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

Adam Maghout

Student number: 7692617

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

*Supervisor: Anton Béguin*

*(International Baccalaureate)*

Word count: 2485

FETC approval: 24-2025

Due date: 22/12/2024

Submission date: 14/01/2025

# 1 Introduction

International assessment studies such as PISA have had a lasting impact on how education systems are perceived (Waldow, 2009). Beyond improving cross-country performance, the quality of tests carries significant financial, legal and psychological implications, with suicide attempts spiking in recent years during exam periods (George, 2024).

Thus, when the cost of assessments for students is substantial, for example with school leaving exams, it is good practice not to rely on a single exam score for pass/fail decisions, as suggested in the Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014). Instead, a composite approach is favoured, wherein a decision rule defines the combination of multiple test results used to make the pass/fail decision.

The International Baccalaureate Organisation (IB0) is a transnational non-profit that offers courses to globally mobile students aged 3 to 19 (Doherty, 2009) and provides internationally recognised qualifications (King and Raghuram, 2013). Among these is the Diploma Programme (DP), offered since 1970 and akin to British A levels (Doherty, 2009). As few analytical studies have assessed the quality of its measurement tools (Hill and Saxton, 2014), this study examined the precision of DP passing rules as part of an internal quality review designed to revise the passing criteria ahead of the 2032 exam session.

Few quantitative evaluations of high-stakes passing criteria exist. This study drew on the work of van Rijn et al. (2014), which examined the passing criteria precision within Dutch secondary education and found that compensatory rules reduced misclassification. Similarly, Yocarini et al. (2018) conducted a comparison of passing decisions on a psychology bachelor's program at Erasmus University Rotterdam.

While these studies considered measurement error at subject level, the impact of viewing measurement error at item level remains unexplored. Comparing both approaches here will thus inform future evaluations of other exam systems. Moreover, retakes in high-stakes contexts remain unexamined quantitatively. The purpose of retakes is twofold: giving non-proficient examinees the opportunity to improve their competency and correcting the grades of proficient examinees who previously failed (Coggeshall, 2021). Therefore, this study also examined how retake exams in the DP influenced missclassification rates.

In sum, this research aimed to determine the extent to which current IB passing conditions accurately classify students and whether increased compensation is warranted through less stringent passing rules or additional opportunities for resits. Precision differences across countries and institution types were also analysed to inform targeted interventions.

# 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. Composite rules always outperform simple ones when tests are correlated (Vermeulen-Kerstens et al., 2012) as similar constructs are being measured. This assumption is however often violated when excess ability in one subject cannot compensate deficiency in another (Mehrens and Phillips, 1989).

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., 2014). This is illustrated in Figure 1. In practice, large-scale assessment rules often combine conjunctive and compensative aspects, which are referred to as complex decision rules (Douglas and Mislevy, 2010).

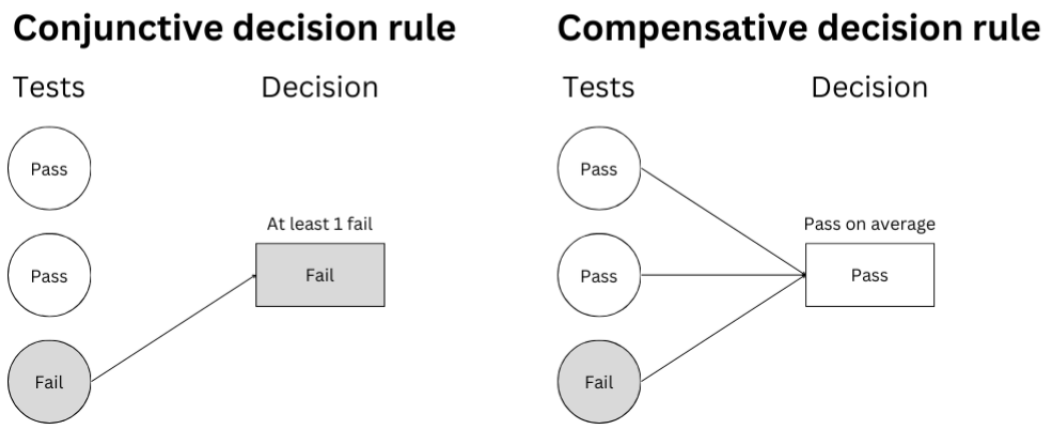


**Figure 1:** Conjunctive and compensative decision rules.

## 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). The exact options 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.

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, grades are derived from results on three examination papers and a practical assessment. An unattempted component yields an N. Students may also retake courses to pass or to improve a previous grade. 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 any subject, including TOK and 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 difficulty but reduce overlap between requirements. The present study seeks to evaluate the impact of modifying these rules on misclassification rates.

## 2.3 Measurement precision

In accordance with van Rijn et al. (2014), 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 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. 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:

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

where $\sigma_X$ is the standard deviation of the test scores. 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. The issue then becomes estimating the true scores, which can be done using simulation-based methods.

Although methods for creating composite scores exist (Gulliksen, 1950), these are ill-suited for the DP's subject-specific criteria. Thus, rather than using reliability only, the notion of measurement precision is extended to the misclassification rate over all subjects. DP criteria are then precise if they yield few classification errors.

# 3 Methods

The research was undertaken in two phases, focusing on decision rule precision and the effectiveness of retakes. Using real examination data, the proportion of students passing was first calculated under each proposed decision rule. This was done as a quality measure to guarantee satisfactory passing rates beyond misclassification. To further evaluate these rules, misclassification tables were constructed through simulations that incorporated measurement error into the observed scores.

## 3.1 Data

The analysis used two datasets of component and item scores for examinees who sat a DP component 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 96 191 took enough components to be assigned a passing decision. Students were anonymised via a candidate number that was linkable between the item and component datasets. Information was also provided concerning the gender and regional office of examinees, the distribution of which is provided in Table 1.

**Table 1:** Gender and regional office distribution (May 2023). Offices: IB Africa, Europe and Middle East (IBAEM), IB Asia-Pacific (IBAP), IB Latin America (IBLA) and IB North America (IBNA).

| | IBAEM | IBAP | IBLA | IBNA | **Total** |
|---|---|---|---|---|---|
| Male | 15 249 (15.9%) | 9 751 (10.1%) | 3 229 (3.4%) | 14 606 (15.2%) | 42 835 (44.5%) |
| Female | 18 461 (19.2%) | 10 090 (10.5%) | 3 901 (4.1%) | 20 339 (21.1%) | 52 791 (54.9%) |
| Other | 109 (0.1%) | 15 (0.0%) | 14 (0.0%) | 427 (0.4%) | 565 (0.6%) |
| **Total** | 33 819 (35.2%) | 19 856 (20.6%) | 7 144 (7.4%) | 35 372 (36.8%) | 96 191 (100%) |

In total, 1055 components were assessed across 219 subjects. Some components, primarily in the core, were taken by most students whilst others were only taken by one examinee, such as Welsh or Guarani. In total, components were taken in 73 different languages. 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.

## 3.2 Proposed Passing Criteria

While 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 risk of failure. To address these concerns, the proposed amendments to the current passing criteria were as follows:

- 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.

A reduction of the minimum total points to 18 was proposed, corresponding to an average grade of 3, 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, five sets of passing criteria were applied to both simulated and observed data, given in Table 2. These sets account for removing HL/SL requirements alongside lowered totals.

**Table 2:** Studied sets of decision rules for the Diploma Programme.

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

## 3.3 Simulations

To evaluate misclassification, a process was derived to obtain observed and true scores to compare. Assuming component grades follow a multivariate normal distribution, one can accept that where some examinees obtain lower scores than their true value, others will obtain higher scores. Thus, the observed scores from the data can be taken as the true scores for the purposes of this study. From there, it is then possible to generate new observed scores by adding measurement error to these true scores, as is described in equation 1.

Measurement error was added at item and component level to compare both methods and to obtain robust misclassification rates. Item-level error preserved variance from inconsistent items in the generated observed scores. Component-level error on the other hand assumed uniform item variance within components. This approach was more straightforward but less sensitive to unreliable items. Both methods assumed uncorrelated errors, meaning that observed correlations between scores arose from correlations in the true scores rather than from the errors themselves.

The simulations were set up as follows, in accordance with van Rijn et al. (2014):

1. Draw error scores from multivariate normal distribution.
2. Obtain observed scores by adding error to true scores.
3. Apply the decision rule to both true and observed scores.
4. Compute the classification table for true and observed decisions.

To ensure the results were stable, this process was repeated 1000 times, for measurement error at item and component level. Mean misclassification rate and standard deviation are then reported for each of the passing criteria proposed above.

## 3.4 Classical Test Theory

To apply measurement error to true scores, the reliability of components must 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:

$$\omega_t = \frac{\left(\sum \lambda_j\right)^2}{\left[\left(\sum \lambda_j\right)^2 + \sum\left(1 - \lambda_j^2\right)\right]} \tag{3}$$

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 $\lambda_j$ analytically, one of which is to use hierarchical factor analysis as performed by the omega function in the psych package (Revelle, 2024) used for this 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.

The mirt package (Chalmers, 2012) provides an alternative method to calculating reliability in the form of marginal reliability. For this, an 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 Graded Response Model was fitted to the items before computing the reliability (Samejima, 1969) as it is suitable for polytomous data.

Once reliability had been calculated using either $\omega_t$ or marginal reliability, the standard error of measurement was obtained and misclassification tables were constructed.

## 3.5 Item Response Theory

To add item-level error, a Graded Response Model was first fit to all components. This yielded a probability to obtain a score $k$ given an ability level $\theta$:

$$P(X = k \mid \theta) = P(X \geq k \mid \theta) - P(X \geq k + 1 \mid \theta) \tag{4}$$

In the mirt package (Chalmers, 2012), $P(X \geq k \mid \theta)$ is computed using a two-parameter logistic model, where Equation 4 is the difference between two known probabilities given by:

$$P(X \geq k \mid \theta) = \frac{1}{1 + e^{-a(\theta - b_k)}} \tag{5}$$

where $a$ is an item-level discrimination parameter and $b_k$ is the item difficulty parameter that is estimated for each threshold (Zein and Akhtar, 2025).

Once the probability of obtaining each score was calculated for students using 4, cumulative probabilities were derived that determined the probability of obtaining a score $k$ or under for each student. A uniform random number $u \sim U(0, 1)$ was then sampled and the smallest $k$ such that $P(X \leq k) > u$ was assigned to students, ensuring the most probable score was usually assigned. Applying this procedure to all items then aggregating them introduced measurement error into component scores given $u$ was random.

## 3.6 Retakes

The simulation of retakes was restricted to students who actually retook a subject in November 2023. Since ability often increases between the initial assessment and subsequent retake, the observed distribution of score increases from the real data was used to generate new true scores for these students. Specifically, for each student who retook a subject, a score increment was drawn from the distribution of observed improvements and added to their original true score to emulate them improving their ability between both tests.

To obtain new observed scores, measurement error was introduced within the CTT framework, using the November 2023 test reliabilities since the May and November tests differed. This process was repeated 1000 times, thereby producing synthetic observed scores for the retakes that could be linked to the original scores generated in section 3.4. Subsequently, to ensure that only improvements were retained, the highest observed score per subject, from the original attempt or the retake, was kept. Finally, a misclassification table was built comparing true scores after retake with observed scores after retake.

Retakes were not evaluated with measurement error added at item level as person abilities were measured on different scales for different items and thus increments could not easily be added.

## 3.7 Software

Data cleaning and analysis were performed in R (R Team, 2014). The mirt (Chalmers, 2012) and psych (Revelle, 2024) packages were used to estimate the IRT models and calculate the various reliabilities. Fortran (Backus and Heising, 1964) was used to speed up simulations.

# References


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

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

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

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

Doherty, C. (2009). The appeal of the International Baccalaureate in Australia's educational market: A curriculum of choice for mobile futures. *Discourse: Studies in the Cultural Politics of Education*, *30*(1), 73–89. https://doi.org/10.1080/01596300802643108

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). Retrieved December 13, 2024, from https://www.ibo.org/about-the-ib/what-it-means-to-be-an-ib-student/recognizing-student-achievement/about-assessment/dp-passing-criteria/

George, D. A. S. (2024). Exam Season Stress and Student Mental Health: An International Epidemic. *Partners Universal International Research Journal*, *3*(1), 138–149. https://doi.org/10.5281/zenodo.10826032

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.

King, R., & Raghuram, P. (2013). International Student Migration: Mapping the Field and New Research Agendas. *Population, Space and Place*, *19*(2), 127–137. https://doi.org/10.1002/psp.1746

Maire, Q. (2021). The IB Diploma from Globalisation to Credential Theory. In Q. Maire (Ed.), *Credential Market: Mass Schooling, Academic Power and the International Baccalau-*

*reate Diploma* (pp. 29–51). Springer International Publishing. https://doi.org/10.1007/978-3-030-80169-4_2

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

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

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.

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

van Rijn, P., Béguin, A., & Verstralen, H. (2014). Educational measurement issues and implications of high stakes decision making in final examinations in secondary education in the Netherlands. In *High-Stakes Testing in Education*. Routledge.

Vermeulen-Kerstens, L., Scheepers, A., Adriaans, M., 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

Zein, R. A., & Akhtar, H. (2025). Getting started with the graded response model: An introduction and tutorial in R. *International Journal of Psychology*, *60*(1), e13265. https://doi.org/10.1002/ijop.13265

# Appendix

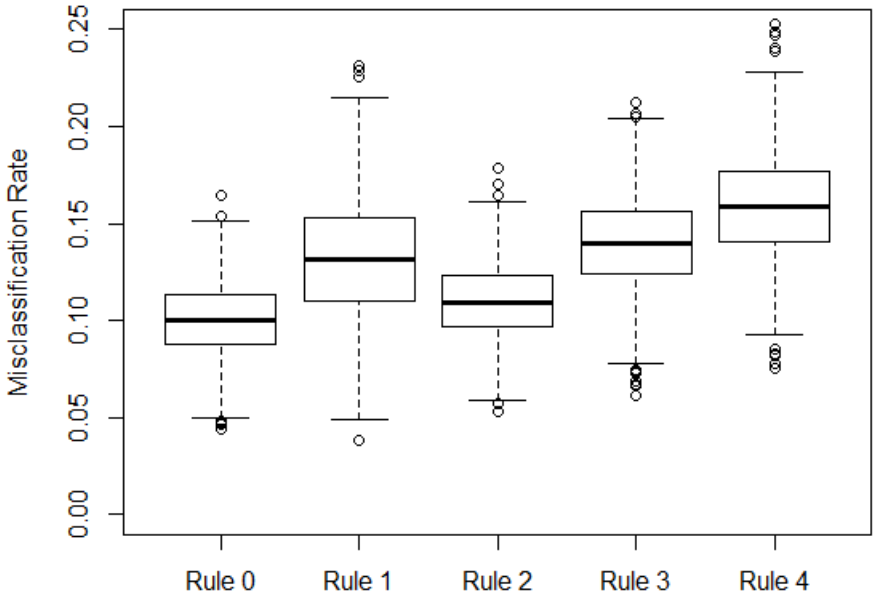


**Figure 2:** Synthetic misclassification rates without retake for the rules in Table 2, over 1000 datasets.