution (May 2023). Regions: IB Africa, Europe and Middle East (IBAEM), IB Asia-Pacific (IBAP), IB Latin America (IBLA), and IB North America (IBNA).

| Gender \ Region | IBAEM | | IBAP | | IBLA | | IBNA | | **Total** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | N | % | N | % | N | % | N | % | N | % |
| Male | 14 170 | 19.2 | 9 081 | 12.3 | 3 089 | 4.2 | 6 878 | 9.3 | 33 218 | 45.1 |
| Female | 17 270 | 23.4 | 9 539 | 12.9 | 3 774 | 5.1 | 9 586 | 13.0 | 40 169 | 54.5 |
| Other | 97 | 0.1 | 14 | 0.0 | 14 | 0.0 | 178 | 0.2 | 303 | 0.4 |
| **Total** | 31 537 | 42.8 | 18 634 | 25.3 | 6 877 | 9.3 | 16 642 | 22.6 | 73 690 | 100.0 |

or as part of other courses internal and external to the IBO. Students were anonymised to the researcher via a candidate number that was linkable between the item and component datasets, as well as between test and retake.

Information was also provided concerning the gender and regional office of examinees, the distribution of which is provided in Table 3. Although not the main point of discussion of this research, identifying differences in misclassification rates between population strands was also carried out to enable targeted policies. High-stakes assessments have been shown to disproportionately disadvantage students from lower-income backgrounds (Smyth and Banks, 2012), with evidence suggesting a correlation between their increased implementation and rising dropout rates (Au, 2008). While such assessments are sometimes promoted as tools for improving equity, research indicates they often reflect disparities in school resources and funding, rather than genuine differences in student learning (Blazer, 2011). Thus the distribution of passing criteria benefits was expected to vary between IB regions.

The dataset for the November session contained results for 2 446 students, all of which were retaking a component they had previously taken in May. From these, 71.2% had not been awarded the diploma, with the remaining 28.8% doing a retake in order to improve their original grade. Only component scores were used for the retake analysis since reliabilities were copied from May.

In total, 1 052 components were assessed across 218 subjects and 73 different languages during the May exam session, of which 473 components across 70 subjects and only 10 languages were re-assessed in November. These were not exclusively language exams as students are able to take general components such as maths and biology in their mother tongue for additional credit. Some components, primarily in the core, were taken by most students whilst others were only taken by a single examinee, such as Welsh or Guarani. Component structures also varied, with items being either dichotomous or polytomous, and the total number of points differing. Items per component ranged from 2 to over 30.

## 3.3 Proposed Passing Criteria

Whilst maintaining the high standards of the DP is essential, internally-led research indicates that some schools are advising students against pursuing the DP due to its perceived high

**Table 4:** Studied sets of decision rules for the Diploma Programme, HL = Higher Level, SL = Standard Level.

| Set | Description |
|---|---|
| Rule 0 | Current decision rule |
| Rule 1 | Rule 0 + minimum for HL subjects set to 9 |
| Rule 2 | Rule 0 - minimum for HL and SL subjects |
| Rule 3 | Rule 2 + minimum total points set to 20 |
| Rule 4 | Rule 2 + minimum total points set to 18 |
| Rule 5 | Rule 4 - maximum number of grade 3s |

risk of failure. Additionally, overlap between passing criteria, such as the 12-point minimum across HL subjects and the overall 24-point minimum, imply the criteria can be simplified. To address these concerns, the amendments to the current passing criteria that the IB suggested to explore were to:

- Reduce the minimum total points required to pass from the current 24.
- Eliminate the minimum point requirements for HL and SL subjects.
- Lower the minimum points at HL to 9 to have the same minimum for all subjects.

In line with these guidelines, a reduction of the minimum total points to 18 was proposed in this study, corresponding to an average grade of 3 across the 6 subjects, where the core components are treated as bonus points. Alternatively, a 20-point minimum was also tested, retaining the same average grade but with 2 expected points from the core components. To evaluate the precision of the current passing rules and their potential revisions, six sets of passing criteria were applied to both simulated and observed data, given in Table 4. These sets accounted for removing HL/SL requirements alongside lowered totals. Additionally, removing the requirement concerning the maximum number of 3s was tested to evaluate the impact of heavily simplifying the rules. Relaxing other requirements was not evaluated as doing so would risk rendering the diploma trivial to obtain.

## 3.4 True score estimation

As it has been established that true component scores cannot be observed, a reasonable estimate $\hat{T}$ of them can be obtained using Kelley's formula (Kelley, 1923):

$$\hat{T}_i = \mu_X + \rho\,(X_i - \mu_X) \tag{3}$$

where $\mu_X$ is the mean component score across all students, $\rho$ is the reliability of the component and $X_i$ and $T_i$ are the observed and true scores for student $i$, respectively, as defined in Equation 1. This pushes unusually high and low component scores towards the mean, in situations where the component's reliability is low. Consequently, this avoided simulating observed scores around an extreme true value that was ultimately unreliable. For simplicity, the estimated true scores $\hat{T}_i$ are referred to as the true scores in the remainder of this work.

## 3.5 Passing rates

Passing rates under each decision rule were calculated by tallying the number of students meeting the criteria and expressing this as a proportion of the total sample. This served as a quality check to ensure that the proposed rules maintained acceptable pass rates beyond simply minimising misclassification.

Passing decisions were determined using two approaches: one based on the real observed scores and the other on the true scores. A misclassification rate was then calculated by comparing decisions from the two approaches to confirm that the use of Kelley's formula did not substantially alter the outcomes. Since passing decisions were assessed both before and after the simulated retake, and under both scoring approaches, each student had four associated passing decisions: two pre-retake (observed and true) and two post-retake (observed and true). Consequently, two misclassification rates were computed per rule, one before and one after retake, to evaluate consistency between the two methods.

## 3.6 Error generation

To apply measurement error to true scores, the reliability of components needed to be calculated as per Equation 2. Following recommendations by Hayes & Couttes (2020), McDonald's $\omega_t$ (McDonald, 1999) was used to calculate the reliability of components:

$$\omega_t = \frac{\left(\sum \lambda_j\right)^2}{\left[\left(\sum \lambda_j\right)^2 + \sum \left(1 - \lambda_j^2\right)\right]} \tag{4}$$

where $\lambda_j^2$ is the communality of item $j$, which quantifies how much individual variance the item has and ranges from 0 to 1. There are several ways to obtain an estimate $\lambda_j$ analytically, one of which is to use hierarchical factor analysis as performed by the omega function in the psych package (Revelle, 2024) used for the present study. Unfortunately, some components contained optional items, meaning students were given the choice to take some items instead of others. This is an issue for calculating reliability as $\omega_t$ relies on pairwise item comparisons, that cannot be performed if two items are mutually exclusive.

Fortunately, the mirt package (Chalmers, 2012) provides an alternative method to calculate reliability in the form of marginal reliability. For this, an Item Response Theory (IRT) model first needs to be fitted, after which a reliability for each item can be computed conditional on the other items. A marginal reliability is then obtained by integrating the newly obtained reliabilities over the latent ability scale. This can be done using the marginal_rxx function in R, which assumes a normal distribution for the abilities and does not rely on pairwise comparisons. A Generalised Partial Credit Model (GPCM) was fitted to the items before computing the reliability as it is suitable for polytomous data (Muraki, 1992) and provides more accurate estimates than the Graded Response Model (Samejima, 1969) in the presence of a large amount of missingness (Dai et al., 2021). A pilot study employing synthetic data with known reliability values was conducted to assess the effectiveness of using

marginal reliability in tests featuring optional items. This approach consistently produced estimates closest to the true reliability in comparison with imputation methods, even for tests that shared common items between versions and could therefore be linked. Consequently, marginal reliability was adopted for the present analysis when components contained optional items. When this was not an issue, however, $\omega_t$ was more precise and faster to compute, so both approaches were retained in order to manage tests with and without optional items. Components with optional items were identified manually by comparing response patterns between students.

For components that were either attempted by few students or for which all students received identical item scores, reliability estimates could not be computed using the methods described above. In these cases, reliability was imputed using the mean of the reliably estimated components. An alternative approach, namely excluding these components from the analysis, was considered but ultimately rejected, as doing so would have necessitated the removal of students who completed these components, thereby significantly reducing the overall sample size. Once reliability had been calculated using $\omega_t$, marginal reliability, or mean imputation, the standard error of measurement was obtained and new observed scores were generated for all students in the dataset following:

$$\hat{X}_i \sim \mathcal{N}(\hat{T}_i, SE_{\text{meas}}) \tag{5}$$

where $\hat{T}_i$ is given in Equation 3 and $\hat{X}_i$ is the new observed score for student $i$ that is different than their original observed score $X_i$.

## 3.7 Retakes

The simulation of retakes was restricted to all students who were not awarded the DP given their observed scores within the simulation. Hence, retake scores could also be generated for subjects that were in practice unavailable for retake during the November session. This allowed the generalisation of the study's findings to any form of retake, rather than the limiting case of November retakes. This also had the advantage of not requiring item-level data for November since reliabilities could be carried over. On the other hand, this meant that subject grade boundaries from May had to be used where grade boundaries for November had not been drawn up due to lack of purpose.

Since ability often increases between the initial assessment and subsequent retake, the observed normal distribution of component score increases from the real data was used to generate new true scores for these students. This distribution was obtained by performing a so-called piecewise linear transformation of the May and November component scores following their respective component grade boundaries, mapping these onto a scale from 0 to 100, then constructing the distribution of resulting score differences. The transformation was chosen because raw component scores could not directly be compared due to grade boundaries and maximum points available varying between sessions. For instance, a 12 in Physics may be awarded a 3 in May but a 2 in November if the difficulty of the test went down between

both. The piecewise linear transformation ensured this was taken into account when devising the improvement distribution. An in-depth overview of the transformation is provided in the Appendix. The Kolmogorov–Smirnov (KS) test was used to assess whether the score differences followed a normal distribution. This test evaluates the goodness of fit between observed data and a specified theoretical distribution, in this case the normal distribution (Berger and Zhou, 2014).

Improvement distributions were not computed for specific components as the November sample size was too small and not all components were available for retake, leading to some undefined distributions for certain components. Since the overall improvement distribution was computed for standardised scores, due to the piecewise linear transformation, using it for all components was permissible but less refined. For each student who retook a subject, a score increment was drawn from this distribution of observed improvements and added to their original true score to simulate them improving their ability between both tests. This then became their new true score for November.

To obtain new observed scores, measurement error was introduced in the same way as for the May observed scores, using the May 2023 test reliabilities to account for components not being present in the November dataset. This process was also repeated 1000 times, thereby producing synthetic observed scores for the retakes that could be linked to the original observed scores generated in section 3.6. Subsequently, to ensure that only improvements were retained, the highest observed score per subject, from the original attempt or the retake, was kept. Precision estimates after retake were then obtained by comparing both decisions in November.

## 3.8 Outcomes

The outcomes of interest comprised the mean and standard deviation of the precision estimates across 1 000 iterations, as summarised in Table 2b. These were compared across passing rules and between pre- and post-retake conditions in order to inform decisions regarding the optimal set of passing criteria for future application within the DP. Higher means indicate the rule performs better, whilst the standard deviation is used as a measure of simulation uncertainty, with lower values indicating the results can reliably be interpreted.

Additionally, a subgroup analysis concerning gender and region was conducted by considering specific subsets of the population. For this, mean precision was the sole outcome considered, under the assumption that observed trends in sensitivity and specificity within the overall population would similarly apply to subgroup analyses, provided comparable levels of precision.

## 3.9 Software

Data cleaning and analysis were performed in R Studio (Team, 2014). The mirt (Chalmers, 2012) and psych (Revelle, 2024) packages were used to estimate IRT models and calculate reliabilities. Fortran (Backus and Heising, 1964) was used to decrease simulation times for

the varying rulesets and methods of error generation. All R and Fortran code is provided on GitHub and can be used as the basis for future studies.

# 4 Results

## 4.1 Student scores

In the May session, 79.5% of participants who completed enough subjects to pass were awarded the DP. When transforming the scores using Kelley's formula, this went up to 80.7%. Both estimates are in line with figures published by the IB ("IB Diploma Stats", 2014). The number of failing students per passing criterion according to the true scores is given in table 5, excluding conditions that could not be tested given the data, such as the CAS requirement and the adjusted minimum for two SL subjects. Often more than two failing conditions overlap, so students may fall in more than one cell of the table. Hence the totals per condition do not correspond to column or row totals. 78.2% of the 13 955 failing participants failed to meet the 24-point minimum, making it the criterion most often responsible for a failure, although it was rarely triggered on its own. Conditions 4 to 7 were mostly triggered in conjunction with the 24-point minimum requirement. Obtaining a passing decision using the additional five proposed rules and the true scores yielded passing rates of 83.7%, 84.8%, 91.5%, 91.6% and 92.4% respectively, as the number of passing conditions decreased. Across every rule, the estimated true scores and real observed scores led to virtually the same pass/fail decisions. Agreement ranged from 97.7% under Rule 0 to 98.2% under Rules 3 and 4, confirming that the true-score shrinkage scarcely affected overall pass rates.

The mean total number of points achieved was 30.12 (SD = 7.73). For non-core subjects, students achieved 4.87 (SD = 1.31) on average, whilst the median grade for core subjects was C. 1.6% of the students retained for analysis did not complete enough components to be awarded a pass, even in the event of them obtaining good results on the components they did take. This is noteworthy for the simulation, as their results will always give them a fail. They were not easily removable as the number of components per subject per session was not provided to the researcher, thus they were retained as their impact was deemed to be minimal.

Concerning retakes, only 12.4% of students that failed in May retook a subject in November. This is probably due to certain subjects being unavailable for retake in November. From those who failed originally and did opt for a retake, 38.6% were able to pass after retake, increasing the overall passing rate after retake and under the current decision rule to 80.4% using the observed scores and 81.7% using the true scores. Using the latter, passing rate after retake was highest for rule 5, at 92.5%. Retaking students retook 2.30 subjects on average in November (SD = 1.23), with only one student being credited with retaking all eight subjects (3 SL, 3 HL, TK and EE). 60.3% of retaken subjects were retaken at HL, 33.5% at SL and the remaining 6.2% were retaken as part of the core. The distribution of subject groups retaken aligns closely with the distribution before retake, with arts subjects being the least

**Table 5:** Number of students failing each DP passing condition, using true scores. Diagonal: only that condition failed; Upper triangle: overlapping failures.

| **Failing Condition** | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 1. 24 point minimum not met | 1 860 | 7 252 | 4 473 | 2 216 | 473 | 1 410 | 4 443 |
| 2. 12 HL point minimum not met | | 1 928 | 2 963 | 1 945 | 413 | 1 317 | 3 709 |
| 3. 9 SL point minimum not met | | | 834 | 1 762 | 366 | 1 198 | 2 682 |
| 4. Grade 1 or E awarded | | | | 131 | 273 | 622 | 1 073 |
| 5. Grade N awarded on core | | | | | 27 | 96 | 167 |
| 6. More than 2 grade 2s | | | | | | 2 | 1 201 |
| 7. More than 3 grades $\leq 3$ | | | | | | | 33 |
| Total | 10 919 | 9 238 | 5 328 | 2 396 | 505 | 1 413 | 4 505 |

popular both before and after retake. The mean total points and non-core subject points for all retaking students during retake were 26.33 (SD = 5.78) and 3.86 (SD = 1.23) respectively. The median grade for core subjects remained C.

## 4.2 Score improvement

When mapping component scores onto a scale from 0 to 100, the mean component scores were 59.68 (SD = 21.08) and 48.67 (SD = 20.94) for all components taken in May and November respectively. The mean score difference between both sessions was 3.58 (SD = 14.57), although a significant number of students did not improve their grade at all, as evidenced by the median 0. This is partly attributable to students being allowed to retake only certain components of a subject. Consequently, certain component scores, such as those from internal assessments, typically completed throughout the academic year, are partially carried forward from the May session and combined with newly obtained scores from the November retake to determine the updated subject grade. As it was not possible to distinguish between carried-forward and newly acquired scores in the available data, all scores were retained in the retake dataset. Some students were also awarded a lower grade than during their original attempt.

The KS test rejected the hypothesis that these improvements are normally distributed (D = 0.20, p < 2.2e-16), although this is likely caused by the aforementioned abundance of zeroes. This is further confirmed by a QQ-plot and acceptable kurtosis (3.21) and skewness (0.03), suggesting a normal distribution may still be used to simulate score improvements.

## 4.3 Reliability

827 and 152 component reliabilities were evaluated using McDonald's $\omega_t$ and marginal reliability respectively. 76 components were not taken by enough students, meaning their reliability was imputed as the mean of the others. The reliability of an additional 6 were also imputed due to lack of differing item scores. The mean reliability was 0.88, whilst the standard deviation before imputation was 0.006. Setting the acceptable reliability threshold at 0.8, 861 components were reliable before imputation. Amongst these, certain may have

**Table 6:** Classification metrics (mean in % (standard deviation)) before retake for each decision rule, PPV = Positive Predictive Value, NPV = Negative Predictive Value.

| Set | Sensitivity | Specificity | PPV | NPV | Precision |
|---|---|---|---|---|---|
| Rule 0 | 97.5 (5e-04) | 90.8 (2e-03) | 97.8 (5e-04) | 89.3 (2e-03) | 96.2 (6e-04) |
| Rule 1 | 97.9 (5e-04) | 90.5 (2e-03) | 98.1 (4e-04) | 89.4 (2e-03) | 96.7 (5e-04) |
| Rule 2 | 98.0 (5e-04) | 90.2 (2e-03) | 98.2 (4e-04) | 89.1 (2e-03) | 96.8 (5e-04) |
| Rule 3 | 98.4 (4e-04) | 88.5 (3e-03) | 98.9 (3e-04) | 83.6 (4e-03) | 97.5 (5e-04) |
| Rule 4 | 98.4 (4e-04) | 88.4 (3e-03) | 98.9 (3e-04) | 83.2 (4e-03) | 97.5 (5e-04) |
| Rule 5 | 99.0 (4e-04) | 65.6 (4e-03) | 97.2 (3e-04) | 84.6 (5e-03) | 96.5 (4e-04) |

been too reliable, indicating perhaps an issue of redundancy of certain items (Streiner, 2003). Indeed, 22 components had reliabilities between 0.98 and 1. On the other hand, 10 components displayed reliabilities below 0.6, the lowest of which was 0.45 for a component in Dutch language acquisition. Upon further investigation, these low values were primarily attributed to the relatively low difficulty of the components for the population taking the exam, which made high marks easily attainable. As a result, minor variations in item scores were more likely to reflect random fluctuations than meaningful differences in student profiles.

## 4.4 Precision without retake

Boxplots of the precision over 1 000 iterations are presented in Figure 2. Overall, precision was remarkably high for all rules compared to the study by van Rijn et al. (2012). Standard deviations are small, indicating results are consistent across iterations and differences between rules can easily be interpreted. Precision before retake was lowest for the current passing criteria and highest for rules 3 and 4. The latter two had near identical precisions, owing to their similar requirements. Lowering and removing the requirement for HL subjects had similar impacts on precision.

A summary of misclassification metrics is presented in Table 6. Reducing the number of criteria increased sensitivity but lowered specificity. This was expected, as relaxing the criteria reduces the chances of a proficient student failing by accident, while increasing the chances of an non-proficient student passing by luck. The precision differences could thus largely be explained by the wide gap between the number of true passes and fails, making sensitivity a good approximation of precision. However, when the number of false positives became too high, this relationship broke down. Thus, in extreme cases, such as with Rule 5, which exhibited a specificity 25.2% lower than that of Rule 0, even a high count of true passes was insufficient to ensure high precision.

## 4.5 Precision with retake

After the retake, precision for rules 1 and 2 decreased slightly, whilst precision for the other rules either increased or remained stable. Notably, precision for rule 5 was 1.7% higher

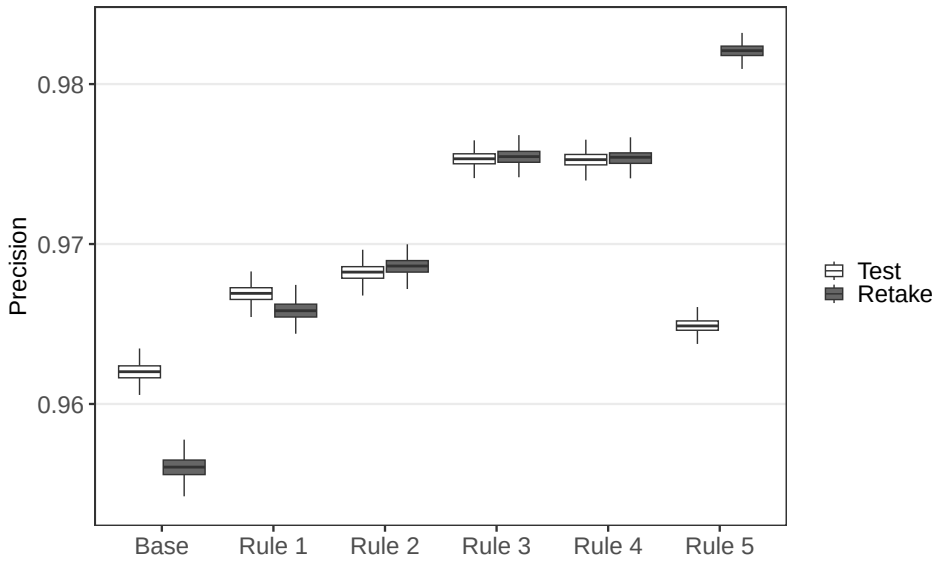


**Figure 2:** Boxplots of precision per decision rule on test and retake.

than before the retake. Given the initially high passing rate under rule 5, it is likely that most students attained the proficiency defined by the criteria after retake. Consequently, the precision is bounded on the lower end by the high passing rate, assuming that sensitivity remains high.

This effect could also be observed for the other rules, but was counteracted by a sharp increase in false positives, as evidenced by the lower specificities reported in Table 7. This increase could be readily explained by non-proficient students being given a second opportunity to pass by chance. Nevertheless, the same pattern emerged connecting sensitivity with simpler rules. This was mirrored in the precision, with rule 5 boasting the highest precision. Although increases in sensitivity were once again associated with losses in specificity, rule 1 demonstrated both higher sensitivity and specificity than the current passing criteria. This was however not the case when removing higher and lower level requirements altogether in rule 2, perhaps due to the additional requirement in rule 1 being more impactful to passing decisions over two sessions. Precision for rules 3 and 4 remained identical.

## 4.6 Demographics

Passing rates before retake differed markedly between IB regional offices but were almost uniform across gender. In Asia-Pacific (IBAP) 91.7% of candidates satisfied the passing conditions, followed by Africa-Europe–Middle East (IBAEM) at 89.1%. By contrast, the proportions fell to 68.2% in Latin America (IBLA) and 59.2% in North America (IBNA). The comparatively low success rates in the American regions stem, in part, from schools entering candidates for individual DP subjects to earn credit toward national qualifications rather than for the full Diploma. 81.8%, 80.2% and 77.9% of female, male and non-binary students respectively passed the DP. Table 8 presents the precision before retake across gender and

**Table 7:** Classification metrics (mean in % (standard deviation)) after retake for each decision rule, PPV = Positive Predictive Value, NPV = Negative Predictive Value.

| Set | Sensitivity | Specificity | PPV | NPV | Precision |
|---|---|---|---|---|---|
| Rule 0 | 99.4 (3e-04) | 71.5 (4e-03) | 95.7 (7e-04) | 95.1 (2e-03) | 95.6 (7e-04) |
| Rule 1 | 99.6 (2e-04) | 71.6 (4e-03) | 96.7 (6e-04) | 95.6 (3e-03) | 96.6 (6e-04) |
| Rule 2 | 99.7 (2e-04) | 70.0 (4e-03) | 96.9 (6e-04) | 95.7 (3e-03) | 96.9 (5e-04) |
| Rule 3 | 99.7 (2e-04) | 67.6 (6e-03) | 97.7 (5e-04) | 94.6 (4e-03) | 97.5 (5e-04) |
| Rule 4 | 99.7 (2e-04) | 67.2 (5e-03) | 97.7 (5e-04) | 94.4 (4e-03) | 97.5 (5e-04) |
| Rule 5 | 99.8 (1e-04) | 62.9 (7e-03) | 98.3 (4e-04) | 95.1 (5e-03) | 98.2 (4e-04) |

regional offices. Non-binary and Latin American students are left out due to smaller sample sizes, although precision related to these can trivially be obtained using figures from the other groups. Standard deviations are small, indicating the results can reliably be interpreted.

The relation between number of true passes and precision was once again highlighted, with higher-scoring IB offices also exhibiting high precision, whereas male and female groups, whose pass rates were virtually identical, displayed comparable precision under every decision rule. The lowest precision was observed in North America using rule 5, for which precision was 2.8% lower than when using rule 4. This may have been due to the larger number of true fails obtaining a pass given the more lenient criteria.

This trend was not reversed when including a retake, although precision was lower in North America for all rules except 5. This was likely caused once again by the high number of true fails being falsely passed across two attempts. Conversely, precision using each rule was higher after retake for the other IB offices and for both genders, as was to be expected to justify the results from Section 4.5. The highest precision was recorded in Asia-Pacific using rule 5, at 99.5%. The current passing criteria fared between 2.2% and 3.1% worse than rule 5 across all subgroups. Granular precision estimates such as sensitivity and specificity were not computed for demographic groups as they were assumed to be tied to precision in an way analogous to that observed on the whole data.

# 5 Discussion

The results presented here reveal that the DP classifies students remarkably well in comparison with other educational systems examined previously (van Rijn et al., 2012). For all rules applied to the simulated data, overall precision was consistently above 95%, meaning few students had their passing decision reversed due to measurement error. This may be due to the hierarchical structure of DP scores, in which a plurality of both component and subject scores are used to reach a decision. In such a system, it is unlikely that an uncharacteristically low component score will not be compensated by one of the other components, or a different subject. Thus, except for the odd unreliable component, students will obtain similar scores to their true score most of the time and therefore the same passing decision under every

**Table 8:** Precision (mean in % (standard deviation)) without retake, by gender and IB regional office. Offices: IB Africa, Europe and Middle East (IBAEM), IB Asia-Pacific (IBAP) and IB North America (IBNA).

| Set | Male | Female | IBAEM | IBAP | IBNA |
|---|---|---|---|---|---|
| Rule 0 | 95.8 (9e-04) | 96.5 (8e-04) | 96.9 (8e-04) | 97.5 (9e-04) | 94.4 (2e-03) |
| Rule 1 | 96.5 (8e-04) | 96.9 (7e-04) | 97.5 (7e-04) | 98.1 (8e-04) | 94.7 (1e-03) |
| Rule 2 | 96.6 (8e-04) | 97.0 (7e-04) | 97.6 (7e-04) | 98.2 (8e-04) | 94.9 (2e-03) |
| Rule 3 | 97.4 (7e-04) | 97.6 (6e-04) | 98.4 (6e-04) | 98.8 (7e-04) | 95.4 (1e-03) |
| Rule 4 | 97.4 (7e-04) | 97.6 (6e-04) | 98.4 (6e-04) | 98.8 (7e-04) | 95.4 (1e-03) |
| Rule 5 | 96.4 (7e-04) | 96.5 (6e-04) | 98.1 (5e-04) | 98.8 (6e-04) | 92.6 (1e-03) |

rule. The only students susceptible to change are those whose true grades sit on the border between passing and failing, thus leading to the low standard deviations observed both before and after retake. This conclusion provides support for the use of composite decision rules within high-stakes assessments.

Furthermore, finding the optimal decision rule within a multiple-subject parameter space is not trivial and it is likely that the DP passing criteria can be optimised beyond what is presented here. Nevertheless, the current research suggests that the requirements imposed on SL and HL subjects may feasibly be relaxed or removed to improve the overall precision of the diploma. This was evidenced by higher precisions both before and after retake when doing so, without drastically reducing the specificity of the diploma. Specificity after retake was even surprisingly higher when reducing the HL requirement to 9 points.

The other rules also exhibited higher precision than the current passing criteria, although this was achieved through increased sensitivity at the expense of specificity. This was most apparent when removing the requirement concerning number of 3s in rule 5. This led to a specificity before retake 25.2% lower than what was achieved with the current rule. Although less pronounced for the other rules, this highlights the requirement to assign weights to false positives and negatives when amending the passing criteria. Given the minutest gains in sensitivity and precision when reducing the minimum total of points, it is likely that the loss in specificity may not be worthwhile. An internal review of the results presented here will decide what can feasibly be amended without compromising the quality of the diploma. Moreover, the specificity is highest before retake for the current passing criteria, indicating it would be reasonable to retain them going forward in light of this study.

Another issue related to lowering the total required to pass is that this would increase the passing rate by over 10%, which could influence perceptions of the diploma's overall standard, potentially enhancing its accessibility but also risking a diminished reputation for selectivity or academic challenge. This passing rate increase can be explained by most student totals hovering around 24 and above before measurement error was added, meaning few students were close to the 18 and 20 thresholds. In contrast, both reducing and eliminating the HL

and SL requirements have minimal impacts on the passing rate, keeping it more or less in line with figures from previous years "IB Diploma Stats", 2014. These findings may however be inconsequential, as students taking the May and November 2023 exams were aware of the passing criteria when taking them. As such, students may not have tried as hard had the requirements been lower. Alternatively, students who did not believe they would pass under the current criteria may have opted to take the exams regardless, rather than postponing. Consequently, little can be concluded about the practical passing rate, although there remains scope to maintain it at around 80% if subject grade boundaries are adjusted in line with the new passing criteria.

Regarding retakes, the true distribution of score improvements could not reliably be determined. While the distribution used in the simulation was plausible, its large standard deviation limited its informativeness. Nevertheless, the large loss in specificity across all rules highlighted the potential danger of allowing unlimited retakes. Conversely, the high sensitivities observed indicated that the IB is successful in ensuring proficient students are passed within two attempts. This mitigates several stakes of the DP, namely the financial and psychological costs associated with having to retake courses unnecessarily. Of course, this hinges on courses being available to retake in November, which may not always be the case. The methods used to simulate retakes are sound and may be applied in other contexts. The approach of adding a score increment to simulate ability change between test and retake has, to the best of the author's knowledge, never been taken before but is realistic in that it may also return score decreases in situations where students did not look at the material enough before the retake.

Precision comparisons between demographic groups did not reveal any differences in how the rules performed, indicating no region nor gender stood to benefit more than another from a potential rule change. The large disparities in passing rates between regions did however enable the juxtaposition of the effect of retakes between high- and low-performing student groups. This revealed that precision went down with retake when the number of true fails was high, whilst it went up when the opposite held true.

# 6 Conclusion

The current research provided evidence to support simplifying the passing criteria of the IB diploma programme. The suggested modifications included removing the requirements for HL and SL subjects, as well as lowering the total number of points required to pass. The latter yielded higher rates of false positives and passing than the current set of passing criteria, casting doubt on the justification of implementing such a change. Retakes improved precision of the diploma but heavily decreased its specificity. Overall, the diploma effectively discriminated between proficient and non-proficient students. No gender or regional effects were identified as mediators of the relationship between ruleset and precision, although passing rate was suggested as an explanation for disparities in the effect of retake on precision in North America compared to other IB offices.

A limitation of the current research concerns the realism of the conditions simulated. Although some of the subjects unavailable for retake in November might have been available in May of the next year instead, others may not trivially have been retakable. For instance, some of the components were based on continuous assessment and would have required retaking the whole course to obtain a new grade. Furthermore, the assumption that students who failed would retake all their courses is unrealistic, as often only improving one of the subject grades would have been sufficient to pass. Students who failed by a lot are also less likely to opt for a retake. A solution that was considered to bridge this issue was to use a logistic model to predict whether a student would choose a retake or not, but the small sample size in November was deemed insufficient to build a model without introducing too much variation between iterations.

Using Kelley's formula to adjust for unreasonably high or low obtained scores represents a novel approach to simulating student performance. While this method effectively reduced the number of students failing due to minimum-related criteria, it may have inflated passing rates across rules and, consequently, led to precision levels higher than those likely to occur in practice. Additionally, the distribution of retake score increments did not account for these adjusted true scores and was based on only a small subset of the May component population. As such, the increment distribution may not accurately reflect performance in components unobserved during the retake. Nevertheless, passing rates derived from true scores closely aligned with official IB figures, suggesting the impact on overall results was limited.

The methods presented here can be applied to other high-stakes assessments both within and beyond the IB. In some contexts, such as driving tests, researchers may prioritise sensitivity or specificity over precision, particularly when minimising false positives is critical. There is also scope to simulate multiple retakes, though this is constrained by the tendency of scores to increase indefinitely with each increment. Another possible extension concerns finding an optimal composite passing rule analytically, as this avoids a sub-optimal passing rule being retained instead. This may also be faster, provided the parameter space to explore is restricted beforehand.

## Note on the author

At the time the research was performed, Adam Maghout was a Master's student at Utrecht University and an intern at the IBO.

## Ethical approval and reproducibility

This research was approved by the Ethical Review Board of the Faculty of Social and Behavioural Sciences at Utrecht University (approval code: 24-2025). The full analysis code is openly available on GitHub. For access to the underlying data, please contact Dr. Anton Béguin at anton.beguin@ibo.org.

## Acknowledgements

The author wishes to express sincere gratitude to Dr. Anton Béguin, Director of Educational Innovation at the IBO, for his valuable supervision and guidance throughout the course of this research. The author is also grateful to Utrecht University for its support in academic writing and for facilitating regular check-ins, which contributed significantly to the progress and completion of this work.

# References


American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.

Au, W. W. (2008). Devising inequality: A Bernsteinian analysis of high-stakes testing and social reproduction in education. *British Journal of Sociology of Education*, *29*(6), 639–651. https://doi.org/10.1080/01425690802423312

Backus, J. W., & Heising, W. P. (1964). Fortran. *IEEE Transactions on Electronic Computers*, *EC-13*(4), 382–385. https://doi.org/10.1109/PGEC.1964.263818

Berger, V. W., & Zhou, Y. (2014). Kolmogorov–Smirnov Test: Overview. In *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118445112.stat06558

Blazer, C. (2011, January). *Unintended Consequences of High-Stakes Testing. Information Capsule. Volume 1008* (tech. rep.). Research Services, Miami-Dade County Public Schools.
ERIC Number: ED536512.

Chalmers, R. P. (2012). Mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*, 1–29. https://doi.org/10.18637/jss.v048.i06

Cho, E. Y.-N., & Chan, T. M. S. (2020). Children's wellbeing in a high-stakes testing environment: The case of Hong Kong. *Children and Youth Services Review*, *109*, 104694. https://doi.org/10.1016/j.childyouth.2019.104694

Coggeshall, W. S. (2021). An Examination of Classification Accuracy in the Continuous Testing Framework. *Educational Measurement: Issues and Practice*, *40*(1), 28–35. https://doi.org/10.1111/emip.12398

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Dai, S., Vo, T. T., Kehinde, O. J., He, H., Xue, Y., Demir, C., & Wang, X. (2021). Performance of Polytomous IRT Models With Rating Scale Data: An Investigation Over Sample Size, Instrument Length, and Missing Data. *Frontiers in Education*, *6*. https://doi.org/10.3389/feduc.2021.721963

Douglas, K. M., & Mislevy, R. J. (2010). Estimating Classification Accuracy for Complex Decision Rules Based on Multiple Scores. *Journal of Educational and Behavioral Statistics*, *35*(3), 280–306.

DP passing criteria. (2023, December).

Gulliksen, H. (1950). *Theory of mental tests*. New York, Wiley.

Hayes, A. F., & Coutts, J. J. (2020). Use Omega Rather than Cronbach's Alpha for Estimating Reliability. But... *Communication Methods and Measures*, *14*(1), 1–24. https://doi.org/10.1080/19312458.2020.1718629

Hill, I., & Saxton, S. (2014). The International Baccalaureate (IB) Programme: An International Gateway to Higher Education and Beyond. *Higher Learning Research Communications*, *4*(3), 42–52
ERIC Number: EJ1133258.

IB Diploma stats. (2014, August).

Kelley, T. (1923). *Statistical method*. The Macmillan Company.
Open Library ID: OL6654500M.

Lord, F. M. (1962). Cutting scores and errors of measurement. *Psychometrika*, *27*(1), 19–30. https://doi.org/10.1007/BF02289661

Maire, Q. (2021). The IB Diploma from Globalisation to Credential Theory. In Q. Maire (Ed.), *Credential Market: Mass Schooling, Academic Power and the International Baccalaureate Diploma* (pp. 29–51). Springer International Publishing. https://doi.org/10.1007/978-3-030-80169-4_2

Marchant, G. (2004). What is at Stake with High Stakes Testing? A Discussion of Issues and Research1. *Ohio Journal of Science*, *104*, 2–7.

McDonald, R. P. (1999, July). *Test Theory: A Unified Treatment*. Psychology Press. https://doi.org/10.4324/9781410601087

Mehrens, W. A., & Phillips, S. (1989). Using College GPA and Test Scores in Teacher Licensure Decisions: Conjunctive Versus Compensatory Models. *Applied Measurement in Education*, *2*(4), 277. https://doi.org/10.1207/s15324818ame0204_1

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an Em Algorithm. *ETS Research Report Series*, *1992*(1), i–30. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Retaking examinations. (2023, July).

Revelle, W. (2024). *Psych: Procedures for psychological, psychometric, and personality research* [R package version 2.4.6]. Northwestern University. Evanston, Illinois. https://CRAN.R-project.org/package=psych

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*(4, Pt. 2), 100–100.

Smyth, E., & Banks, J. (2012). High stakes testing and student perspectives on teaching and learning in the Republic of Ireland. *Educational Assessment, Evaluation and Accountability*, *24*(4), 283–306. https://doi.org/10.1007/s11092-012-9154-6

Streiner, D. L. (2003). Starting at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency. *Journal of Personality Assessment*, *80*(1), 99–103. https://doi.org/10.1207/S15327752JPA8001_18

Tay, I. (2023). The Movement of International Education Towards the Globalising Approach: Comparing the International Baccalaureate Diploma Programme and the International A-Levels. *Journal of International and Comparative Education (JICE)*, 87–102. https://doi.org/10.14425/jice.2023.12.2.1205

Team, R. (2014). R: A language and environment for statistical computing. *MSOR connections*.

van Rijn, P., Béguin, A., & Verstralen, H. (2012). Educational measurement issues and implications of high stakes decision making in final examinations in secondary education in the Netherlands. *Assessment in Education: Principles, Policy & Practice*, *19*(1), 117–136. https://doi.org/10.1080/0969594X.2011.591289

Vermeulen-Kerstens, .,, Scheepers, .,, Adriaans, .,, Arends, L., van den Bos, R., Bouwmeester, S., van der Meer, F.-B., Schaap, L. (, Smeets, G., van der Molen, H., & Schmidt, H. (2012). Nominaal studeren in het eerste jaar. *Tijdschrift voor Hoger Onderwijs*, *30*(3), 204–216.

Waldow, F. (2009). What PISA Did and Did Not Do: Germany after the 'PISA-shock'. *European Educational Research Journal*, *8*(3), 476–483. https://doi.org/10.2304/eerj.2009.8.3.476

Yocarini, I., Bouwmeester, S., Smeets, G., & Arends, L. (2018). Systematic Comparison of Decision Accuracy of Complex Compensatory Decision Rules Combining Multiple Tests in a Higher Education Context. *Educational Measurement: Issues and Practice*, *37*(3), 24–39. https://doi.org/10.1111/emip.12186

# Appendix: Piecewise Linear Transformation Algorithm

1. **Collect grade boundaries.** For each grade $g \in \{1, \ldots, 7\}$ record its lower and upper component-score cut-scores $\ell_g$ and $u_g$.

2. **Locate the grade band.** Given a component score $x$, determine the unique grade $g$ for which $\ell_g \leq x \leq u_g$.

3. **Set the band's starting point.**

$$B_g \;=\; \frac{100}{7}\,(g-1).$$

   This is the score on the 0–100 scale that corresponds to the lowest mark of grade $g$. Thus g=1 $\Rightarrow B_1 = 0$, $g = 2 \Rightarrow B_2 \approx 14.29$ for the first two grade boundaries.

4. **Add the within-band offset.**

$$D = \frac{100}{7}\,\frac{x-\ell_g}{u_g-\ell_g}, \qquad y = B_g + D.$$

   The offset $D$ places $x$ proportionally within its band, so the lowest and highest scores in grade $g$ map to the band's lower and upper limits, respectively.