An Roinn Oideachais agus Scileanna
Department of Education and Skills

Calculated Grades for Leaving Certificate 2020

Discussion paper for SEC-DES Technical Working Group on Calculated Results

Preface

This report was prepared as part of contingency planning for the Leaving Certificate examinations in 2020. The initial intention of the then Minister, Mr. Joe McHugh T.D. and Government, as announced on 10th April 2020, was that the examinations could not proceed in June 2020 and that they would take place in late July/August 2020. Contingency planning was also underway.

Conscious that there was a risk that it would not prove possible to hold the Leaving Certificate examinations as planned, the Minister also asked that a Technical Working Group be established, involving experts from the State Examinations Commission, the Educational Research Centre, the Department of Education and Skills and independent statistical and psychometric expertise, to examine the feasibility of establishing a system of Calculated Grades for Leaving Certificate students. At the end of April, that group reported to him advising that a system of Calculated Grades could be put in place as an alternative to running the conventional examinations. The Minister considered that advice alongside the advice of senior officials of his own department and of the Department of Health when forming his views for the postponement of Leaving Certificate Examinations 2020 (including the Leaving Certificate Vocational Programme and the Leaving Certificate Applied Examinations) and the introduction of an optional system of Calculated Grades, which led to a Government decision to that effect.

This discussion paper was prepared as part of the work of the Technical Working Group. The group considered the paper at a meeting on 24 April and it formed an important part of the basis of the advice of the Working Group to the Minister.

Although this preface has been added at time of publication, readers should nonetheless note that the discussion paper itself was completed before the 24 April and, apart from minor corrections and clarifications, is presented here as it stood at that time. In particular it should be noted that:

- At the time of this paper's preparation, it remained the Minister's intention to hold the Leaving Certificate examination in late July and August. Accordingly, the Technical Working Group considered two possible scenarios: that it might not be possible for the examinations to proceed as scheduled, (leading to a need for calculated grades for all students,) or that the examinations might proceed but with large numbers of students being unable to attend (leading to a need for calculated grades for some but not all students). Both of these scenarios are explored in this discussion paper.
- The paper pre-dates the receipt of the advice of the Attorney General to the effect that there was no legal basis for the State Examinations Commission to be involved in the running of the calculated grades scheme, and that it would therefore need to be operated under the Minister's executive powers through the establishment of a non-statutory executive office in the Department of Education and Skills. In line with the understanding at the time, the paper refers to proposed actions that were

envisaged to be carried out by the State Examinations Commission but which subsequently were instead the responsibility of Calculated Grades Executive Office in the Department.

- Much of the detail of the proposals as described in this discussion paper were subsequently refined or amended as a result of further deliberations and engagement with stakeholders after the announcement was made to postpone the examinations and to introduce the optional system of calculated grades.

I would like to thank the members of the Technical Working Group for their thoughtful and incisive consideration of the issues involved in preparing their advice to the Minister. I commend them in particular for the work involved in preparing a background paper of such high quality in the short time that was available to them, which facilitated the framing of that advice on a rigorous and solid foundation.

Seán Ó Foghlú
Secretary General
Department of Education and Skills

Discussion paper for SEC-DES Technical Working Group on Calculated Results

1 Background and Rationale

As a result of the COVID-19 pandemic, the Leaving Certificate Examination cannot proceed as scheduled in June 2020. While every effort is being made to ensure that a set of examinations that is as normal as possible can proceed later in the summer, it has been agreed that it would be prudent to explore possible alternative ways of issuing fair and valid Leaving Certificates, including those for the Leaving Certificate Applied, if such examinations cannot proceed. In particular, the possibility of issuing results based on alternative forms of evidence, incorporating estimates from teachers and/or data that already exists in schools, is to be explored.

Additionally, even if such examinations do proceed later in the summer, it is inevitable that some candidates will be unable to attend because of COVID-19. The potential use of a similar methodology to provide certification for these candidates is also to be explored.

The Department of Education and Skills (DES) and the State Examinations Commission (SEC) established a joint technical working group tasked with carrying out scoping and preliminary work in anticipation of one or other of the above two scenarios arising. The membership of the Technical Working Group expanded over the course of its work. The following were members for some or all of its duration:

- Dr Tim Desmond, Head of Examinations and Assessment, SEC (chair)
- Dr Harold Hislop, Chief Inspector, DES
- Aidan Farrell, Chief Executive Officer, SEC
- Andrea Feeney, Director of Operations and IT, SEC
- Hugh McManus, Assistant Head of Examinations and Assessment, SEC
- Elaine Sheridan, Assistant Head of Examinations and Assessment, SEC
- Dr David Millar, Examinations and Assessment Manager (Research and Development), SEC
- Orlaith O'Connor, Assistant Chief Inspector, DES
- Jason Kelly, Senior Inspector, DES
- Dr Jude Cosgrove, Chief Executive Officer, Educational Research Centre (ERC)
- Fernando Cartwright, Principal at Polymetrika, Inc., Ottawa, Ontario, Canada (external consultant)

Following initial discussions and ancillary work exploring objectives, options, international comparators, and design principles, the group agreed that it would be valuable to prepare a discussion paper on possible approaches, fleshing them out in sufficient detail to allow proper consideration of the issues and implications that arise, so as to support sound decision-making. A subgroup was tasked with preparing this paper for consideration by the full group.

Two different scenarios are described in the first two paragraphs above that may give rise to a need to produce certification based on evidence other than results from an examination sat and marked in the normal way. While the task at hand might appear at first glance to be similar in both cases, it should be noted that the two scenarios yield different technical problems that may require significantly different solutions.

Among the most notable differences between the two scenarios that affect the premises of the arguments underpinning the selection of procedures and models are as follows:

If no examination is sat:	**If an examination is sat and we are generating estimates for missing candidates only:**
• All candidates are treated in the same way by whatever process is agreed	• Two groups of candidates will have followed two very different routes to certification, which places a significant additional comparability burden on the processes employed
• The results to be estimated are the results that candidates would have been expected to achieve in the June examinations if the pandemic had never arisen	• It seems most likely that the results from the late-summer sits will stand as results in their own right, rather than being treated as estimates for further statistical treatment. If so, then the results to be estimated for the missing candidates are the results they would be expected to have achieved in the late-summer form of the examination under comparable conditions to those who actually sat them, which is a more difficult estimation problem
• No actual live examination performance information is available to combine with teacher judgment data when estimating results	• Both estimated performance data (on one examination form taken under one set of conditions) and actual performance data (on another examination form taken under different conditions) will exist for a subset of candidates. This would be highly relevant bivariate data in estimating the likely performance of missing candidates, so it could not credibly be ignored

While it is conceivable that a single sufficiently general and adaptable model could deal with both scenarios, this should not be assumed *a priori*, and the above differences are of sufficient significance that it is appropriate to treat the two scenarios separately in this paper.

2 Teacher Estimates or Pre-existing School-based Data

There is a wide range of information and evidence available in schools that could be used for the purpose of generating an estimated mark. The question arises as to whether data that is already available in school administration systems and teacher records, which is likely to include results of mock examinations, results of Christmas and summer tests, and various other assessment outcomes, could be harvested from schools and fed directly into a statistical estimation algorithm, without any further requirement for teachers to mediate these to form predictions based on judgment. The argument might be made that, rather than being 'mere opinion', this is 'hard data', that the more data that is collected from schools the more accurate the outcome will be, and that a properly designed and universally applied algorithm can process this data more fairly and effectively than individual teachers.

However, these arguments fail to take account of the degree to which the quality, contexts, and circumstances surrounding the generation of these data varies dramatically from school to school, that these tests will have varied widely in their content, structure and style, that different teachers will have marked them in different ways – some considering, for example, that mock exams should be marked harshly to motivate students, while others might mark them more leniently to encourage

students who lack confidence, and so on. Furthermore, the degree to which different schools systematically capture and record such data varies considerably. We consider in these circumstances that the arguments above in favour of the direct use of such data give a false comfort. Used raw, these data are not reliable. To be properly interpreted in terms of what they say about individual student achievement, it is critical that they be mediated by the informed and considered professional judgment of teachers, who know their own students and are familiar with their work, who understand the nature, content and contexts of the school-based tests that have generated these data, and who have experience in interpreting and reporting on them to students and parents.

Teachers regularly use both formative and summative assessment of students' work as part of teaching and learning in their classrooms. Additionally, teachers spend two years preparing students for these high-stakes examinations and are familiar with the format and demands of the examinations as well as the quality of work of their students. In the absence of a Leaving Certificate examination, teachers are the ones best placed to collate and interpret all available evidence in relation to the expected performance of their students. Teachers know their students and are able to balance a variety of evidence in arriving at a holistic professional judgement in relation to each student's expected performance.

Furthermore, the use of teacher professional judgment facilitates incorporating evidence from evaluating partly or wholly completed coursework that has been carried out as part of the Leaving Certificate examination and which has not to date fed into any datasets already held by the school.

3 Broad Approaches to Using Teacher Estimates

In circumstances where estimated or predictive information from teachers about likely candidate performance is to form a dominant or important part of the evidence base on which grades are to be awarded, mechanisms are clearly required in order to address likely differences between teachers and schools in the degree to which they may be severe, lenient or otherwise inaccurate in their estimates. The details of the issues that arise and possible mitigation measures are teased out in more detail in Sections 5 to 7, but it is worth initially drawing attention to some broad questions in relation to how one sets about this task.

It may be considered that there are two different approaches:

1. approaches that entail moderating, by statistical or other means, teacher estimates in a manner that seeks to align teacher judgmental standards with each other but strictly constrains the moderation process to maintain rank-order aspects of teacher judgment intact.
2. approaches that allow the statistical moderation process to adjust the degree to which the final estimates are constrained by the teacher judgments. The results from this approach may be optimal from a probabilistic perspective but may not preserve the rank order assigned by teachers.

3.1 Moderation Approaches

If a moderation model is used, the primary form of evidence is the estimate produced by the school. The moderation process seeks only to align the standards across teachers and schools to each other and potentially to a national standard reflected in established grade distributions. Ideally, one seeks to moderate robustly across teachers within each school in addition to across schools. However, the requisite data to statistically moderate properly across teachers within a school may not be available in the current case, so *procedural* forms of within-school moderation may need to be combined with

statistical forms of cross-school moderation. The two different potential scenarios under consideration would almost certainly require two different moderation models (the full-cohort scenario: no examinations occur and estimates are required for all candidates, and the partial-cohort scenario: examinations are taken in August but substantial numbers of candidates are unable to sit and their likely results are to be imputed.)

One possible moderation model for scenario A is the one currently proposed for A-level results in England by Ofqual. Information about this approach is available at: https://www.gov.uk/government/organisations/ofqual.

Within the set of potential moderation models, further comparatively broad-brush decisions are required. For instance, does one automatically adjust all school distributions with each other to align the judgments as accurately as the model allows, or does one allow a tolerance level below which no adjustments are made? Another question that arises in the case of some potential models is whether there is a minimum number of students in the group concerned required in order that we can safely apply the moderation process, and the related question of what to do if this criterion is not met.

3.2 Estimation Approaches

Typically, the data are fed into a statistical estimation model that seeks to optimise the information from all data sources to maximise the statistical accuracy of each estimate. This approach allows for arbitrary assignment or statistical optimization of the relative importance ascribed to the judgments of the teacher or school in the estimation process. Although it remains an option, no specific weighting need be ascribed to any particular form of data. By default, it is the observed interactions across all the data in the data set that determine how each datum affects each estimated score.

With this type of approach, it is important to distinguish between data that are used to estimate scores, data that are used to scale scores, and data that are used to validate scores. Data that are used for estimating scores must be characterized by dependence on individual student attainment, such as observed student performance or teacher estimates of student performance. Data that are used for scaling scores should describe the frame of reference for which the estimates are valid, such as subject, class and school membership of each student. Data that are used for validation may be more broad, but they should have an historical and known relationship to the distribution of test scores, such as national means and standard deviations, and long-run average distributions of examination results in schools or different demographic groups.

Any model, whether it represents an estimation or moderation approach, will need to take account of alignment of standards across levels, especially since 'entry' and 'sit' patterns are not identical to each other in a typical year, and the 'confirmed intended level' that will be captured as the best estimate of the likely 'sit' level is likely to result in candidates remaining at Higher level in the calculated results process who would probably have ultimately taken Ordinary level.

4 Psychometric Framework

A description of a psychometric framework within which the problem at hand may be situated is given in Appendix A. In summary, its implications for the details of the selection of a model and associated procedures, as they would apply in the case of the 2020 Leaving Certificate Examinations, are as follows.

Conceptually, all sources of information about a student may be used to estimate test scores. The most defensible are those based directly on student behaviours and characteristics in the domains measured in the Leaving Certification examinations. Test data, even if they are biased or incomplete representations, are the strongest form of evidence because they are direct estimates. There are many indirect sources of student-level data relevant to estimating examination performance to replace the missing examination data, but the only data universally available are teacher judgments about student proficiency.

However, teacher data tend to be nested within a local frame of reference that makes standardized interpretations difficult. Therefore, the estimation process must use additional sources of information to correct for the localized bias of teacher judgments. The likelihood-based approach to scoring provides a consistent approach to using all relevant information while minimizing dependency on arbitrary assumptions about how they contribute to the estimation process.

5 Summary of Relevant Research on the Use of Teacher Judgment

While this section summarises research relevant in this field, it should be born in mind that the context of much of this research will have been quite different from the one at hand, so caution needs to be exercised in relation to the degree to which conclusions may be transferrable to the present context. Furthermore, while the research might identify certain features of a model or procedure as being desirable, such features may not be practicable to implement in the present context. For completeness, the research summary below is provided without prejudice to these observations.

There are four main branches of relevant research on using teacher judgment to replace standardized test results[1]:

1. estimating the accuracy of teacher judgments
2. factors affecting the accuracy of teacher judgments
3. how to increase the validity of teacher judgments
4. how to increase the reliability of teacher judgments

5.1 Accuracy of Teachers' Judgments

Accuracy has two facets: validity and reliability. Validity, in this case, is the degree to which the teachers' judgements predict standardized test results. Reliability is the degree to which the judgment of teachers produces consistent results: do different teachers provide similar estimates for the same students and are the ratings of each teacher consistent for students with the same objective level of performance.

Typically, validity has been easier to estimate, because the criterion for validity (agreement with tests) is directly observable. With few exceptions[2,3,4], existing research has consistently found moderate to high correlations between teacher judgments and test results[5,67,8,9,10,11]. However, high correlations inflate the perception of accuracy because they describe consistency in relative ranking, not agreement in absolute level[12,13]. Because teacher judgments tend to over-estimate student performance, and the test results are (often) interpreted against absolute standards, this bias reduces the validity of the raw, unadjusted estimates.

Teacher judgments also use local frames of reference[14,15,16,17]: if the judgments are criterion-referenced, they tend to refer to school, classroom, or even student-specific expectations of

performance; if the judgments are norm-referenced, they tend to reference school or classroom-specific distributions. However, the more familiar the teachers are with their given frame of reference, the more reliable their estimates are: within-classroom (or within-teacher) estimates are extremely reliable[18,19,20], whereas the reliability of estimates between classrooms is dependent on the degree of similarity between the reference frames used to make the judgments[21,22,23,24].

5.2 Factors Affecting Accuracy of Teacher Judgments

There are three class of factors affecting teacher judgments:

1. the type of judgment
2. teacher characteristics
3. student characteristics

5.2.1 Types of Teacher Judgments

There are 5 types of teacher judgments with respect to degree of specificity[25]:

1. ratings – qualitative ordinal statements, typically selected from predefined descriptors (e.g., "Excellent, very good, …")
2. rankings – relative position in a population of students' peers (e.g., "first, second, …, last") or position relative to a norm-referenced distribution (e.g., "well above average")
3. grades (ordinal scores) – standards-based interpretations of performance corresponding to a predefined, publicly-known grading system
4. interval or ratio summary scores – holistic estimates of scores on a specific assessment framework, performance rubric, or test form
5. item-level scores – estimates of item level performance or probability of item-level performance on a specific set of items.

Note that evidence from any of the latter (i.e., more specific) types of judgments may be processed to a less-specific form of judgment with little loss of accuracy, but the reverse is not true.

There are two types of teacher judgments with respect to student test performance[26]:

1. direct estimates – teachers estimate how students perform on a specific test form, and
2. indirect estimates – teachers estimate performance on the same construct that the test claims to measure.

Direct estimates with greater specificity tend to be more accurate, both in terms of absolute error and correlation. As in other contexts, treating judgment data incorrectly reduces its accuracy (e.g., increasing the number of ratings categories produces a pseudo-interval scale but shrinks the classification stability[27,28], and binning an interval-level set of scores produces ordinal categories and shrinks correlations[29]).

5.2.2 Teacher Characteristics

Teacher judgments tend to be more accurate when the teachers have more teaching experience, both in terms of teaching the subject matter and teaching the specific students for whom they are providing estimates[30]. Across a wide range of high and low stakes assessment, teachers tend to overestimate their students' test performance[31,32,33].

Across all age ranges and subject areas, teacher judgments are consistently more accurate when the teachers are better-informed about the test for which they are providing estimates, including

content as well as the underlying assessment framework, assessment purpose, and testing conditions[34,35,36,37,38].

Teacher judgments are also more accurate when the teachers believe their judgments are important, either because the use of the results is high-stakes, or they are being used to reconcile other inconsistent information sources[39,40].

5.2.3 Student Characteristics

Teacher judgments tend to be more accurate for higher-performing students than lower-performing students[41,42,43,44]. This bias is often interpreted as an artefact of testing, not of teacher judgment; higher performing students will converge on a single correct response, whereas low-performing students may produce a wider array of incorrect responses.

Teacher judgments tend to magnify existing disparity, by underestimating performance of low-performing students and over-estimating performance of high-performing students[45]. This estimation bias is also influenced by social identification associated with low performance; teachers tend to underestimate performance of students with special needs identification or from historically disadvantaged social-demographic groups[46,47,48]. At the same time, judgements also tend to be more accurate when they are informed by historical evidence of student performance in the classroom and other assessments.

Student characteristics that are correlated with performance, such as classroom behaviours, engagement with teachers, and performance in other subjects, influence teachers' estimates of performance independently of their actual performance in the subject being estimated.

5.3 How to Increase Accuracy of Teacher Judgments

As noted previously, there are two distinct methodologies that are associated with more reliable teacher judgments:

1. Increasing specificity
2. Formalizing comparative judgments

5.3.1 Increasing Specificity

As with validity, the more informed teachers are about the reference frame (i.e., the test and the students), the more stable their estimates will be[49]. Without adequate guidance, teachers are influenced by classroom behaviours, student-teacher relationship, performance in other subjects, and performance relative to other students rather than solely by the relationship between students and test content[50,51].

5.3.2 Comparative Judgments

It is difficult for teachers to make judgments that reference absolute standards. Teachers tend to make judgments about students by comparing different students; as a result, within-class student rankings tend to have lower overall correlations with test performance than test-based predictions[52,53].

However, many studies have found that the stability of rank-based judgments, particularly pairwise comparisons between students, tend to be extremely reliable – even more so than score estimates or ratings[54,55,56,57,58]. Unfortunately, this increased reliability maximizes validity within the frame of reference (i.e., the body of students being directly compared), but not across a larger population. These findings suggest that an ideal teacher-based estimation process would benefit from both the specificity of test-oriented numeric scores and student-oriented rank-ordering.

5.4 Implications for the 2020 Leaving Certification

The findings suggest some strategies that are likely to produce the most accurate results:

1. judgments should be estimates of examination performance, rather than estimates of school or curriculum performance.
2. data collection should use concrete formats that are specific to test form content – ideally item-based predictions or questionnaire rating scales – with which teachers are familiar rather than holistic judgments of general subject domain achievement.
3. guidance should discourage teachers rating a single student across multiple subjects (instead, if they teach multiple subjects, complete each subject independently).
4. guidance should inform teachers about how their estimates may be used to estimate student performance (if this is not exactly known at the time of data collection, the underlying operating principles should be clear).
5. guidance should discourage consideration of classroom behaviour and social or cognitive disadvantage.
6. although ranking is useful within classrooms, teachers should be strongly discouraged from ranking students against populations with which they are less familiar (i.e., unless all teachers have the same degree of familiarity with the national distribution, they should focus on their immediate reference frames: test and classroom).
7. guidance should inform teachers of which forms of evidence may be used to inform their judgment.

Due to operational constraints, not all these strategies may be realistic; however, the validity of the results will increase to the degree that they are realized.

In addition to the previous research specific to this issue, other developments in statistics and psychometrics that were either not available at the time of or (for whatever reason) were not applied in primary research can remedy many of the limitations noted by the original authors. For example, estimation techniques for multilevel data can estimate and correct for bias that results from shifting and nested frames of reference, and likelihood-based estimation for latent data can efficiently combine data from multiple sources while avoiding conflicts in statistical assumptions.

6 Relevant Data Sources – Full-cohort Scenario

There are many sources of data potentially relevant to the estimation of student examination performance. To be incorporated into the estimation or moderation procedure, data must be linkable to a common unit of analysis, such as student, school, or teacher, and a reference time period. The data sources recommended for this procedure are listed in Table 1. The data sources are listed in order of expected usefulness with respect to estimating individual student examination results, with teacher estimates being the most relevant. This order also implies a hierarchy of credibility; in the event that evidence from less-relevant sources disagrees with evidence from more-relevant sources, higher credence is given to the more-relevant sources.

Table 1 Data sources

Data Type	2020 Grads*	Pre-2020 Grads*	Intended Use**
1. Teacher estimates of student performance	Y		S
2. Teacher/Class membership	Y		C
3. School membership	Y	Y	C
4. Junior Certificate Examination results	Y	Y	C
5. Student 2-or-3-year programme status	Y	Y	C
6. Student demographic characteristics	Y	Y	V
7. Full LCE results		Y	C

*"Y" indicates data are linked at the student level to globally unique student identifiers

**How the data will be used in the procedure:

S=estimating specific student scores

C=estimating conditioning distributions

V=validation (not directly contributing to estimates)

Some data sources will be linked to students in the 2020 graduating class, some to previous graduating classes, and some to both. Only one data source will be used to differentiate between individual students within classes in the event that no examinations transact: teacher estimates. All other data sources will be used to estimate conditioning distributions at class level and higher or used in a validation process to evaluate the credibility of the overall set of estimates. Implicitly, conditioning variables also provide validating information, but some data marked explicitly for validation will not be used for estimation of scores, because they risk producing biased results for individual students.

Although teachers have access to rich information about students, in-class formative assessment and mock examination results are insufficiently standardized in content or accuracy to provide a fair comparison of students. However, these data sources are valid sources of evidence to inform the teacher estimates of examination performance.

Junior Certificate exam results are strong predictors of Leaving Certificate performance but are inadequate by themselves to estimate individual student performance; they are not credible representations of second-level academic performance, and data are missing for some students. However, because of their near-universal coverage, they provide a useful means of determining the objective performance distributions of classes and schools. The linkage between Junior Certificate and Leaving Certificate examination results in previous years facilitates the calibration of the Junior Certificate results for the purpose of providing conditioning information for the current cohort of 2020 candidates. Conceptually, the logic of this calibration is:

1. Use the Junior Certificate data linked to students who graduated in previous years to determine the conditional distributions of Leaving Certificate performance that correspond to different levels of Junior Certificate performance.
2. Use Junior Certificate performance data linked to the 2020 candidates to estimate the conditional likelihood of Leaving Certificate performance for each 2020 candidate with valid data.
3. Use these conditional likelihood functions to estimate the conditioning distributions for Leaving Certificate examination results at the class and school levels for the 2020 candidates.

Note that this procedure only uses student-level data for the purpose of estimating information at the class and school levels.

Similarly, although programme length has a relationship to opportunity-to-learn and other achievement-related factors, this relationship occurs at the system/macro level rather than at the individual student level. Therefore, it may be used to define conditioning distributions at a macro level, as individual teachers with classes that are homogeneous with respect to programme length may not be able to consider the effects of program length on their students relative to students in other classes. The linkage between programme status and Leaving Certificate results in previous years facilitates the calibration of programme length in the same manner as the Junior Certificate results. The conditioning information based on programme length is used at an aggregate, rather than individual, level.

7 Methodological Alternatives

The following sections suggest several methodological alternatives. Some alternatives are not mutually exclusive and may be combined for greater accuracy. Additional methodological alternatives that might arise in the partial-cohort scenario are dealt with in Section 10.

The alternatives fall within the same broad strategy:

1. Use conditioning information to estimate and correct for the bias in the teacher-sourced information
2. Use validating information to evaluate the credibility of the estimated results and revise the estimation procedures, if required. Credible results will produce macro-level distributions of performance that are within the ranges typically observed in previous years.

7.1 Teacher-sourced Information

For the purposes of this subsection, all data collection assumes that data from teachers will use a secure web application that require multi-factor authentication using numeric school and/or teacher identifiers. All methods retain data linkages that allow teachers to use multiple sessions and correct previous data entries. All methods assume that teachers have access to and are familiar with the cognitive requirements of previous years' Leaving Certification Examination test forms. Consideration is also being given to how the preliminary work that teachers will be required to engage in – including the collation and consideration of school-based data along with other information in order to form their professional judgments – is to be structured and recorded (see Section 9).

1. Item based performance - Teacher estimate the probability (or proportion-of-total) score of each student on each test item.
2. Aggregate test subsection performance – Teachers estimate the percent-of-total score for cognitively-distinct subsections of the test (corresponding to meaningfully distinct element of the test blueprint).
3. Aggregate test overall performance – Teachers estimate the percent-of-total score for the entire test. Data are collected using an interface that requires one mouse-click per student.
4. Student relative ranking – Teachers rank students within their class relative to their expected overall test performance.
5. Judgment rationale – Teachers write in qualitative information rationalizing their data input (referencing specific sources of evidence).

6. Class calibration – Teachers estimate the range of their class distribution against the range of possible test scores.

Each of these alternatives has trade-offs. The choice of which method(s) to use should consider the relative costs and benefits, summarized in Table 2. Each option is characterized in terms of expected accuracy, transparency, amount of time expected for data collection per each student/subject combination, operational risk, and whether the data are essential to estimating valid results. Accuracy of each option is relative to the baseline of aggregate estimation of overall performance and is estimated based on review of existing literature. Transparency describes the degree to which the estimates are verifiable and provide a means of estimating data quality. Collection time is based on typical times required for holistic judgment in human marking and assume that each test has approximately 130 items and 5 distinct subsections and class sizes average 20 students. Operational risk describes the chance that the complexity of data collection operations will either discourage participation and engagement from teachers (and other key stakeholders) or result in incorrect or inconsistent data entry.

The rating categories are listed in order of increasing importance: because the viability of any procedure is contingent on acquiring the necessary data, Operational Risk should be considered with similar weight as whether the data are essential to estimation.

Table 2 Teacher-sourced information options costs and benefits

	Accuracy	**Transparency**	**Collection time** (minutes)**	**Operational risk**	**Essential**
Item based performance	Higher	High	30	High	Yes*
Aggregate test subsection performance	Higher	High	10	Medium	
Aggregate test overall performance	na	Low	5	Low	
Student relative ranking	Lower	Low	0.5	Low	
Judgment rationale***	na	High	tbd	Medium	No
Class calibration	na	Medium	tbd	Medium	No

*At least one of these options is required.

** Collection times are for data entry only and are exclusive of time spent collating school-based data, reviewing student work, and considering other information in order to form professional judgments.

***Collection times and level of detail of this information will be determined by the protocols and guidance provided to teachers.

Note that options 1, 2, and 3 are equivalent measures; if one of these is selected, the others are redundant. Option 1 and 2 are relatively similar in terms of the information and accuracy they provide. Both provide a means of describing internal consistency of data and estimating a student-level likelihood function; in contrast, option 3 does not. Without student-level likelihood functions, the estimation process must treat the teacher estimates as error-free within the class context (only applying conditioning at the class level) or arbitrarily assign a degree of imprecision to the estimates so that they can be used consistently with other data.

Option 4 (student relative ranking) is sufficiently distinct from the first three options that including it with any of the first three should increase the accuracy of the student level estimates. Research

indicates that teachers are more reliable in paired rank ordering of students in their classrooms than in estimating numeric scores, (see Section 5.3.2 above). Without any of the first three options, student relative ranking is likely to provide less accurate estimates than holistic estimates of overall test performance, because the test scores are expressed on an objective interval scale, but ranking is only ordinal within the context of each class. Test scores have the same meaning across classes, and differences between non-adjacent test scores have consistent interpretations. However, rankings do not have the same meaning across classes; non-adjacent ranks cannot be meaningfully compared, and they are more susceptible to "halo" effect judgments.

7.2 Conditioning Data

Conditioning data include variables that uniquely identify classes, teachers or schools, or nest students in groups larger than classes. The only conditioning variable considered at this time other than the group identifiers is whether students are in 2-year or 3-year programs. This variable is in student-level data already available to SEC.

Operationally, linking students with classes and teachers adds some operational complexity, but arguably, since the data must originate in the context of a specific class and must be provided by a specific teacher, the operational challenge relates to data capture rather than data collection. Under the assumption that all data collection is performed using a secure online interface, these linkages may be established implicitly during the process of data collection without requiring explicit data entry from teachers.

7.3 Limitations of Available Conditioning Data

It may be noted that only the current year entry data and any data gathered in respect of the 2020 cohort as part of the process currently under consideration can be linked to teachers as distinct from schools. All data from previous years, whether Leaving Certificate or Junior Certificate, is only linked to candidates and schools, not teachers. The existing examinations data can be used to generate two types of conditioning distributions:

1. national or school level distributions based on the historical distributions of Leaving Certificate results, or
2. school, teacher, or class level distributions based on the prior Junior Certificate examination results of the current 2020 candidates.

Both types of distributions are limited in their sensitivity to different factors. The historical distributions are valid to the extent that the current cohorts of students at each level are randomly equivalent to students in previous years, an assumption that may hold to a lesser extent for smaller schools. The distributions based on the Junior Certificate results are valid to the extent that the Junior Certificate results are predictive of Leaving Certificate results, which may be less applicable if students in different classes experienced systematic differences in their educational experiences in the years since the Junior Certificate examination. These limitations imply that actual distributions of student performance could deviate from the historical or predicted distributions.

To ensure that the data collection accurately records meaningful deviations and the conditioning process does not subsequently remove them, the data collection should include non-statistical moderation procedures to align teacher judgments of absolute (as distinct from relative) performance, and, where distributions of teachers' estimates deviate substantially from historical or predicted results, require confirmation from school management (described in section 9.1) to provide corroborating evidence that the deviation is warranted to reflect actual student outcomes.

In any event, conditioning distributions at the sub-national level should not impose strict constraints on the calculation of student scores; rather, they should adjust and refine the more-specific evidence from teacher judgments.

7.4 Validation Data

SEC routinely publishes examination results at different levels of aggregation. Over the life cycle of the examination program, norms have emerged that are used by SEC internally and the public at large to judge the credibility of results. The norms may relate to equity issues, such as relative performance of gender groups, or to fairness, such as the relative performance of students in one year compared to students in previous years. If the results of an examination in a given year were to violate these established norms, stakeholders would question the credibility of results.

In order to ensure that the estimation process produces credible results, the estimation procedure may apply principles of reinforcement learning[59] to constrain the estimation process. The distributions that characterise established norms define a set of macro-level likelihood functions that describe the likelihood that any given set of results will be interpreted as a valid set of Leaving Certificate examination results. Without explicitly contributing to the estimated results, iterative evaluation of sets of estimates against the validation data may be used to modify estimation parameters to ensure that the results produced have the greatest possible credibility given the input data.

7.5 Score Estimation

There are two appropriate methods of estimating scores, depending on the constraints imposed by the available data. The first method of score estimation applies if the teacher-sourced data is considered error-free at the within-class level and is produced using non-linear rescaling. The second method of estimation applies when the teacher-sourced data is combined with other data using the likelihood-based approach.

7.5.1 Nonlinear Rescaling

When teacher-sourced data are considered error-free at the within-class level, the conditioning data is used to estimate the shape and location of the population-referenced distribution of performance for each class. This estimation attempts to find the distributional parameters of each class such that the aggregation of the distributions of all classes are consistent with both the predicted distributions based on conditioning data and the expected distributions based on validation data.

The teacher-sourced data map individual students to a teacher-referenced interval-level distribution. Optionally, the shape (variance, skewness, kurtosis) of the teacher-referenced distribution may be used to inform the shape of the population-referenced distribution either by applying them to or blending them with existing shape of the population-referenced distribution.

The individual score estimates from the teacher-referenced distribution are mapped to the population referenced distribution such that the percentile rank of each student within each class is preserved. This transformation ensures that the absolute distances between scores are consistent with the properties of the expected distributional parameters of the Leaving Certificate examination results, but the relative distances and rank ordering of students are consistent with the teacher-sourced data.

7.5.2 Likelihood-based Estimates

When teacher-sourced data are not considered error-free, each student level estimate is represented by a continuous function describing the likelihood of each possible score value. These student-level estimates are rescaled following the same criterion as the adjustment of class-level distributions described for nonlinear rescaling: the aggregation of the student level distributions should produce class, school and population distributions that optimally match both the predicted distributions based on conditioning data and the expected distributions based on validation data.

To produce final student estimates for students, the student-level functions based on teacher-sourced data are multiplied successively, following Bayes Theorem, with each of the additional distributions in which they are nested to produce a final full-information posterior density function, from which there are three possibilities for estimating scores:

1. Maximum - the most likely score to have resulted in the collected data,
2. Mean - the score that is associated with the smallest estimation error, or
3. Biased - the highest estimate that is consistent with a tolerable degree of error (this is a biased estimate that is higher than the other estimates, where the bias is greater for students with greater uncertainty in their posterior density functions).

The degree to which the rank-ordering of teacher-sourced data is preserved may be controlled by parameters of the estimation process to allow for a range of possible constraints. Initial rank order may be strictly preserved, used as conditioning information to minimize changes in rank order of final estimates, or used to inform the initial student-level likelihood functions but impose no subsequent constraints on rank order.

7.5.3 Final Score Estimates for Reporting

Scores may also be produced by averaging score from more than one type of estimator. Regardless of the type of estimate, the final step of the process adjusts the collection of score estimates to match the distributional properties of the validation data and produce the final reported values.

8 Selecting a Practicable and Defensible Model and Procedure

The data-capture and statistical processing procedures described above are sufficiently general to allow a range of decisions to be made later without undermining the capacity of the model to be applied. Nevertheless, at some point some significant decisions need to be made surrounding what forms of data can reasonably be captured in the current circumstances, in order to strike the right balance between the practicalities of generating and collecting the data and the utility of the data collected.

It is now proposed, in light of all of the above arguments and issues, that the model selected should be based on the following premises:

Premise 1: when balanced against practicability and operational risk, maximum utility in this context is achieved by collecting an estimated percentage mark for each student in each class, along with a strict rank ordering of the students in the class.

Premise 2: prior attainment data at student level should be used only in aggregate form to inform conditioning distributions, and not to affect the individual student's calculated result.

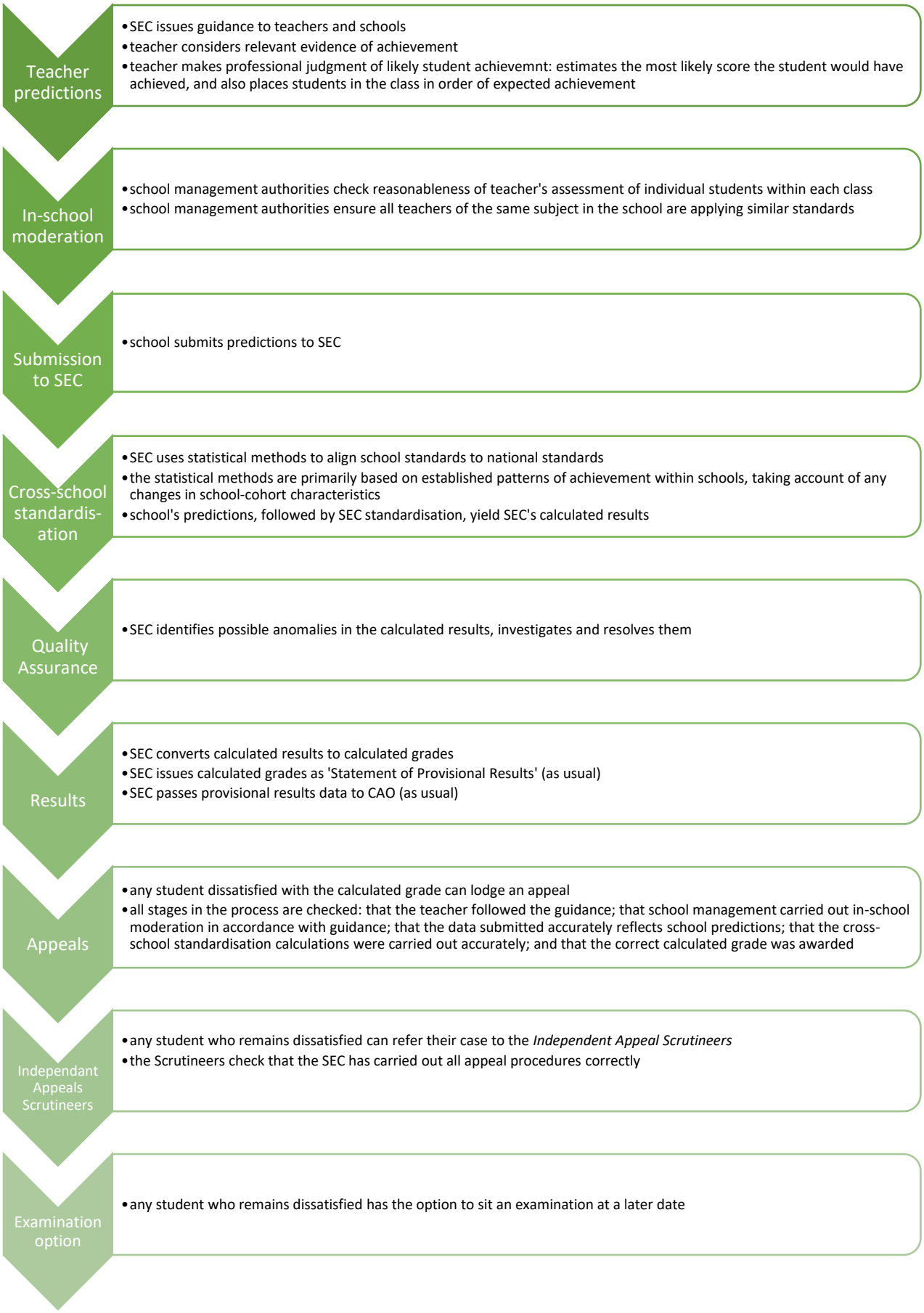
Premise 3: teacher-estimate data for one subject should not be allowed to influence student likelihood functions or conditioning distributions for another subject.

Premise 4: the model must adequately accommodate intra-school teacher effects on likely attainment.

Premise 5: adequate records should be generated to facilitate reasonable oversight by school authorities and to facilitate the implementation of a suitable appeals process.

In accordance with these premises, the following process is proposed. It is described in terms that suggest that the first two stages are predominantly paper-based, but could be adapted in the event that a suitable and practicable digital means of integrating and supporting the first three stages can be made available and readily accessible to all teachers and schools.[1]

[1] As noted in the preface, it was expected at the time of drafting this paper that the SEC would implement this process. Since then, the Calculated Grades Executive Office was established within the Department of Education to do so.


Teacher predictions
•SEC issues guidance to teachers and schools
•teacher considers relevant evidence of achievement
•teacher makes professional judgment of likely student achievemnt: estimates the most likely score the student would have achieved, and also places students in the class in order of expected achievement
In-school moderation
•school management authorities check reasonableness of teacher's assessment of individual students within each class
•school management authorities ensure all teachers of the same subject in the school are applying similar standards
Submission to SEC
•school submits predictions to SEC
Cross-school standardis-ation
•SEC uses statistical methods to align school standards to national standards
•the statistical methods are primarily based on established patterns of achievement within schools, taking account of any changes in school-cohort characteristics
•school's predictions, followed by SEC standardisation, yield SEC's calculated results
Quality Assurance
•SEC identifies possible anomalies in the calculated results, investigates and resolves them
Results
•SEC converts calculated results to calculated grades
•SEC issues calculated grades as 'Statement of Provisional Results' (as usual)
•SEC passes provisional results data to CAO (as usual)
Appeals
•any student dissatisfied with the calculated grade can lodge an appeal
•all stages in the process are checked: that the teacher followed the guidance; that school management carried out in-school moderation in accordance with guidance; that the data submitted accurately reflects school predictions; that the cross-school standardisation calculations were carried out accurately; and that the correct calculated grade was awarded
Independant Appeals Scrutineers
•any student who remains dissatisfied can refer their case to the *Independent Appeal Scrutineers*
•the Scrutineers check that the SEC has carried out all appeal procedures correctly
Examination option
•any student who remains dissatisfied has the option to sit an examination at a later date

Of course, the devil remains in the detail of the 'statistical methods' to be used in the cross-school standardisation step, but the stated premises and the preceding discussion serve to clarify that in large measure. The proposed model deals with intra-school teacher effects procedurally as well as statistically, with emphasis given to maximizing the validity of school-based procedures and data.

It is noted that, in a few subjects, some examination components have already been completed by candidates. In one such case, the material has already been marked, while in the remaining cases it remains unclear as to whether it will be feasible to mark it. Notwithstanding the potential availability of this additional data, it is considered that in all cases the teacher estimates should relate to expected performance across all components of the examination. It is anticipated that components already marked, or which may be marked later, can be used as part of a validation process rather than feeding directly into the calculated results.

9 Data collection

9.1 Data to be collected from schools

Notwithstanding the earlier arguments as to why teacher professional judgments based on school data and other locally held information and knowledge should be favoured over the use of pre-existing raw school-based data, relying on teacher judgment carries risks that need to be mitigated. The quality and accuracy of the information provided is of utmost importance. Furthermore, it is important that the task set out for teachers is both manageable and achievable. Given that this is the first time that schools and teachers will be required to provide information on students' expected performance, it is essential that the process is as clear and straightforward as possible. We do recognise and have considered the indications from the research outlined in Section 5 to the effect that composite data built from multiple granular judgments can in many circumstances be more reliable than single overall judgments. Nevertheless, we consider that in order to ensure that teachers are confident and comfortable with the task that they are presented with, it is essential to both restrict the number of data points sought in respect of each student and to ensure that type of information sought is of a straightforward and familiar form.

Accordingly, teachers and schools will be required to provide two key pieces of information which they will generate by considering a range of quantitative and qualitative evidence and relying on their professional judgement. This information will provide the basis for further processing that will lead to the generation of the results:

(i) Estimated percentage mark for each candidate

(ii) A rank order for each class for each subject

The estimated percentage mark will reflect what the teacher considers to be the most likely score the student would have achieved in the event that the COVID-19 pandemic had never occurred, schools had proceeded as normal to the end of the year, and the examinations happened as usual. In the case of multi-component subjects, it will be a single overall judgement in respect of the examination as a whole, including relevant coursework, practical work, oral and aural components.

In addition to the two pieces of data that directly inform the score calculation, teachers and school management will also be required to record information describing the evidentiary basis for the teacher-sourced estimates and confirm the accurate application of estimation procedures in each

school. These additional data are evidence of process validity that must be available to support the appeals process.

The order in which the subtasks that lead to the generation of the two data points for each student in each subject may differ depending on what preliminary work one anticipates the teachers engaging in and the degree to which a school uses digital tools to facilitate the process.

The practicalities surrounding the school-based data that feeds into the judgmental process, the formats in which it is available to teachers, and the anticipated ways in which they are expected to consider and weight the evidence are more appropriate to the estimation of percentage marks. In this case, the assignment of ranks would primarily serve to break ties in assignment of percent scores that cluster around arbitrary thresholds (e.g., multiples of 5 or 10 and benchmark scores). The rank order will, in large measure, flow directly from the estimated percentage marks, but tied ranks within a class/subject group will not be permitted, with the consequence that teachers will be required to give an explicit rank-ordering for any two or more candidates who have been placed on the same estimated percentage mark.

The other option is to ask teachers to perform the ranking first and then refine this into an estimated score for each student. Within-class ranking tends to be more reliable than percent-score estimates of performance on an external test, which suggests initially rank-ordering may reduce the burden of percent-score estimation. However, it may also reduce the specificity of the teacher judgments because rankings may be more susceptible to influence of non-achievement factors.

A pragmatic consideration is the degree of effort required to perform an adequate set of pairwise comparisons of students to accurately assign rank order. Realistically, managing teacher response burdens will require that teachers either use digital tools to facilitate the rank-ordering or perform the percentage mark estimation first.

9.2 Data Collection Tools

All guidance and supporting process documentation will be transmitted to schools digitally, and all data will be collected from schools and transferred to SEC digitally. However, different schools and teachers may employ a combination of paper-based and digital tools to facilitate the within-school data collection process. This process would encompass the following activities:

1. Collective review and training using guidance documentation (see section 9.3)
2. For each student in each subject, recording the evidentiary basis for teacher judgments
3. For each student in each subject, recording the percentage and rank-order information for each student
4. At the school level for each subject, confirming the accuracy and correctness of the in-school procedures. It should be noted that the school management is responsible for ensuring the integrity of the process, but not for the assignment of estimates to individual students.

Schools and teachers may perform all these tasks digitally or, alternately, in paper and either individually record the results digitally or pass the paper records to school management for digital data entry. Given that the data must be ultimately stored and used in digital form, the more steps that enter data directly onto a digital platform, the less overall effort will be required to support the process. However, for many reasons (e.g., connectivity, level of technological comfort), different schools and teachers may wish to perform more tasks using paper and pencil. Therefore, the data collection tools should be able facilitate both approaches to elicit and capture equivalent data.

The set of data tools will depend on whether the process performs percentage mark or rank ordering estimation first. Both workflows are described conceptually below. The development of the tools should be informed by feedback on usability from practicing teachers and school management prior to full implementation.

9.2.1 Initial percentage mark

The following description of the percentage-mark-first process applies to both digital and paper-based approaches. Having identified the evidence to be used for the estimation of a percentage mark, this evidence should be reviewed and a form completed as part of this process (Form X). There should be an individual form for each student. Once these have been provisionally completed, the candidates could be ordered according to their estimated percentage mark. In the digital tool, this would be performed automatically, and teachers using the paper-based process would place the completed Form X's in order with the form/student with the highest estimated percentage mark on top and the lowest at the bottom.

On a separate form (Form Y), the teacher would verify the rank orders and resolve any ties. Using the paper-based approach, the candidate number or other identifier with the highest estimated percentage mark should be entered as number one on the form Y and so on working through the class set of Form X's entering the students in the order of their estimated percentage mark. If, during this process, there are two or more students with the same estimated percentage mark the teacher should consider whether the mark that they have estimated for each is the correct mark or should the mark be moved up or down for any of them (and adjust the appropriate Form X's and re-order the candidates). If, following such consideration, it is concluded that the mark estimated is the correct mark, the teacher must then make an active decision in relation to rank ordering these students. Using a digital interface, the teachers would review the order of students in an automatically generated Form Y and either click the student identifiers to review and modify their respective Form X's or manually drag-and-drop the student identifiers with the same percentage marks into appropriate ranks.

In order to ensure that the standard is appropriate and satisfactory at an individual student level and across teachers/classes, the estimated marks for each class group should be reviewed by school management to confirm as best they can that the standard applied by the teacher is fair and appropriate and the Form X signed off on. This can be achieved by a member of the school management co-signing each of the Form X's having reviewed them. This is of particular importance where a school has a large number of candidates entered for a subject and more than one teacher teaching the subject.

This approach scaffolds the teacher through the process of estimating a percentage mark. Through the completion of the form the teacher is engaging with the various evidence identified for the estimation of the mark and is confirming what it is that they have reviewed as part of the process. The use of the Form X for the rank ordering of the students is similar to a pairs-analysis process when two or students are on the same estimated mark.

The nature of the process from estimating the percentage mark, the assembling of a rank order and the review of the estimated percentage mark by school management, results in the generation of a record of the process under taken to arrive at the two sources of information required for the calculation of a mark for each student in each subject. This record would be very useful as part of the quality assurance process and also at appeal stage.

If a school has opted for all tasks to be conducted using a paper-based process, it will be critical that the centralized data entry process enters data separately from each teacher, mimicking the direct data entry of schools using the fully digital approach. This data entry protocol is required to maintain accurate linkages between students, teachers, and classes.

9.2.2 Initial Rank Order

Rank ordering is an inherently more complex process than estimation of percentage marks, because, in its extreme, it could involve a complete set of pairwise comparisons between all students in a class (380 unique comparisons for a class of 20; fortunately, the actual number of comparisons can be reduced substantially using efficient algorithms). Because respondent burden affects accuracy of results, fairness dictates that all teachers should follow the same procedure, and the complexity of efficient sorting algorithms requires that the rank-order-first approach be facilitated by a standardized digital tool.

An example of one such process is the application of an efficient sorting algorithm, such as Quicksort[60], and where all student classifications are performed by drag-and-drop operations on icons with student identifiers into ranks.

Initial rank ordering will require the same consideration and documentation of evidence as the percentage mark-first approach, using a similar Form X. However, the initial judgment on Form X will be classification of each student into one of two ranks: higher than average or lower than average. After each student's initial classification, the interface randomly selects a student for each of the two distinct ranks, and the teacher classifies the remaining students with the same rank as above or below the selected student, producing two sub-ranks. Within each sub-rank, the process is repeatedly recursively until there is only one student in each distinct sub-rank. This algorithm has been shown to be, on average, approximately 7 times faster than pairwise ranking[61].

The estimation of percentage mark scores could be performed concurrently with the rank order estimation, where the teacher only estimates the percentage mark for each student that the interface randomly selects, such that, at the end of the process, all students have been assigned both a rank order and percentage mark. Alternatively, following a discrete estimation of rank order, the teacher could work sequentially from the top of the ordered students and assign percentage marks, which the interface would constrain to be less than the lowest score of any higher-ranked student.

The conclusion of this process requires the same evidentiary record and review and confirmation from school management as the percentage-mark-first approach. Although it produces data with comparable utility compared to the percentage-first process, its greater complexity and dependence on digital tools may reduce its practicality.

9.3 Advice for teachers and schools

To assist teachers and schools with the process it is essential that clear guidance is provided to them which sets out their role and the role of school management in the process. The guidance should include:

- a description of the certification that students will receive on results day
- an outline of the process for generating those results, making clear the importance of their role in it
- guidance on how to determine an estimated percentage mark and on what sources of evidence should be used

- guidance on how to determine the rank order
- information on how and when to submit the data
- guidance on dealing with students in particular circumstances, such as students who have recently come to the school, students who would be in receipt of a bonus for taking examinations through Irish, students who have been approved for examination accommodations, students studying subjects in another school or outside of school, and so on
- an outline of the appeals process
- information in relation to access to data

10 Additional Considerations that Apply to the Partial-cohort Scenario

The preceding sections are all formulated in terms of the primary task that the technical working group was set up to consider: the use of calculated results for the award of Leaving Certificate grades in the event that no examinations could proceed. We turn now to the additional issues that arise in the case of the additional task remitted to the group: the potential use of the same or a similar approach in the event that some form of examinations do proceed at some point later in the summer, and results need to be generated for candidates who are unable to sit for reasons related to COVID-19.

The most significant differences that arise were tabulated at the outset of this paper in Section 1 as follows:

If no examination is sat:	**If an examination is sat and we are generating estimates for missing candidates only:**
• All candidates are treated in the same way by whatever process is agreed.	• Two groups of candidates will have followed two very different routes to certification, which places a significant addition comparability burden on the processes employed.
• The results to be estimated are the results that candidates would have been expected to achieve in the June examinations if the pandemic had never arisen.	• It seems most likely that the results from the late-summer sits will stand as results in their own right, rather than being treated as estimates for further statistical treatment. If so, then the results to be estimated for the missing candidates are the results they would be expected to have achieved in the late-summer form of the examination under comparable conditions to those who actually sat them, which is a more difficult estimation problem.
• No actual live examination performance information is available to combine with teacher judgment data when estimating results.	• Both estimated performance data (on one examination form taken under one set of conditions) and actual performance data (on another examination form taken under different conditions) will exist for a subset of candidates. This would be highly relevant bivariate data in estimating the likely performance of missing candidates, so it could not credibly be ignored.

Some of the assertions made in earlier sections need to be amended in the case of this different context.

In addition to the methodological alternatives already described in Section 7 above, the following possibility arises:

10.1 Use of Student-sourced Information as Part of an Estimation Model

In the table above it is assumed that the results of any examination taken later in the summer would stand as results in their own right, having precisely the same status as the results that would have been generated if the pandemic had never occurred. However, there is another option. The data from such examinations could, in theory at least, not be regarded as 'proper' examinations that directly yield results, but as additional data that provides information about likely candidate performance on the regular June examinations under normal conditions.

If fed into the same conceptual model as previously described, these examination outcomes are an additional source of student-sourced data. If the examinations use shortened forms, the results are unavoidably less reliable with respect to typical Leaving Certificate examination results. This takes the form of both reduced accuracy and content coverage. When content coverage is reduced, the results cannot fairly support interpretations that are consistent with results from typical Leaving Certificate examinations.

The previously presented Table 1 from Section 6 above now becomes:

Table 3 Data sources

Data Type	2020 Grads*	Pre-2020 Grads*	Intended Use**
1. LCE results from partial cohort	Y		S
2. Teacher estimates of student performance	Y		S
3. Teacher/Class membership	Y		C
4. School membership	Y	Y	C
5. Junior Certificate Examination results	Y	Y	C
6. Student 2-or-3-year programme status	Y	Y	C
7. Student demographic characteristics	Y	Y	V
8. Full LCE results		Y	C

*"Y" indicates data are linked at the student level to globally unique student identifiers

**How the data will be used in the procedure:

S=estimating specific student scores

C=estimating conditioning distributions

V=validation (not directly contributing to estimates)

The reader is reminded that the data sources are listed in order of expected usefulness with respect to estimating individual student examination results, with LCE results from the partial cohort now being considered the most relevant. As previously noted, this order also implies a hierarchy of credibility; in the event that evidence from less-relevant sources disagrees with evidence from more-relevant sources, higher credence is given to the more-relevant sources.

If such examination data for part of the cohort are available, there is justification for them to be augmented with additional data from the other sources above in order to produce final estimates of student scores. However, given the credibility hierarchy of data sources, the statistical methodology would need to ensure that the inclusion of additional data refine the accuracy of results without fundamentally contradicting the direct response data. This approach is possible in a likelihood-based estimation framework defined earlier.

10.2 Alternative Model for Partial-cohort Scenario

Notwithstanding the theoretical attractiveness and defensibility of using partial-cohort LCE results as input data in the manner descried above, it seems highly unlikely that this approach would prove acceptable to stakeholders. It would not be acceptable in the current context to award any candidate who sat an examination a grade that is lower than the one that their raw score would generate in the event that it was marked and graded in the ordinary way as a normal examination.

It should also be noted that, in the considered opinion of senior management of the Examinations and Assessment Division of the SEC, the alternate forms and dramatically altered conditions under which any late-summer sit would take place are such that the outcomes of those examinations will not be comparable to those that the original process would have generated under normal conditions. In particular, it is considered unlikely that the SEC's normal standard setting processes will be adequate to the task of yielding a comparable grade distribution in each subject to those that occurred in the past. Even if they were, this national distribution could not be assumed to disaggregate down to school level in the same way as would have been the case in any other year.

Furthermore, if the candidates who sit the examination are undergoing a process that generates results that are significantly less equivalent to other years than is usual, and these results are not being adjusted to instead provide estimates of likely performance on the normal examination under normal conditions, it seems most reasonable to suggest that the estimation task for those who did not sit should be to estimate their most likely score in the event that they sat the same examination and under the same conditions as their peers, as distinct from estimating their likely performance on the normal examination under normal conditions.

Accordingly, it is considered that the data and assumptions that form the basis of the conditioning distributions in the full-cohort scenario become a somewhat less reliable and appropriate basis for estimation in the partial-cohort scenario. On the other hand, a new and potentially very valuable mechanism for estimating the missing scores becomes available in this scenario, assuming that a reasonable proportion of each class takes the examination. That is, models that treat teacher predictions as an independent variable and performance on the examination form taken as the explained variable provide a sound basis for such predictions, because now there is a set of observations of the explained variable upon which to build the model. Such models could be based on population data (which presents a large data set) or the data within the class (which encapsulates the teacher's 'local standards') or some combination thereof.

We consider this to be the more promising approach in this scenario. It may be noted that the same type of data collected in the same way can serve the needs of the preferred model in either scenario.

11 Dealing with measurement error

No matter which scenario is involved, consideration needs to be given to how measurement error is to be dealt with and reported on. Measurement error can be quantified at an individual level or at a general level. At a general level, the 'overall' level of measurement error in a given model can be quantified by various methods, such as by comparing the amount of 'explained' variance to 'unexplained variance' when the model is applied to a data set with known outcomes.

In some of the models under consideration here, it may be difficult to quantify the true levels of measurement error, as teacher estimates and actual outcomes from an examination of the type originally envisaged will not be simultaneously available in order to measure it.

Likelihood-based estimation methods have inbuilt mechanisms for establishing measurement error based on the degree of fit of the data to the model. Additionally, each individual estimate of a trait parameter will have its own associated level of measurement error, which will depend on the extent to which the data pertaining to that individual fits the model.

We consider that, irrespective of the model chosen, the reporting of estimation error at the individual level would be problematic in this context, and we take the view that transparency is adequately served by dealing with measurement error in full in the overall technical report that will follow implementation of the process.

12 Reporting outcomes – statements of provisional results, certification, and access to further data

Notwithstanding that the basis on which grades are issued will be different from in other years, certification and related processes will remain as normal. On 'results day', candidates will receive a statement of provisional results in the same format as usual – that is, a statement in respect of each subject taken, the level at which it was taken and the grade awarded. The data will be passed to the Central Applications Office as usual. Likewise, the certificates that issue later in the year after appeal processes are complete will be indistinguishable from those of other years.

Consideration also needs to be given to the 'explanatory notes' (annotations) associated with certain examination accommodations. It may be argued that, since the teachers' estimates will be based on the premise that accommodations sought and granted would have been made available, the most reasonable and consistent thing to do is for the annotation to stand. If the teacher does as instructed, the explanatory note remains true, albeit with a somewhat different meaning to the term 'were assessed' from the usual one.

Irrespective of the model used for generating the results, consideration needs to be given to what information is made available to candidates automatically, what information might be available to them through the procedures that might otherwise have given them access to their component marks, and what further information might be available to them under freedom-of-information or data-protection legislation.

Component marks will not exist, but these are in any event not reported on either the statement of provisional results or the subsequent Leaving Certificate. However, component marks were made available to candidates through a separate online process in 2019, with the intention that this would continue.

It seems reasonable to suggest that transparency demands that candidates be entitled to know the final calculated percentage mark that led to the grade awarded. What is less clear is whether they should be entitled to know (after the event) what the estimated mark that their school submitted was, and/or what was their rank order in the class. Non-disclosure of this information helps protect the process from an integrity perspective, because if the teachers know that the information will later be disclosed, the estimates they make may be inappropriately influenced by that knowledge. This poses a risk to the quality of the data. The correct balance between this need for transparency and the aim of getting the best possible estimates will need to be considered and resolved.

13 Appeals and further recourse

An appropriate transparent appeals process will need to be in place as part of any process for arriving at a set of calculated results for the 2020 Leaving Certificate.

There are three possible scenarios:

13.1 Appeals and further recourse – whole-cohort scenario

In the scenario where no examinations are held and all candidates are issued with calculated results, the following points of recourse will be available to candidates who are dissatisfied with the grades awarded:

(a) Appeals

Candidates may appeal the calculated grade awarded by the SEC

This will trigger an appeal process comprising

(i) a desk-check that all SEC-based processes were carried out correctly: that the data handed off to the moderation algorithm in respect of the candidate was that captured from the school; that all statistical and other processes were transacted correctly; that any or all interventions made in respect of the processing of the data were appropriate; that the calculated final grade awarded corresponds to the final calculated percentage mark

(ii) direct contact with the school to seek confirmation of the following based on the records generated during the process: that the teacher followed the correct process and the SEC guidance in arriving at the professional judgement made; that the school authorities carried out the in-school moderation processes in accordance with SEC guidance; that the data held by the SEC in respect of the candidate matches the school's record of the data they intended to submit; and accordingly, that the school authority is satisfied that the data as transferred is correct.

(b) Independent Appeals Scrutineers

Candidates who remain dissatisfied can refer their case to the Independent Appeals Scrutineers. The scrutineers will check whether the SEC has carried out all appeal processes correctly.

(c) Later examination

Candidates who remain dissatisfied with the outcome of this process will have recourse to sitting a set of Leaving Certificate examinations at a later date in any subjects in which they feel they have not been awarded the grade they believe they should have received.

Recourse (c) above represents a 'backstop' for those who consider that they have not received an appropriate grade and have accordingly been disadvantaged by the non-availability of the normal examination-based procedure for achieving certification.

13.2 Appeals and further recourse – partial-cohort scenario

In the scenario where formal examinations of some form are held and calculated results are issued only to those candidates who, for reasons related to COVID-19, were unable to present, the following points of recourse will be available to candidates who are dissatisfied with the grades awarded:

(a) Appeals

Candidates may appeal the grade awarded by the SEC, whether it is a calculated grade or otherwise

The standard appeal process for written examinations will apply to candidates who presented for examination.

For candidates who were awarded a calculated grade, the appeal process outlined under heading (a) in the preceding subsection will apply.

(b) Independent Appeals Scrutineers

Candidates who remain dissatisfied can refer their case to the Independent Appeals Scrutineers. The scrutineers will check whether the SEC has carried out all appeal processes correctly.

It may be noted that, since an opportunity to be certified by examination has already been made available, the rationale for recourse (c) in the preceding subsection does not apply in the partial-cohort scenario. Accordingly, there will be no recourse to a later examination (other than the normal option to 'repeat' by taking the 2021 examinations).

14 Messaging and communication

The associated messaging will have to be timely, understandable to the public at large, and have all the critical elements associated with openness and transparency. In the context of timely communication, information should be released at the point that it is announced that work is being undertaken in respect of the possibility of calculated results being used to certify some or all of the 2020 Leaving Certificate cohort.

This information should describe clearly the context and purpose of producing calculated results. It should also clearly describe how results would issue and how the results would be used for transition to higher and further education and the world of work. While being as reassuring as possible, It should be made clear from the outset that, though at a systemic level it is possible to arrive at a set of results nationally for each subject which are fair and reasonable, there is no available methodology which eliminates all error at an individual candidate level and which could accordingly completely prevent candidates from receiving results which either underestimate or overestimate their attainment in any given subject. However, the communication should assure stakeholders that every effort will be made to apply a broad and understandable statistical process that minimises such error. It would also be indicated at this stage that candidates would have recourse to an appeals process through which they could appeal any results with which they were dissatisfied.

The process through which calculated results will be produced should be described. This would include describing the information that will be sought from schools, the range of other available information that will be used in arriving at a calculated result for each subject, and an overview of the actual processing of data itself. The existence of internal quality assurance checks throughout the process and the appeal process should also be described. If, as a result of no examinations transacting in July/August, there is an option of sitting full-scale examinations at the earliest feasible opportunity as a final backstop to cater for a candidate who remained dissatisfied with their calculated results, then this should also be made clear.

At a minimum a further similar set of communications should issue prior to the issue of provisional results. In this instance more detailed information on the appeals process should be provided, specifically, how to appeal and the detail of what process entails made fully clear.

Appendix 1: Psychometric Framework

The challenge of estimating student examination performance in this case, where directly observed examinations response data are completely missing, is an extreme example of the assumptions underlying any examination performance. All assessments define a construct as the universe (or domain) of observable behaviours and characteristics that represent the subject being assessed. For most assessments, it is not possible to include all the possible behaviours and tasks (collectively, items) that define the construct in a test form. Instead, assessments sample from the universe of items, and generalize to the entire universe of items, including those that are not included in any test forms. Linear scores (i.e., scores calculated by some form of average of the observed item performances) that the test produces estimate how students are expected to perform on the entire domain. Under this assumption, results from the 2019 Leaving Certificate examination are considered equivalent to all previous years of the Leaving Certification examination.

In most domains, the items can be clustered into classes that require similar sets of skills and knowledge – in most cases, the classes of items reflect the structure of curricula (threads, strands, themes, etc.). Assessments with the greatest validity use a sample of items that are randomly and fully representative of the set of classes that comprise the domain, as illustrated in Figure 1. If test performance is equal to the percentage of items answered correctly, then it is clear that performance on the sample of items in the test will approximate performance on the entire domain of items.

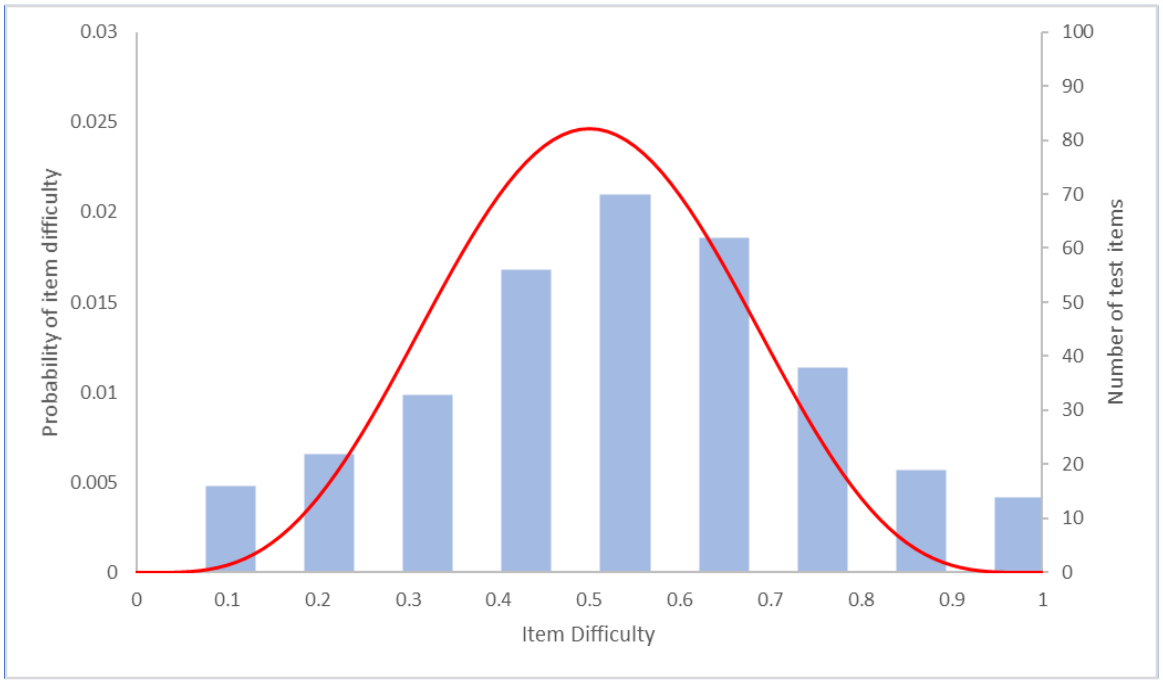


Figure 1 Example: the difficulty of randomly sampled items reflects the distribution of difficulty in the universe of items. Item Difficulty is the proportion of students expected to perform incorrectly. Height of bar indicates number of items.

However, for practical reasons, such as development cost, response burden or marking time, it may not be possible to have the set of items on a test form be randomly representative of the domain. As a result, the test form under-samples from some classes of items (e.g., complex performances tend to be under sampled, not because they are less important, but because they are time consuming during both testing and marking). To remedy this imbalance, some items are assigned score weights that reflect the relative number of similar items that are not included in the test form. The resulting

linear test scores produced by this weighted process again generalize to the entire domain, as illustrated in Figure 2. Although the items are no longer truly representative, the weighting can recover reasonable estimates that generalize to the entire domain.

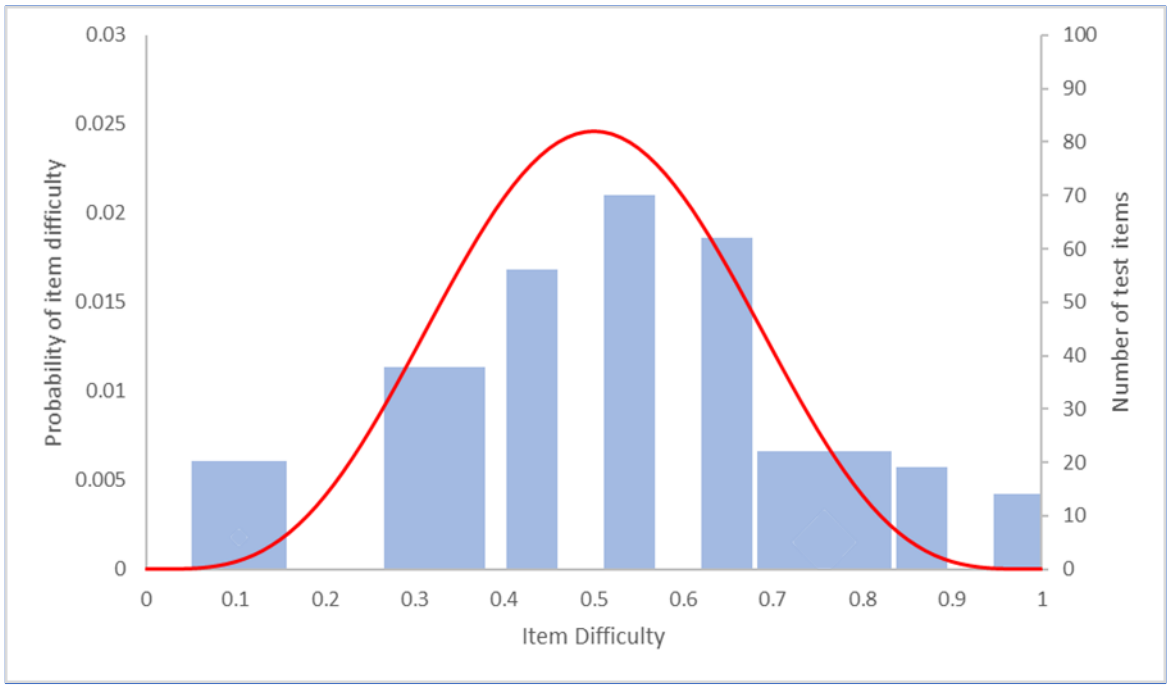


Figure 2 Example: unequal weighting of sampled items corrects for sample bias. Item Difficulty is the proportion of students expected to perform incorrectly. Width of bar is proportional to weight.

However, in less ideal circumstances, the samples of items are too biased to estimate scores consistently for all students. For example, a test that includes a disproportionate number of easy items may be able to estimate consistent scores for low performing students but not for high performing students; low proficiency students may have estimated scores that are reasonably close to their true score, but the estimates would be biased for higher proficiency students, and all students with proficiency above a certain level would be expected to have scores of 100% (see Figure 3). When there are large differences between assessment domain and the scope of the test form, it is unlikely that any simple weighting scheme will be able to adequately correct for the missing items using linear scoring. Consider 3 students, A, B and C, with true scores 57%, 65%, and 74%, respectively, expressed on the Item Difficulty scale (e.g., a true score of 0.57 for student A indicates the student is will answer 57% of the items in the entire domain correctly). If each of the three students correctly answers no item with a difficulty higher than their true score, 50% of items with difficulty equivalent to their true score, and all items with a lower difficulty, the test would provide the estimates above each arrow: 70%, 85% and 100%, corresponding to the cumulative proportion of items with difficulty at or below a student's true score. All estimates have bias that increases dramatically as the true score increases.

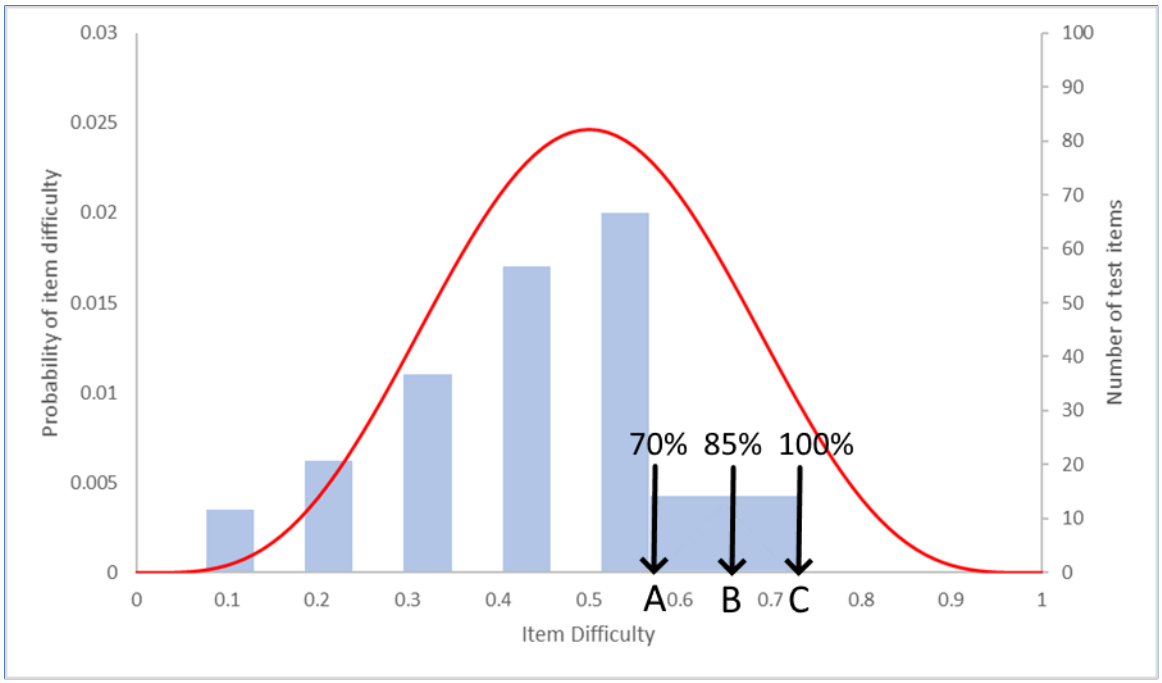


Figure 3 Example: sample bias cannot be reasonably corrected with weighting

A more appropriate scoring method that can make appropriate use of the properties of the test items uses the concept of *likelihood* to estimate student proficiency. For each item, a certain score is more likely to be produced by students within a certain range of proficiency. In the simplest case, for an item scored as either correct or incorrect, a correct response is more likely to be produced by high proficiency students and an incorrect response is more likely to be produced by low proficiency students. When a score is observed for a student with an unknown proficiency, a correct score would indicate the student is more likely to have a high proficiency and less likely to have a low proficiency. By combining the information about the likelihood of the student's proficiency across many items with both correct and incorrect scores, it is possible to determine that the most likely location of the student's proficiency is in a very narrow range. Thus, knowledge about the statistical properties of items can be used to generate consistent estimates of domain performance even if the sample of items is biased. Figure 4**Error! Reference source not found.** illustrates how likelihood functions describe student scores in the case of an unbalanced test like the example in Figure 3. The score estimates in this case are illustrated by the arrows indicating where the maximum values of the likelihood functions land on the Item Difficulty scale. As long as a student has both correct and incorrect items responses, much of the sampling bias is corrected, but the bias for students with no variation in item scores remains large.

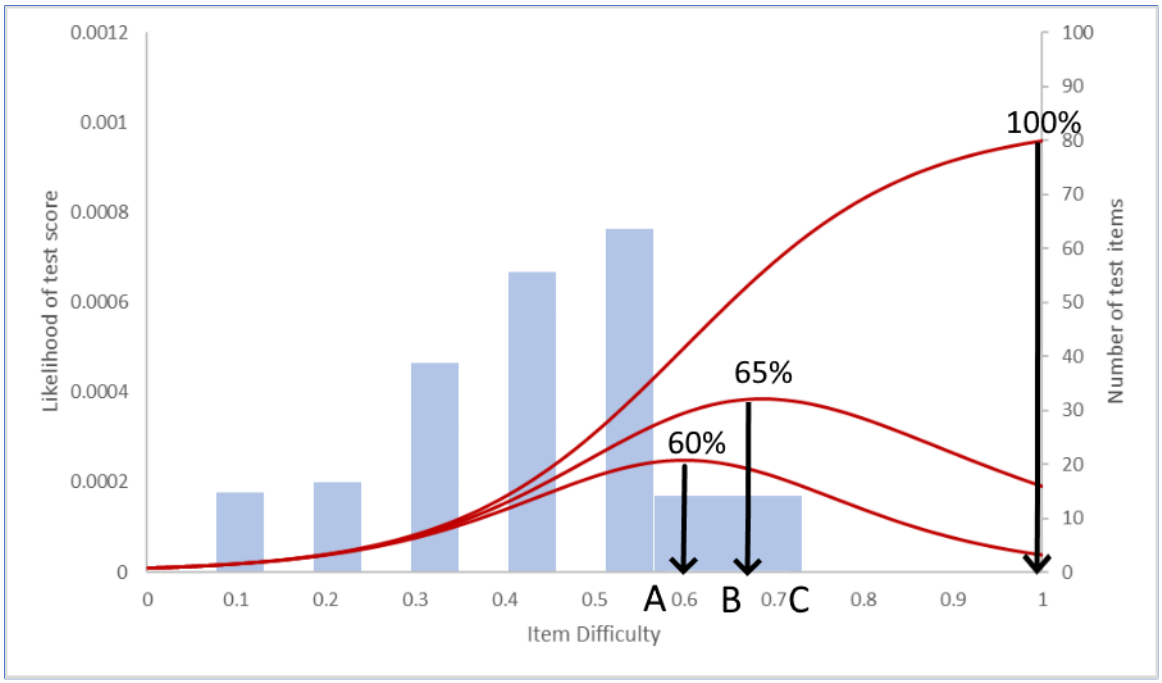


Figure 4 Example: likelihood-based scoring produces more consistent estimates

The concept of using likelihood to estimate test scores can be extended to incorporate information that indirectly describes student proficiency. For example, if a student is known to belong to a population of students with a known distribution of proficiency, it is still possible to make reasonable statements about that student's performance even in the absence of any observed test data. Indeed, the common practice of reporting population averages of test results does exactly that – the population average is the expected test score of any member of the population, and may be interpreted as the most likely score to be produced by a randomly-sampled member of the population.

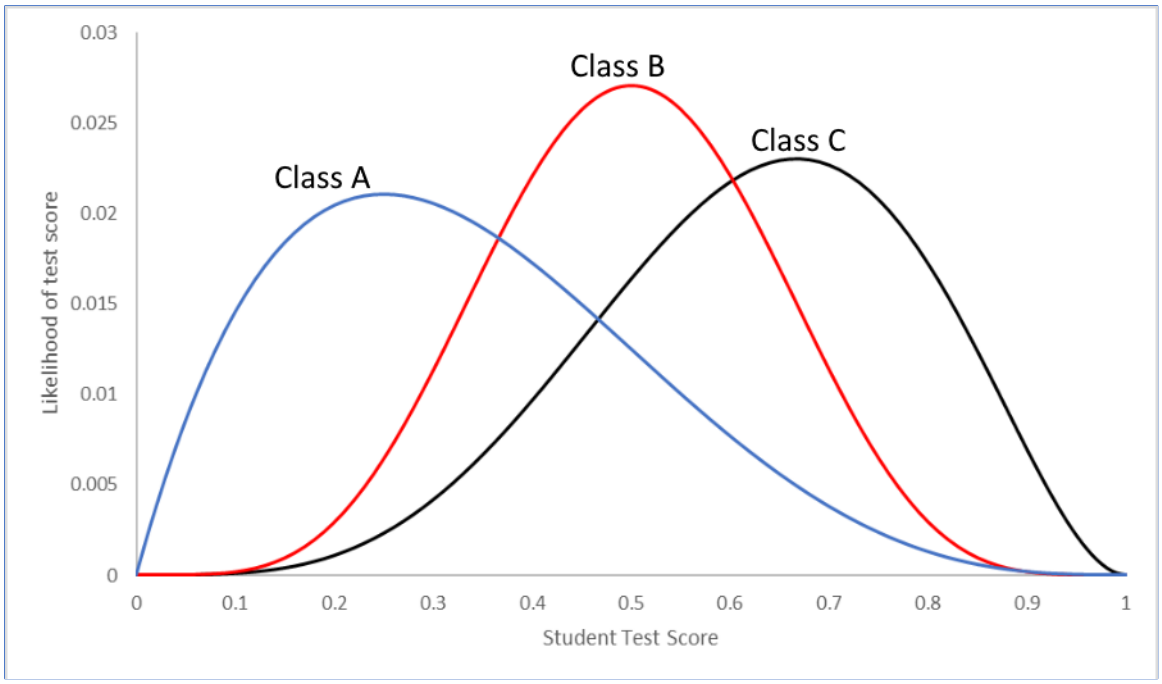


Figure 5 Example: group membership in classes informs individual score estimation with group distributions

However, the overall population distribution is not useful for discriminating between members of the population, because it applies to everybody equally. Instead, it helps to consider that the overall population distribution is composed of many smaller component distributions: the distributions of different schools, and the distributions of different classes within a school, as well as the distributions of students with different individual characteristics. Even if individual student results are not available, if information about the various distributions that apply to each student are known, that information may be used to estimate the likelihood of scores for students that are nested within the sample families of distributions (e.g., class or school and other student characteristics). This principle is illustrated in Figure 5**Error! Reference source not found.**, where the score distributions of three classes, A, B and C are clearly distinct. Based on this information it is reasonable to infer that a random student from Class B is likely to perform higher than a random student from Class A and lower than a random student from Class C.

Finally, the likelihood of specific test scores may also be estimated indirectly using data from one or more correlated measures. When two measures are correlated, each value of one variable corresponds to a specific distribution of the other. A common example in education is the relationship between holistic grades and percentage grades. Typically, holistic grades are determined through a simultaneous consideration of multiple sources of evidence, to which expert judgment is used to determine a single letter grade (e.g., A, B, C, D, F). Conversely, percentage grades are calculated by combining numeric information from several numerically scored sources of evidence (which, in the case of examinations, may be single test items). However, if the two variables are measures of the same construct, then each letter grade will correspond to a distribution of percentage scores, which means that, if a student has only a letter grade, it is possible to estimate not only their most likely percentage score, but also the likelihood of all other possible percentage scores, as in Figure 6.

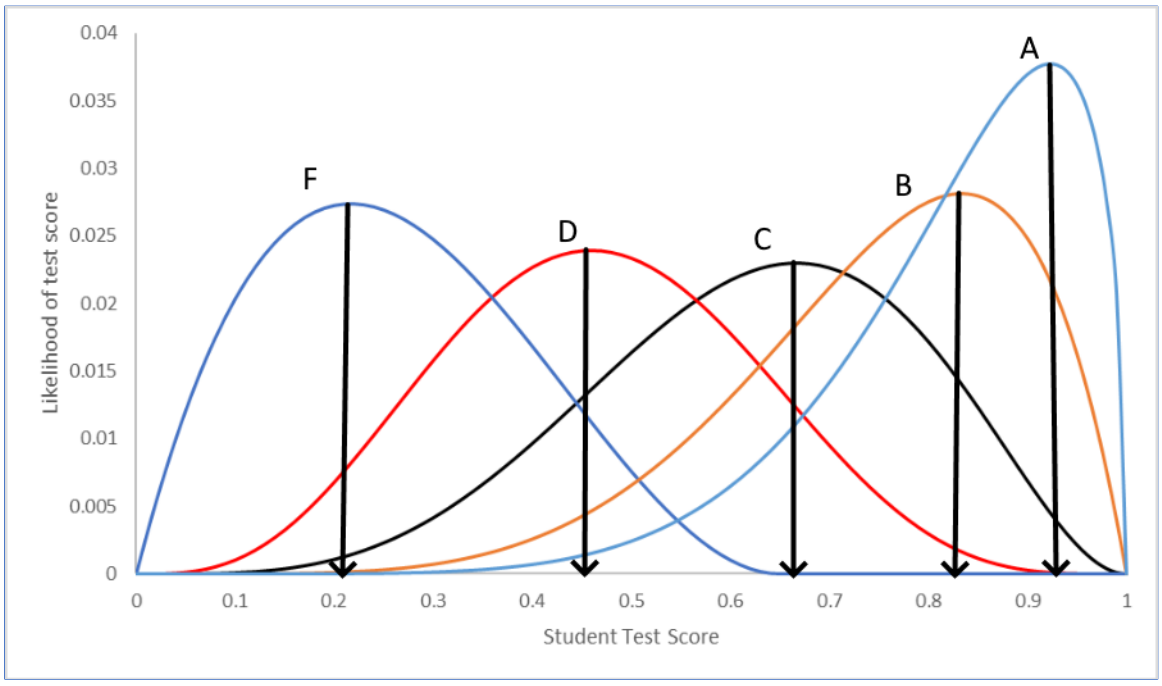


Figure 6 Example: conditional likelihood functions based on correlated data for assigned letter grades

In the absence of some or even all item response data, the likelihood-based approach to scoring provides a methodology for estimating test scores that incorporates non-test data yet remains consistent with the estimation of test scores using a fully representative sample of items from the

item universe. Using non-test data in this manner to estimate test results is commonly referred to as *conditioning*.

Unlike linear scoring, which assumes that either all variables are randomly equivalent or all variables have a mutually compensatory relationship (meaning high values in one variable can offset low values of another), likelihood based scoring allows the unique information from each variable to be carried through to the estimation process. For example, in the previous examples, even though the test data alone do not accurately estimate score for student C, if that student were in Class C and is also assigned a letter grade of A, the three separate likelihood function can be combined, as in Figure 7**Error! Reference source not found.**, to produce a more reasonable estimate for student. The relative contribution of each variable to the estimation of the final score is proportional to its precision with respect to the location of each student's proficiency. For example, if one student is nested within a school with a narrow distribution of proficiency, the information from school membership is intrinsically more useful for the estimation process than for a student nested within a school with a very broadly distribution of proficiency.

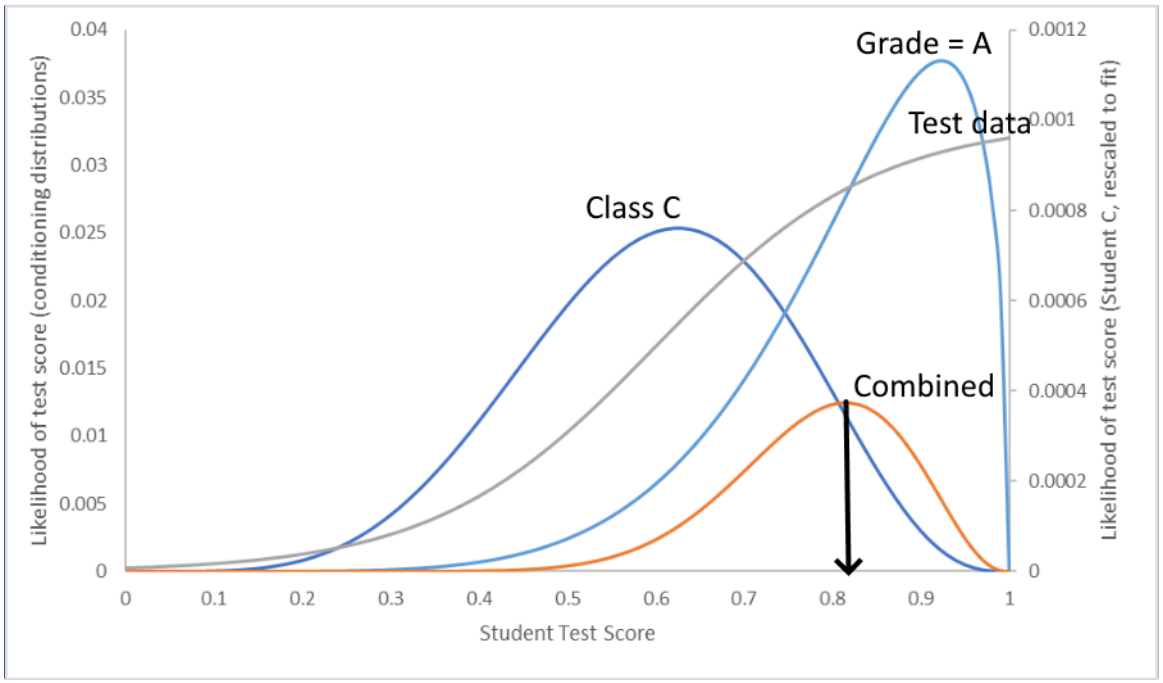


Figure 7 Example: Combined likelihood function using test data and non-test data

Application to the 2020 Leaving Certificate Examinations

Conceptually, all sources of information about a student may be used to estimate test scores. The most defensible are those based directly on student behaviours and characteristics in the domains measured in the Leaving Certification Examinations. Test data, even if they are biased or incomplete representations, are the strongest form of evidence because they are direct estimates. There are many indirect sources of student-level data relevant to estimating examination performance, but the only data universally available are teacher judgments about student proficiency to replace the missing examination data.

However, teacher data tend to be nested within a local frame of reference that makes standardized interpretations difficult. Therefore, the estimation process must use additional sources of information to correct for the localized bias of teacher judgments. The likelihood-based approach to

scoring provides a consistent approach to using all relevant information while minimizing dependency on arbitrary assumptions about how they contribute to the estimation process.

[1] Hoge, Robert D., and Theodore Coladarci. "Teacher-Based Judgments of Academic Achievement: A Review of Literature." Review of Educational Research 59, no. 3 (1989): 297-313. Accessed April 19, 2020. www.jstor.org/stable/1170184.

[2] Schroder, Carole & Crawford, Patricia. (1970). School Achievement as Measured by Teacher Ratings and Standardized Achievement Tests.

[3] Sharpley, C.F. and Edgar, E. (1986), Teachers' ratings vs standardized tests: An empirical investigation of agreement between two indices of achievement. Psychol. Schs., 23: 106-111. doi:10.1002/1520-6807(198601)23:1<106::AID-PITS2310230117>3.0.CO;2-C

[4] Shah, Rajesh & Brown, Gavin & Keegan, Peter & Burakevych, Nataliia & Harding, Jane & Mckinlay, Chris. (2020). Teacher rating versus measured academic achievement: Implications for paediatric research: Academic success in paediatric research. Journal of Paediatrics and Child Health. 10.1111/jpc.14824.

[5] Hoge & Theodore Coladarci (1989).

[6] Hopkins, K., George, C., & Williams, D. (1985). The Concurrent Validity of Standardized Achievement Tests by Content Area Using Teachers' Ratings as Criteria. Journal of Educational Measurement, 22(3), 177-182. Retrieved April 19, 2020, from www.jstor.org/stable/1435031

[7] Coladarci, Theodore. (1986). Accuracy of Teacher Judgments of Student Responses to Standardized Test Items. Journal of Educational Psychology. 78. 141-146. 10.1037/0022-0663.78.2.141.

[8] Demaray, Michelle & Elliott, Stephen. (1998). Teachers' judgments of students' academic functioning: A comparison of actual and predicted performances. School Psychology Quarterly. 13. 8-24. 10.1037/h0088969.

[9] Mingo, Maya & Bell, Sherry & Mccallum, R. & Walpitage, Dammika. (2019). Relative Efficacy of Teacher Rankings and Curriculum-Based Measures as Predictors of Performance on High-Stakes Tests. Journal of Psychoeducational Assessment. 073428291983110. 10.1177/0734282919831103.

[10] Martínez, José & Stecher, Brian & Borko, Hilda. (2009). Classroom Assessment Practices, Teacher Judgments, and Student Achievement in Mathematics: Evidence from the ECLS. Educational Assessment. 14. 78-102. 10.1080/10627190903039429.

[11] Dhillon, Debra. (2005). Teachers' estimates of candidates' grades: Curriculum 2000 Advanced Level Qualifications. British Educational Research Journal - BR EDUC RES J. 31. 69-88. 10.1080/0141192052000310038.

[12] Coladarci (1986).

[13] Begeny, John & Eckert, Tanya & Montarello, Staci & Storie, Michelle. (2008). Teachers' Perceptions of Students' Reading Abilities: An Examination of the Relationship Between Teachers' Judgments and Students' Performance Across a Continuum of Rating Methods. School Psychology Quarterly. 23. 43-55. 10.1037/1045-3830.23.1.43.

[14] Martinez *et al* (2009).

[15] Hopkins *et al* (1985)

[16] Martin R. Delap (1994) An investigation into the accuracy of A-level predicted grades, Educational Research, 36:2, 135-148, DOI: 10.1080/0013188940360203

[17] Johnson, Sandra. (2013). On the reliability of high-stakes teacher assessment. Research Papers in Education. 28. 91-105. 10.1080/02671522.2012.754229.

[18] Julia Klug, Simone Bruder & Bernhard Schmitz (2016) Which variables predict teachers diagnostic competence when diagnosing students' learning behavior at different stages of a teacher's career?, Teachers and Teaching, 22:4, 461-484, DOI: 10.1080/13540602.2015.1082729

[19] Fleckenstein, Johanna & Leucht, Michael & Köller, Olaf. (2018). Teachers' Judgement Accuracy Concerning CEFR Levels of Prospective University Students. Language Assessment Quarterly An International Journal. 15. 10.1080/15434303.2017.1421956.

[20] Begeny *et al* (2008)

[21] Delap (1994)

[22] Wright, D., & Wiese, M. (1988). Teacher Judgment in Student Evaluation: A Comparison of Grading Methods. The Journal of Educational Research, 82(1), 10-14. Retrieved April 19, 2020, from www.jstor.org/stable/27540330

[23] Thiede, Keith & Brendefur, Jonathan & Osguthorpe, Richard & Carney, Michele & Bremner, Amanda & Strother, Sam & Oswalt, Steven & Snow, Jennifer & Sutton, John & Jesse, Dan. (2015). Can teachers accurately predict student performance?. Teaching and Teacher Education. 49. 10.1016/j.tate.2015.01.012.
[24] Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. Journal of Educational Psychology, 104(3), 743–762. https://doi.org/10.1037/a0027627
[25] Adapted from Hoge & Coladarci (1989)
[26] Hoge & Coladarci (1989)
[27] Wasik, B. H., & Loven, M. D. (1980). Classroom observational data: Sources of inaccuracy and proposed solutions. Behavioral Assessment, 2, 211–227.
[28] Hoge, Robert. (1983). Psychometric Properties of Teacher-Judgment Measures of Pupil Aptitudes, Classroom Behaviors, and Achievement Levels. The Journal of Special Education. 17. 401-429. 10.1177/002246698301700404.
[29] WISEMAN, S. (1967), THE EFFECT OF RESTRICTION OF RANGE UPON CORRELATION COEFFICIENTS. British Journal of Educational Psychology, 37: 248-252. doi:10.1111/j.2044-8279.1967.tb01933.x
[30] Südkamp *et al* (2012).
[31] Dhillon (2005)
[32] Doherty, Jim & Conolly, Michael. (2006). How Accurately can Primary School Teachers Predict the Scores of their Pupils in Standardised Tests of Attainment? A Study of some non-Cognitive Factors that Influence Specific Judgements. Educational Studies. 11. 41-60. 10.1080/0305569850110105.
[33] Hopkins, K., George, C., & Williams, D. (1985). The Concurrent Validity of Standardized Achievement Tests by Content Area Using Teachers' Ratings as Criteria. Journal of Educational Measurement, 22(3), 177-182. Retrieved April 19, 2020, from www.jstor.org/stable/1435031
[34] Demaray, Michelle & Elliott, Stephen. (1998). Teachers' judgments of students' academic functioning: A comparison of actual and predicted performances. School Psychology Quarterly. 13. 8-24. 10.1037/h0088969.
[35] Sudkamp *et al* (2012)
[36] Hoge & Coladarci (1989)
[37] Harlen, W. (2005). Trusting teachers' judgement: research evidence of the reliability and validity of teachers' assessment used for summative purposes. Res. Pap. Educ. 20, 245–270. doi: 10.1080/02671520500193744
[38] Harlen, W. (2004). "A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes," in Research Evidence in Education Library. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.
[39] Glock, Sabine & Krolak-Schwerdt, Sabine & Klapproth, Florian & Böhmer, Matthias. (2012). Improving Teachers' Judgments: Accountability Affects Teachers' Tracking Decision. International Journal of Technology and Inclusive Education. 1. 86-95. 10.20533/ijtie.2047.0533.2012.0012.
[40] Pit-ten Cate, Ineke & Hörstermann, Thomas & Krolak-Schwerdt, Sabine & Gräsel, Cornelia & Böhmer, Ines & Glock, Sabine. (2019). Teachers' information processing and judgement accuracy: effects of information consistency and accountability. European Journal of Psychology of Education. 10.1007/s10212-019-00436-6.
[41] Feinberg, Adam & Shapiro, Edward. (2009). Teacher Accuracy: An Examination of Teacher-Based Judgments of Students' Reading with Differing Achievement Levels. Journal of Educational Research - J EDUC RES. 102. 453-462. 10.3200/JOER.102.6.453-462.
[42] Coladarci (1986)
[43] Demary & Elliot (1998)
[44] Begeny et al (2008)
[45] Martin & Thorpe, Andy & Hoskins, Sherria & Chevalier, Arnaud. (2008). Teachers' perceptions and A-level performance: Is there any evidence of systematic bias?. Oxford Review of Education - OXFORD REV EDUC. 34. 403-423. 10.1080/03054980701682140.
[46] Glock *et al* (2012)
[47] Martinez *et al* (2009)
[48] Meissel, K., Meyer, F., Yao, E., & Rubie-Davies, C.M. (2017). Subjectivity of teacher judgments: Exploring student characteristics that influence teacher judgments of student ability.
[49] e.g., Demeray & Elliot (1998)
[50] Kaiser, Johanna & Südkamp, Anna & Möller, Jens. (2016). The effects of student characteristics on teachers' judgment accuracy: Disentangling ethnicity, minority status, and achievement. Journal of Educational Psychology. 109. 10.1037/edu0000156.
[51] Martinez *et al* (2009)
[52] Hoge & Coladarci (1989)

[53] Demaray & Elliot (1998)
[54] Hopkins *et al* (1985)
[55] Tarricone, P. &Newhouse, C.P. (2016). Using comparative judgement and online technologies in the assessment and measurement of creative performance and capability. Int J Educ Technol High Educ 13, 16. https://doi.org/10.1186/s41239-016-0018-x
[56] McMahon, Suzanne & Jones, Ian (2015) A comparative judgement approach to teacher assessment, Assessment in Education: Principles, Policy & Practice, 22:3, 368-389, DOI: 10.1080/0969594X.2014.978839
[57] Harlen (2004)
[58] Heldsinger S, Humphry SM (2013) Using calibrated exemplars in the teacher-assessment of writing: an empirical study. Educational Res 55(3):219–235
[59] *Reinforcement Learning*: An *Introduction*. Second edition, in progress. Richard S. Sutton and Andrew G. Barto c 2014, 2015. A Bradford Book. The MIT Press.
https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf
[60] Hoare, C. A. R. (1961). "Algorithm 64: Quicksort". Communications of the ACM. 4 (7): 321. doi:10.1145/366622.366644
[61] Hoare, C. A. R. (1961).