A high-level review of the fairness of the methodology used to adjust for the disruptions of the 2025 Selective High Schools (SHS) and Opportunity Class (OC) Tests

Report prepared for the NSW Department of Education by
Professor Jim Tognolini
Director of the Centre for Educational Measurement and Assessment (CEMA) at the University of Sydney, July 2025

Contents

INTRODUCTION

Entry decisions into selective education rely on formal placement tests that measure a student's overall academic ability. The Selective High Schools (SHS) Tests have four components. The Opportunity Class (OC) Tests have the same first three components but do not include writing. The components are:

- **Reading**
  Assesses a variety of comprehension skills using diverse texts.

- **Mathematical reasoning**
  Evaluates the ability to apply mathematical understanding and problem-solving across topics.

- **Thinking skills**
  Measures critical thinking and general problem-solving aptitude.

- **Writing (SHS only)**
  Requires one task to demonstrate creativity and effective writing for a given audience and purpose.

In 2025, 17,559 students applied to sit the SHS Tests and 13,263 applied to sit the OC Tests. Of these, 15,573 and 11,458 students respectively actually sat the SHS and OC Tests. The 2025 administration dates were 2, 3 and 4 May with the "Make-up-Test" scheduled for the 19th of May 2025.

On the first day of the test administration there were several technical and well-being issues which warranted intervention in the nominated 2025 SHS and OC Test process. One of the key interventions was that some students were offered the opportunity to re-sit the SHS and OC tests at a later date.

The purpose of this report is to provide advice on the fairness of the testing methodology that:

1. will be used to make the scores of students sitting the different test versions comparable; and,
2. was designed to be used to provide adjustments in the case of disruptions during the original tests.

The first section of this report summarises and reviews the methodology at a high level and provides an informed and evidence-based comment regarding the fairness of the testing methodology. It does not attempt to evaluate the methodology used from a psychometric point-of-view.

The second section describes the adjustments that were used to accommodate the disruptions and addresses their relative fairness by discussing some of the concerns that have been raised by stakeholders.

It is important to stress up-front, that there is no ideal solution to the problems created by the disruptions that occurred and that ultimately the decision regarding fairness must be evaluated in the context of the situation that existed at the time.

2025 TESTING METHODOLOGY FOR SHS AND OC TESTS

Changes to the 2025 SHS and OC test administrations

In 2025, the Reading, Mathematical Reasoning, and Thinking Skills tests were delivered online and automatically marked for the first time, while the writing task was completed by the students online but scored by human markers.

The final measure of students' academic ability is captured in a score comprising a weighted, combined Reading/Mathematical Reasoning/Thinking skills component and a Writing component (for the SHS score) to provide an overall score out of 120 for each student, which was then weighted to produce a final score out of 100.

This process is consistent with previous years apart from some adjustments to the weightings of the SHS. In 2025 all 4 components are weighted equally, whereas in 2024 the weighting of the components for the SHS were 35%, 25%, 25% and 15% for the Mathematical Reasoning, Reading, Thinking skills and Writing respectively. This change had been disseminated to students well in advance of the testing date. and as such would have not adversely affected the fairness of the 2025 administration.

A second change in 2025 was that there were four test dates planned for the SHS and OC Tests to be delivered online. A different version of the tests was to be delivered online each day for the three nominated administration days, and the fourth version of the test administered later for students who for one reason or another could not attend the original administration dates. This meant that there were four different SHS tests (and OC tests) to be administered in 2025.

The limitations of traditional testing

A serious limitation of traditional testing is that the scores that students obtain on the test can only be interpreted in terms of the test that they received. When any new versions of a test are constructed, even those that are designed to be equivalent to the original test, they will still be comprised of different test items (questions), such that the scores of the students that completed the different test versions cannot be fairly compared without transforming (i.e. linking) them. The move to online testing in 2025 generated a situation where it was necessary to compare the scores of students who had taken *different* versions (i.e., tests with different sets of items) of the OC and SHS tests.

Linking the tests

The process of transforming scores on one test so that they can be compared directly to scores on another is referred to as statistical test-form *equating, or, depending on the assumptions, linking*. As an integral part of the test construction process, test linking has received widespread coverage in the measurement and

psychometric literature. It is beyond the scope of this report to provide a survey of all evidence and rationale required for test equating and linking. However, the methods used to compare the performance of students who have completed different versions of the same test are well-researched, implemented for almost all assessments, and are fair.

The method used to link the different versions of the SHS (and OC) is widely used and includes the use of *common items* in the different versions of the tests. Using these common items and responses of the students in conjunction with Rasch Measurement Theory and the Rasch Model, student scores can be transformed into ability measures that can be legitimately compared even though the students have taken different test versions. There are at least six common items (sometimes referred to as link items) in the Reading, Thinking Skills, and Mathematical Reasoning sections in each of the SHS and OC tests. This represents approximately 20% of the total number of items in the tests and is relatively consistent with the number of common items used to link tests in similar programs. The same process is used to link different versions of the National Assessment Program – Literacy and Numeracy (NAPLAN) and several high stakes international tests (e.g., Programme for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS)).

The writing component of the SHS tests

Each of the test versions in the SHS has one writing task, but these tasks are different across the versions. The writing tasks have the same marking rubric and are marked by two human markers out of a score of 25. The scores from the two markers are added together to give a total score out of 50. Students complete the writing tasks online. The score distribution for the writing results is then adjusted to make it consistent with previous years. The distribution of the scores from the Reading/Mathematical Reasoning/Thinking skills tests are transformed onto the writing distribution to give some degree of comparability across years. The scores from the Reading/Mathematical Reasoning/Thinking skills tests and the writing component of the tests are then combined and weighted to produce the final scores for the students out of one hundred.

Weighting and combining the test scores to produce a final score

The process used to weight and combine the scores from the various components to generate the final scores for students has been made slightly more complex by the need to convert and weight scores from the different metrics generated using traditional test theory and Rasch theory in the same process. As I understand it, this was a transition methodology used to align 2025 practices with the way scores were produced in the past. Once again, these practices did not detract from the fairness of the 2025 testing process which was designed to reflect the previous psychometric methodology.

The use of tiebreak scores

In the event of tied scores, the final placement process may consider student level performance on the common items.

Technical difficulties during the first administration of the tests

One hundred and thirty one (131) technical disruptions (at various levels of magnitude) were reported at the major centres during the administration of the tests. Where appropriate, adjustments, in the form of extra time for students were applied during the test. More severe disruptions were managed through the established illness and misadventure process. The process for accommodating these adjustments had been agreed upon before the testing and was designed to mitigate any adverse effects on students' scores.

Summary

The 2025 SHS and OC testing program, being a transitional phase toward full online delivery, was designed to retain the psychometric robustness of prior administrations. The methodology aimed to ensure fairness by upholding established measurement principles despite technological changes.

Presumably, one of the main aims of moving to online delivery is to enhance test security. Having students sit different versions of the SHS and OC Tests moves some way to increasing the security of the tests. However, having different versions of the same online tests linked by common items administered at different times and days has the potential to create a security risk if students comment on the questions that they have sat for in the earlier version of their tests. Some evidence of this "sharing of items" practice has emerged now that students are aware that there are common items on the different versions of the tests. It might be worthwhile moderating the impact against such potential security breaches by considering in the future moving towards a process of using computer adaptive tests (CAT). Such tests tailor each assessment specifically for an individual. No students necessarily see all the same test items as other students and hence memorisation, collusion or cheating is all but eliminated.

ADMINISTRATION OF THE SHS AND OC TESTS ON 2nd of MAY 2025

Disruptions to the test administration on 2nd of May

There were significant disruptions to the administration of the SHS and OC Tests conducted on this date across the 3 large test centres. The disruptions were characterised by students having lengthy delays getting access to the centres, high levels of anxiety of parents and students attending these centres, crowd control and ultimately police intervention.

There are likely to be varying levels of impact on students' scores. The impact would have been difficult, if not impossible to measure and accurately accommodate using the existing protocols for illness and misadventure or other more sophisticated psychometric methods. The disruptors on this occasion at these centres were not directly related to technical issues associated with the administration, but were more related to the intangible impacts on student performance associated with their adverse reactions to anxiety caused by the delays. The decision to offer all students attending these centres the opportunity to re-sit the tests on another occasion was seen as the fairest solution to the problem. It is a solution that has been used on similar occasions where there have been disruptions to high stakes testing programs that could not be solved after the fact, psychometrically.

The second decision, related to the first, was that students who sat twice would be given the better of their two scores. While this decision is fair given that it gives the "benefit-of-doubt" to students, nonetheless, they are being asked to sit the test twice within a brief period. This may, in itself, potentially negatively impact their second test score and raises a question about fairness. It could be argued, that ***all students*** should have been given the opportunity to sit the test twice and given their best score. It could also be argued, of course, that students who were not affected by the disruptions at the three large centres did not have a reason to be given a re-sit and that giving them a re-sit would undermine the fairness of the process. In addition, test theory suggests that without directed practice, it is highly likely that students not affected by the original disruptions would score similarly, within measurement error, on both tests. Ultimately, the decision to offer only those students from the affected centres (where the majority of students sat the tests) a re-sit was made on the grounds of presiding test theory and logistics. It was considered to be impossible to run such a large-scale testing program for all centres within the brief period of time that the providers had to run another test administration.

It is important to repeat that there is no ideal solution to the problems created by the disruptions that occurred and that ultimately the decision regarding fairness has to be evaluated in the context of the situation that existed at the time and that, as a consequence, there will always be some stakeholders who will feel as though they have been disadvantaged or that others have been advantaged. Some stakeholders expressed their concerns about the process used to adjust for the disruptions with several questions challenging the fairness of the procedure. I will address these challenges from a test-fairness point of view later in this report.

The decision to re-sit the SHS Tests on the 17 and 25 May 2025 and the OC Tests on the 18 and 24 May has created distinct groups of students with different profiles of results.

The test score profile of student groups

The following table summarises the number of students re-sitting the tests at the rescheduled test dates of the 17 and 25 May.

	Total number of students
Applied to sit 2025 tests for 2026 SHS placements	17,559
Applied to sit 2025 tests for 2026 OC placements	13,263
Sat the first administration of the SHS Test (Versions 1, 2 and 3)	7240
Sat the SHS re-sit* (Versions 4 and 5 on the 17 and 25 May)	9602 (8333 for first time, 1269 second attempt)
Sat the first administration of the OC Test	6284
Sat the OC re-sit* (Versions 4 and 5 on the 18 and 24 May)	6550 (5174 for first time, 1376 second attempt)
Sat two tests	1269 SHS 1376 OC
Were eligible to re-sit OC and did NOT elect to	1429 (eligible to re-sit 2808)
Were eligible to re-sit SHS and did NOT elect to	1213 (eligible to re-sit 2482)

*Note: These re-sit figures only include complete test attempts

In summary, there were 15,573 students who sat the SHS Tests and 1269 students sat the re-sit and eventually received two scores. That is, 8% of the total number of students who sat the SHS Test eventually received two scores. In the case of the OC Tests, a total of 1376 (12%) of the 11,458 students who sat the test received two scores. The remainder of the students (92% and 88% of students) received only one test score. It is also worthy of note that 2482 and 2808 of the students in the SHS and OC Tests were eligible to re-sit the tests, but for one reason or another only 1213 (49%) and 1429 (51%) of these students chose to accept the opportunity to re-sit.

The next section considers the actions that were conducted to enable the re-sit of the SHS and OC Tests to take place on the nominated dates.

Actions to enable the re-sit of the SHS and OC Tests

A fourth version of the SHS and OC Tests was constructed as part of the original administration plan. This version was to be used as a “Make-up-Test” scheduled for administration on the 19th of May 2025. A fifth version of the test was also

constructed to accommodate the new circumstances and administered to students as part of the re-sit. This effectively meant that five versions of the SHS and OC Tests were linked to the respective scales rather than the four that were originally intended. This adjustment, given that the fifth version of the test was constructed and linked with the same care as the previous versions, would not adversely impact the fairness of the testing process.

The statistics summarising the functioning of the fifth version of the test support the argument that all versions of the test meet the requirements of the Rasch Measurement Model that underpins the placement test process. This means that students can take different versions of the test (i.e. sit different items) and the results are still comparable. No students were advantaged or disadvantaged from a psychometric point of view by sitting versions four and five of the SHS and OC Tests.

Technical difficulties during the re-sit

During the Mathematical Reasoning Test on 17 May, a technical issue occurred that impacted the test administration. Immediate remediation included providing students with additional time to compensate for the disruption, and the chief invigilator formally lodged an incident report. This report, alongside a psychometric analysis of test metadata - including time-stamps, response patterns, and system logs - served as the evidentiary basis for assessing the severity of the incident. These insights were used to guide decisions regarding illness and misadventure applications, ensuring that responses were fair, data-informed, and aligned with the principles of test equity and integrity.

Summary

On 2 May 2025, significant disruptions at the three largest test centres affected the administration of the SHS and OC Tests, leading to delays, crowd control issues, and heightened anxiety. To address the unpredictable impact on student performance, students at the affected centres were offered a re-sit opportunity, with their best score retained. Around half of the eligible students accepted the re-sit, and five test versions were ultimately used to ensure consistent scaling and comparability. Despite a technical issue during one re-sit session, I believe that remediation measures supported by sound psychometric analyses have upheld test integrity. The ultimate decision regarding the re-sit solution to the major disruption was guided by practical constraints and test theory, recognising that perceptions of fairness may still vary among stakeholders.

The next section of this report examines the fairness of the testing methodology used to provide adjustments where disruptions occurred during the original administrations of the tests. The final section addresses concerns that stakeholders have had with the decisions that have been made.

FAIRNESS OF TESTING METHODOLOGY USED FOR ADJUSTMENTS TO RESULTS BECAUSE OF DISRUPTIONS

Testing methodology

Version 4 of the SHS and OC Tests (referred to earlier as the “Make-up-Test”), used in the re-sit, was one of the original versions developed and subjected to the validation procedures that are typical of such high-stakes tests. These procedures drew upon multiple forms of evidence, consistent with those applied to the other versions.

Version 5 was constructed in a relatively short time frame using the same template for item and test construction as the earlier versions. A brief analysis of the statistical data used to show fit to the Rasch Model shows that versions 4 and 5 had the same consistent summary fit to the model as the earlier versions.

The same common items were used in Mathematical Reasoning and the Thinking Skills Tests to link versions 4 and 5 of the tests to the measurement scale as the other versions. In the case of reading, a different item stem with its associated options, was used for the linking. The case for this change is sound because of the higher potential of a single item being memorised from the previous administration of versions 1, 2 and 3 of the tests.

Students *retaking* the tests were excluded from the linking process used to link versions 4 and 5 to the measurement scale. This step helped to support equitable outcomes by avoiding any systematic bias that might favour the cohorts sitting versions 4 and 5. A comparison of the mean difficulties for common items across all versions suggests they are roughly equivalent, reinforcing the integrity of the linkage.

A different writing task was used for versions 4 and 5 of the SHS Test, but it was marked by human markers applying the same standards to common marking rubrics.

The overall conclusion is that the testing methodology used to adjust the results produced was fair using the traditional psychometric measures of fairness. However, in this particular case there were a number of issues raised by stakeholders that could potentially detract from the fairness of the process. These issues are addressed in the next section.

Advantage gained by students re-sitting the tests because of the common items used in the tests for equating

Concerns have emerged around possible advantages for those students who sat more than one administration of the SHS and OC Tests, due to familiarity with the common items used to link the versions of the tests. This potential advantage was mitigated by the original decision to exclude the results of these students on the common items across all test versions - ensuring no student gained a scoring benefit from prior exposure of the common items.

The test methodology used to adjust the results due to the disruptions accommodated this perceived issue associated with recognition of common items in

the tests and did not provide an advantage or disadvantage to students who sat the test once or twice.

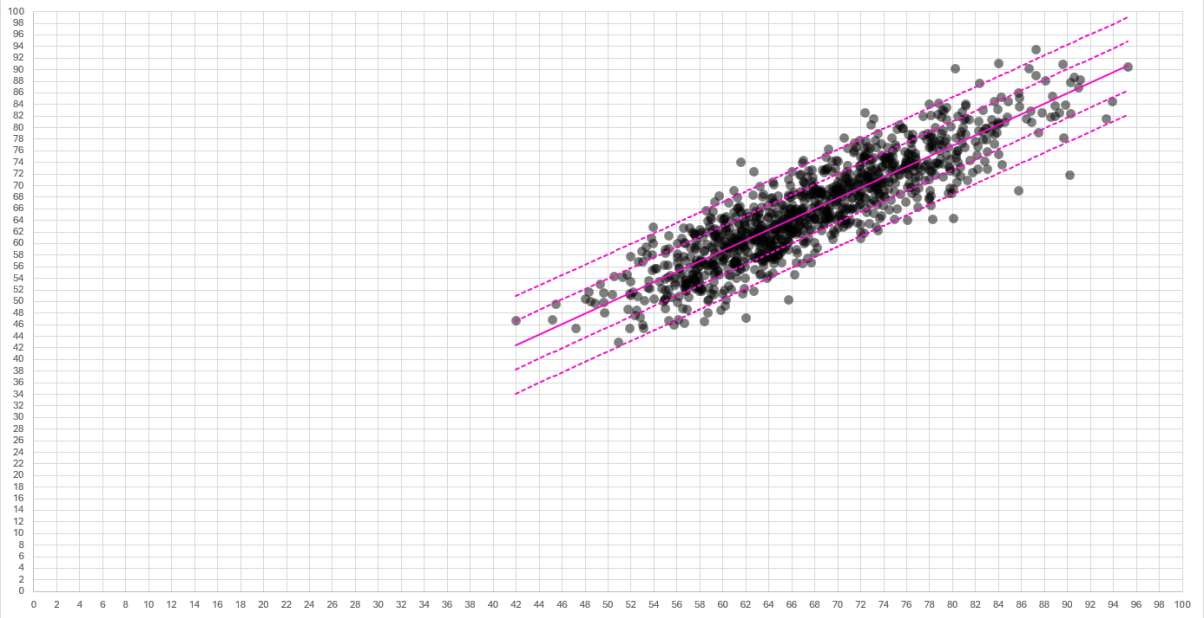
It is interesting to note that the mean percentage correct scores on the common items across versions 1/2/3, 4 and 5 for the SHS and OC Tests were 49%, 54% and 50%, and 47%, 51% and 53% respectively. The relative consistency of these scores suggests that there is no systematic bias introduced by retaining the same common items in the Mathematical Reasoning and Thinking Skills Tests. The results are consistent with what might be expected from random variation that occurs with every test administration.
These analyses add weight to the argument that the methodology functioned in accord with expectations and there was no evidence at a group level that there was systematic advantage to those students who resat the tests due to the inclusion of common items.

Advantage gained by students re-sitting the tests because of familiarity with the test and the testing experience

There were some concerns that students who had the opportunity to re-sit a version of the test (even though no student received the same test during the re-sit) may have been advantaged by their familiarity with the test and the testing experience. To assess the veracity of such claims, data were collected from the students with two test scores and a scatter plot constructed which shows the relationship between the scores.

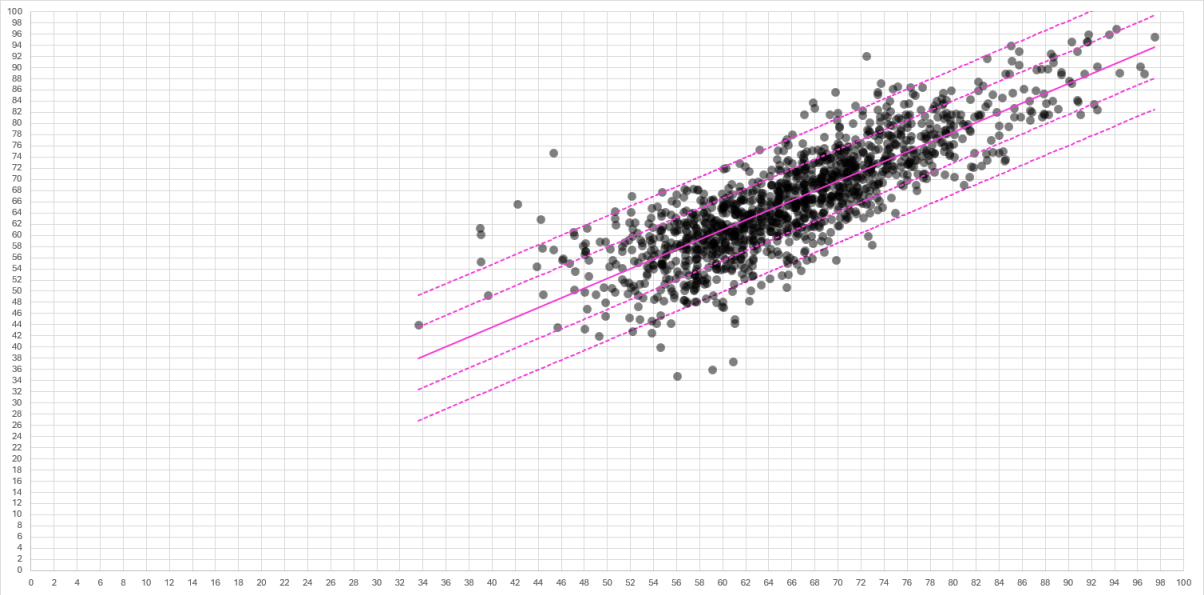
The mean scores of students who resat the first administration of the SHS Tests (versions 1/2/3) and then the second (versions 4 or 5) were 67.6 and 65.7. That is, students who resat, on average, scored worse on the second administration. In the OC Tests, the mean scores were 66.1 and 66.4. There was no perceptible difference between the scores (on average) of the same students on the two administrations.

Figures 1 and 2 show scatterplots of the scores of the 1269 students who sat two versions of the SHS Tests and the 1376 students who sat two versions of the OC Tests. The correlations between the two administrations of the SHS and OC Tests are relatively high, 0.88 and 0.83 respectively. This means that the majority of students performed approximately the same on both occasions, although as stated above, in the case of the SHS Tests the students who resat the test scored slightly worse. This does further suggest that there is no systematic advantage caused by the second test administration and also adds weight to the expectation that suggests that unless directed practice is provided on exposed content, it is unlikely familiarity will lead to an increase in score.



Note: The horizontal axis shows the scores on the first administration and the vertical axis shows the scores of the same students on the re-sit

Figure 1: Scatterplot showing the scores of students who resat the SHS Test



Note: The horizontal axis shows the scores on the first administration and the vertical axis shows the scores of the same students on the re-sit

Figure 2: Scatterplot showing the scores of students who resat the OC Test

The (outside) pink dotted lines on Figures 1 and 2 show what is called the 95% confidence bands of the results. The scores of students within these bands are consistent with what might be psychometrically expected, given the random variation that occurs from test-to-test.

It can also be seen that approximately 3% of the 1269 students who sat two versions of the SHS Tests and approximately 3% of the 1376 students who sat two versions of the OC Tests may have scored high enough on the re-sit to suggest that they may have benefited from the second administration. However, the majority of the SHS and OC test students have performed within measurement expectations. This further suggests that there is very little, if any, systematic advantage accorded to students who resat the tests because of the methodology used to adjust for the disruptions.

Advantage gained by students re-sitting the tests and having extended time to prepare for the second test

Some stakeholders have raised concerns that students who resat the test at a later date may have had an advantage due to additional preparation time. However, this perceived advantage is not necessarily a function of the adjustment solution used to accommodate disruptions in the testing process. Historically, “Make-up-Tests” have always been scheduled later - even in years without disruptions - and score comparisons between original and make-up sittings were never flagged as problematic. There has been no precedent of claims suggesting students gained an unfair edge from the short delay in test timing.

Disadvantage of students’ motivation being reduced after delaying the test administration date

Some stakeholders have argued that students were “primed” to perform at their peak during the original test sitting and were "over-their-peak" at the time of their later re-sit. Such claims overlook a fundamental principle of assessment: all measurements contain a degree of error. When students sit the same test on different dates within a short timeframe, minor variations in scores are expected. This outcome - known as **measurement error** - is directly tied to the **reliability** of the tests.

The SHS and OC tests demonstrate **high reliability**, indicating that score fluctuations across administrations are generally minimal. While minor individual differences may exist, they do not warrant broad adjustments or accommodations in a large-scale, high-stakes context. Predicting performance outcomes a priori, especially based on speculative assertions, undermines the integrity and fairness of standardised testing procedures.

Summary

While stakeholder concerns are acknowledged, empirical data suggest minimal, if any, systematic advantage for re-sitting students. The test construction and adjustment methodology have tended to support the argument that the methodology has produced relatively fair and reliable results in a high-stakes context which for one reason or another needed adjustments to be made because of issues that disrupted the administration of the tests.