AlphaPlus

Statistical grade boundary setting approaches

Literature review for the IB

December 2021

Contents

1. Executive summary

This literature review forms the first stage of a project aiming to review and improve the International Baccalaureate's (IB's) Statistically Recommended Boundary (SRB) setting procedures. The ultimate goal of the project is that, ideally, SRBs would provide an accurate estimate of where grade boundaries should be that rarely needs adjusting (or at least, needs more minor adjustments applying than current SRBs do).

In light of this, this literature review aims to accomplish the following:

1. Map out the 'universe' of statistical standard setting procedures, including:
    a. Any requirements for them being able to be utilised
    b. Any advantages and disadvantages relative to other approaches
2. Make initial judgments as to which procedures might be most suitable or unsuitable for the IB's contexts

A wealth of literature was reviewed to gather information on the statistical standard setting methodologies in use. Broadly they fall into one of two categories, score equating and prediction-based approaches. We summarise the techniques covered by each category, including but not limited to:

- Score equating
    - Basic equating techniques (mean, linear, equipercentile, etc)
    - Smoothing techniques
    - Nonequivalent groups designs
    - Item response theory
- Prediction-based
    - Ways of deriving a prediction
    - External indicators of cohort differences
- Ways of combining multiple approaches

Based on this, we draw some initial conclusions about the standard setting approaches likely to be viable (or not) for the IB's contexts. Prior attainment-based approaches seem infeasible due to a lack of such information, whilst nonequivalent groups designs are also likely impractical since anchor items would compromise the security of IB's assessments.

This leaves, to generalise, three broad approaches which seem promising for the IB:

a. Basic equating techniques
b. Concurrent attainment approaches
c. Approaches seeking to maintain the prior outcome (i.e., via common centres)

Basic equating techniques, as a whole, are suitable in situations where the two cohorts are comparable in ability. However, only around half of IB's contexts meet this assumption. It is worth noting that basic equating approaches *can* be applied in almost any circumstance (they need only a small sample size), which might mean that in some cases they are the only viable option. The question is whether it is advisable to do so (i.e., if cohorts are likely to be dissimilar), or whether it would be preferable to rely on judgemental approaches alone.

Concurrent equating approaches like the Instant summary of achievement without grades (ISAWG) are powerful, and are suitable for the IB's programmes due to their featuring a broad range of subjects. Further, it offers (by some margin) the most convincing equating approach for some of the most awkward contexts, including very small subjects, those with complete cohort change, and completely new subjects. However, ISAWG approaches are extremely complex, with a huge wealth of available options and modifications (even when compared to the other approaches in this

paper). It seems likely that ISAWG would be a method that *can* offer solutions for IB's most challenging contexts, but would require a substantial amount of effort to adequately trial and implement it – effort which might be disproportionate to the benefits it offers. The approach also has other drawbacks, being tricky to implement and a black box in terms of ease of explanation to laypersons.

Approaches seeking to maintain the prior outcome for a subset of the cohort (such as the 'common centres' approach) are a well-established means of attempting to account for cohort changes that is viable as long as there is a large enough cohort, and sufficient centres taking the subject from one year to the next. Whilst not as strong of a method as prior attainment for maintaining outcomes, it is still superior to many other approaches as it aims to account for any change in cohort ability over time. It is also appropriate in just about all of IB's contexts, with the exception of very small cohorts and completely new subjects (though there is the possibility of using common centres to link to a similar existing subject, dubious as this may be).

Later stages of the project can draw upon this review to determine the approaches which are worthwhile carrying out further modelling on to evaluate their appropriateness for the IB's varied awarding contexts.

2. Definitions and terminology

Before delving into the specifics of individual statistically driven standard setting approaches, it is important to set the scene and discuss key terms and concepts.

2.1. Criterion- and norm-referencing

Standard setting is the process of assigning thresholds, cut-scores or grade boundaries to assessments, with the aim of ensuring that only the candidates that deserve to achieve a particular grade or status are allocated it. There are, broadly speaking, two different paradigms for setting standards in assessment (Bond, 1996), which we initially define to contextualise the subsequent discussion.

The first is criterion-referencing. In this paradigm, a standard is linked to a particular criterion (or several criteria). This approach is well-suited to vocational or practical qualifications, and is often used in such cases. For instance, in a practical medical assessment a doctor may be required to be able to: administer CPR, accurately measure medicine volumes, and carry out procedures like injections and catheter removals.

From another point of view, in a criterion-referenced system a candidate must be able to demonstrate competence in the criteria in order to be awarded a particular status or grade. Note that whilst many vocational and practical assessments are pass/fail in nature, this is not an inherent feature of criterion-referencing. It is possible for the criteria for different grades to relate to different levels of competence in less binary ways, such as perhaps “carry out an injection safely”, as compared to “carry out an injection safely whilst ensuring the patient is comfortable”.

The second key paradigm is norm-referencing. In this paradigm, a standard is linked to a group norm, such as ‘the top 10 per cent of the national cohort’ – who might be allocated the top grade. The key distinction from criterion-referencing is that there is no explicit link to candidates’ level of competence in a norm-referenced system.

Norm referencing is often utilised in national systems where there is a need to stream or select candidates based on their level of performance but there are a finite number of places available in the next stage of education, as it helps ensure a relatively stable number of candidates achieve each status or grade. Ultimately however, most education systems and assessments which focus more on norms tend to use somewhat of a mixture of norm- and criterion-referencing, as it is generally undesirable to have no link to a level of competence whatsoever.

Understanding these concepts is important in order to understand how IB’s current approach and several other statistical standard setting techniques function. As mentioned below, the IB’s current standard setting approach is weak criterion-referencing, but its SRB setting methodology is an example of pure norm-referencing.

2.2. Standard setting and maintaining

In the literature a key distinction is between standard setting and standard maintaining. Standard setting is the process by which the threshold of passing (or achieving at a particular level) on an assessment is agreed upon – and this is an exercise that cannot be done statistically. Statistics and statisticians simply do not know what the appropriate level of knowledge (for instance) a qualified medical professional should have, or what level of competence is worthy of a grade 7 in Geography. Only subject experts can determine this, via standard setting procedures. A myriad of such approaches exist (Baird and Scharaschkin, 2002; Benton & Elliot, 2016; Black and Bramley, 2008; Bramley and Gill, 2010; Curcin et al, 2019), but the focus on specifically *statistical* standard

setting methods in this work means that only approaches which do not utilise any form of expert judgement or qualitative input are within the review's scope.

This means that this review can realistically only focus on standards maintenance activities, which are chiefly statistically based[1]. A key point to bear in mind is that this inherently means that an initial standard *must already be defined* within a given subject in order for statistical approaches to be able to function. If the IB was a new organisation just beginning to deliver assessments, it would not be feasible to use statistics to set the standard purely using statistics.

There are broadly two different schools of standards maintenance activities in terms of their underlying operation:

- Score equating approaches
- Prediction-based approaches

Score equating approaches function by working out what mark on one assessment is equivalent to another mark on a different assessment; equating one mark to another. In the IB's context, this would be "what mark on this year's assessment is equivalent to the grade boundary mark on the prior year's assessment" in most cases.

Prediction-based approaches still produce the same result – a mark on one assessment that is notionally equivalent to that on another, but do so via different means. They rely on some external measure of cohort ability to 'predict' how well the current cohort should do relative to the previous one, i.e. that 10 per cent of candidates should achieve a 7, 20 per cent a 6, and so on. Based on the proportion of candidates expected to achieve each grade according to the prediction, boundaries can be set to best 'meet' the predictions. The IB's current norm-referenced SRB setting procedure is a prediction-based method, with the inherent assumption that the cohort's ability is unchanged relative to the prior year (and thus that similar proportions of each grade should be awarded).

2.3. Classical test theory and item response theory

There are broadly two dominant theories of psychometrics (the science of measurement) in an educational assessment context. We define them here as some of the statistical standard setting techniques described in this paper are founded in each theory.

Classical test theory (CTT) is, as the name suggests, the 'original' theory of measurement codified by Lord and Novick (1968). It is founded on the belief that a candidate's mark on an assessment has two components; a 'true score' and an error component. In other words, if a candidate took the same assessment day after day (with their memory somehow wiped between) they would not always achieve the same mark – this is the error component around their true score coming into play. Readers are likely to be familiar with some of the key statistics arising from CTT, even if they do not realise they are CTT statistics – item facility indices, discrimination indices, and Cronbach's Alpha.

CTT is mark focused – a candidate's overall ability level is measured by their total mark on the assessment, which makes disentangling it from the error component very challenging. Being mark focused also leads to arguably its biggest shortcoming – that it is not possible to disentangle

[1] Note that whilst one scenario on the prior page "new subjects" generally lacks statistical data, there are a very limited number of approaches which can be used to statistically carry forward the standard from other comparable subjects into a new qualification. Arguably this is standard setting, but in all other cases only standards maintenance is considered in this paper.

candidate ability from item difficulty. For instance, if an item is responded to correctly by many candidates, is it very straightforward, or are the candidates just very able?

Item response theory (IRT) by contrast, does away with raw marks and moves both measurements of item difficulty and candidate ability onto a different scale[2] (Rasch, 1933). Key to the theory is the relationship between the two this establishes – a candidate with a given level of ability has an exactly 50 per cent chance to get an item of the same difficulty level correct. This focus on item rather than overall assessment performance is what gives IRT its name.

Crucially, through placing these two metrics onto the same scale (but not a raw mark scale), IRT disentangles candidate ability and item difficulty, meaning psychometricians can definitively say whether items are straightforward or candidates are able. This is a key advantage for many test equating scenarios where it is useful to be able to establish what marks on one assessment form are equivalent to marks on another form. Though many techniques exist which accomplish the same in a CTT framework too, IRT inherently lends itself to test equating and other scenarios, and permits some specific approaches not possible in a CTT framework that might be valuable in some situations.

[2] Known as a logit scale, and often referred to as theta in IRT publications.

3. Context

3.1. IB's current standard setting procedures

The International Baccalaureate (IB) is a major international nonprofit foundation which offers a suite of educational programmes to students aged between 3 and 19. These educational programmes are alternatives to "in country" programmes, with their own curricula and assessments. As a result, one of the myriad roles for the IB in their programmes' running is in setting and maintaining the standard of these assessments, in order to ensure fairness and comparability from year to year. IB's standard setting is done via a process called 'grade awarding'.

Historically the IB's grade awarding model can be described as "weak criterion referencing". In other words, a balance of the criterion (competence in the domain at hand) with how candidates have performed in prior years is used to set the standard. The following diagram (IBO, 2018) shows the three key sources of information that input into IB's grade awarding process.

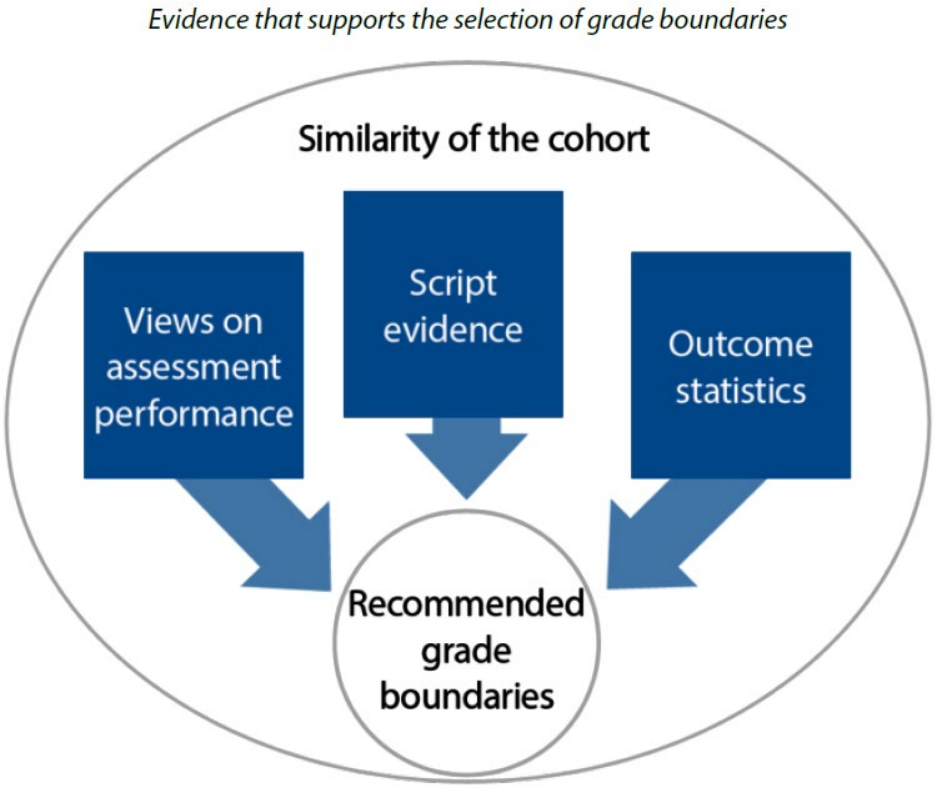

Figure 1: IB's current inputs into grade awarding activities

The first input into grade awarding is 'outcome statistics', which uses data from the prior and current session to ask the question "If the prior cohort had sat this year's exam, what grade boundaries would be needed to maintain the same overall grade distribution?". This input is entirely quantitative in nature, and is used to derive statistically recommended grade boundaries (SRBs) for key judgemental grades (3, 4, and 7). Notably, this input is actually norm-referenced as opposed to criterion-referenced, as in the absence of other inputs would result in a maintenance of outcomes from year to year.

The criterion-referencing element of IB's grade awarding is introduced by the other two inputs into the process, views on assessment performance from key personnel, combined with evidence from

scrutiny of candidate scripts. These qualitative inputs are used by the awarding committee to decide whether the difficulty of the assessment and/or the ability of the cohort has changed relative to last year, and to adjust the SRBs to generate final grade boundary positions.

3.1.1. Issues with current SRB setting procedures

The current norm-referenced SRB procedure is reliant on two factors remaining stable over consecutive session in order to be completely valid: a) the difficulty of the assessment remaining constant relative to the prior session, and b) the ability of the cohort remaining constant relative to the prior session. Whilst it is likely that in many of the IB's awarding contexts this is the case, in many others it will not always be – and further to this, in some contexts there will be no prior session to benchmark against.

When the assessment's difficulty and the cohort's ability are not stable over time (or there is no prior session to refer to) the current purely norm-referenced SRB setting method ceases to provide the best possible estimate of where grade boundaries should be placed. Whilst a comprehensive review of the IB's awarding contexts forms a later stage of this project, an initial scrutiny of some awarding contexts the IB faces makes plain that these two elements remaining stable is not the case in many contexts. Below we list some common contexts and how they violate these assumptions, leading to SRBs potentially being inaccurate.

1. Large stable subjects (whilst large cohorts are likely to be stable, even the best constructed assessments tend to vary in difficulty slightly over time)
2. Small subjects (small cohorts are inherently less stable in ability over time)
3. Growing subjects (the 'new' centres starting a subject are likely to cause a shift in cohort ability profile over time)
4. Changing curriculum or assessment models (the assessment's difficulty may change with a shift to a new model; the cohort may also initially struggle with a new assessment reflecting a drop in effective ability)
5. New subjects (in these cases there is no prior data on which to base SRBs)

3.2. Aim of this literature review

This literature review forms the first stage of a project aiming to review and improve the IB's SRB setting procedures. The aim of the project is that, ideally, SRBs would provide an accurate estimate of where grade boundaries should be that rarely needs adjusting (or at least, needs much more minor adjustments applying than current SRBs do).

In light of this, this literature review's aims are as follows:

3. To map out the 'universe' of statistical standard setting procedures, including:
    a. Any requirements for them being able to be utilised
    b. Any advantages and disadvantages relative to other approaches
4. To make initial judgments as to which procedures might be most suitable or unsuitable for the IB's contexts

Later stages of the project can then draw upon this review to determine the approaches which are worthwhile carrying out further modelling on to evaluate their appropriateness for the IB's varied awarding contexts.

4. Score equating

As outlined above, score equating focuses on converting marks on one assessment form into the scale of another (or sometimes onto a common scale). There are a myriad of ways to do this. In this section we will outline various score equating methods, starting with the most straightforward and moving on to more complex ones. The many equating techniques are however often overlapping and somewhat confusing; we have aimed to lay out this section in as logical a fashion as possible, and not to overwhelm the reader with formulae or very technical details.

4.1. Basic equating techniques

4.1.1. Mean equating

The simplest form of score equating, mean equating assumes that there is a constant difference between two assessment forms across the mark scale. I.e. a mark of five on one form equates to seven on the second form; a mark of 42 on the first would also equate to 44 on the second. As the name suggests, the constant shift applied in mean equating is determined by the difference between the means of candidate marks on the two assessment forms.

Plainly, this is an extremely simple form of equating; the operation is a simple addition or subtraction – this is its main advantage. As readers are likely already conscious of however, there are many assumptions inherent in such an approach. The main assumption is that the difference in difficulty between two assessment forms can be described with a single constant. I.e. that it is not the case that five actually equates to six, and 42 to 44.

Secondly, it assumes that the ability of candidates sitting each assessment form is identical. If the same cohort is indeed sitting both forms being equated then this substantial assumption can be eliminated, but this is often not the case.

Thirdly, it is largely only applicable when the total number of marks of both assessment forms being equated are very similar (if not identical). This is because converted marks at the extremes of the range will exceed the limits of the marks available on the assessment in mean equating (i.e. if the max mark on both forms is 100, then a mark of 100 on the first form would equate to 102 on the second). This is a minor problem in and of itself that requires some form of capping rule to manage – but if the total number of marks of one form were (for example) 50, the equating exercise would be completely invalid, as half the marks on the higher total form would have no equivalent on the lower total form.

4.1.2. Linear equating

With mean equating based on how the mean of the mark distribution on two assessment forms differ, the next logical extension of the method is to also account for the spread of marks around each mean. This essentially means that a constant difference between marks on two assessment forms is not required.

Technically speaking, this is accomplished using the SD of marks on each assessment form; marks +1 SD from the mean on each form are set as equivalent, marks -1 SD from the mean on each form are set as equivalent, marks +2 SDs from the mean are set as equivalent, and so on.

This approach is still relatively straightforward to explain and implement, and evades the key assumption of constant difference made in mean equating – which is important as the distribution of marks on different assessment forms is rarely identical in practice. However, it still falls foul of assuming candidates sitting both assessment forms are of equivalent ability, requiring capping of extreme values, and being inadvisable if the two forms have differing total number of marks.

Notably, for both linear and mean equating the greatest equating accuracy occurs near the mean (Kolen and Brennan, 2014) – rendering this a strong method for assessments with a single cut score that candidates tend to cluster around. For graded qualifications or vocational ones where the standard of candidates tends to be far higher than the cut score, equating accuracy near the mean is much less of a useful feature.

4.1.2.1. ANCOVA equating

A modification to the linear equating approach uses an analysis of covariance (ANCOVA) to attempt to mitigate a key assumption – that the candidates sitting each assessment form are of similar ability (van Onna, Jongkamp & Lamprianou, 2021). It does this through the use of background variables; some examples of which might be: gender, length of time in education, type of centre, and so forth. If a particular background variable is associated with stronger or weaker performance, then an ANCOVA can 'correct' for this.

This means that if (for instance) a particular centre type is associated with weaker performance, and more candidates from that centre type sit one assessment form, then we would expect lower marks on that form as a result – and in ANCOVA the equating relationship between forms will adjust to account for this. As such this approach is valuable in cases where there are known group differences in performance on an assessment (and data on group membership is readily available), and the proportion of candidates from said groups tends to vary from sitting to sitting.

There are however some drawbacks to utilising ANCOVA to try to account for differences in cohort ability. Firstly, it only achieves its aim if performance variations are actually explained by background variables available to the analyst. Secondly, it introduces a new assumption that the relationship between background variables and ability are consistent across both cohorts. Using the above example, if the centre type that performs poorly on one form actually performs very well on the second, then the approach would be invalidated.

The main drawback however is an ethical one; predicting groups' performance based on their group membership can be problematic. If for instance the above example did come to pass – a centre type we expected to perform poorly actually performing quite well, then we would have effectively deflated candidates' outcomes on the second paper unfairly. It is easy to see how this becomes problematic if more sensitive characteristics are used in the approach. The example of 2021's extraordinary awarding session in the UK, where the statistical approach used the centre candidates attended as a key factor in the grades awarded in the absence of exams, serves as a cautionary tale about the public perception of similar uses of background variables (Priestly et al, 2020).

4.1.2.2. Resit analysis

Another novel adaptation of a linear equating methodology is resit analysis (van Onna, Jongkamp & Lamprianou, 2021). Here, if there is a sufficient volume of resitters between assessment form A and B, these candidates being common to both assessment forms can be used to gain some information about the relative difficulty of the two forms. Again, this approach is useful in attempting to account for variations in the difficulty of forms, which is assumed in typical linear equating.

Here the assumption is that, in typical circumstances, candidates resitting an assessment will not (overall) have reduced in ability between sit A and sit B. They may not have progressed much, but they should not have fallen backwards in terms of overall competence (Covid-19 being a notable violation to this generally safe assumption). If we accept this, then one can consider the relative performance of the cohort on both assessment forms and factor this into the model to obtain a *lower limit* of the true difference in difficulty between the forms.

As an example, if the cohort achieved an average of 24 marks on form A, then resat and scored an average of 26 marks on form B, if we assume little progress has been made by the resitters then form B is two marks easier than form A. This is the aforementioned lower limit – we would conclude from the resit analysis that form B is no more than 2 marks easier than form A.

Obviously this is an approach which needs to be coupled with others, as it does not give a true estimate of equated values, just a bound for them. This is a drawback, but it does also mean that there are few other issues with this approach in and of itself; it is reasonable to use it to derive a lower bound, assuming there has not been a substantial slip in ability between sits. A more pragmatic reason for not using the approach is that often large numbers of resitters are not available in equating situations (and they may also be unrepresentative of the cohort as a whole).

4.1.3. Equipercentile equating

Notably, the relationship between marks on two assessment forms equated with mean or linear equating can be described with a straight line (mean equating changes the intercept, linear equating also adjusts its slope). Equipercentile equating differs in that it describes the relationship via a curve rather than a straight line. This means that assessment form A could, instead of consistently being harder/easier than assessment form B depending on the difference in the mean mark on each form, be harder at some points on the mark range but easier at others.

The way this is accomplished uses the percentiles of candidates that achieve a given mark to link them across forms; if 20% of candidates on form A achieve a mark of 12 or more and 20% of candidates on form B achieve a mark of 15 or more, then these two marks would be considered equivalent under equipercentile approaches. Challenges arise when percentiles do not exactly match (as is common with integer marks), but the method gets around this through treating the mark distribution as if it were continuous and effectively finding the closest percentile match for a given mark.

Using a curve rather than a line to equate has clear advantages. Consider an assessment form B which, relative to assessment form A, has 10 items far easier than any on A, 80 of comparable standard, and 10 far harder than any items in A. This is not an unrealistic scenario, and in such a case it would be valid to want to treat the lowest 10 marks on B as easier than those marks on A, but the highest 10 marks as harder. In a nutshell, the increased complexity of the relationship between assessment forms in equipercentile approaches allows for better modelling of possible scenarios in equating. It is useable when total number of marks differ, and prevents the capping issues present in mean and linear equating. However, a key assumption of the prior two methods is still present – that the ability of the cohort sitting each assessment form is equivalent.

It's also worth noting that whilst linear and mean equating might seem inferior to equipercentile equating, they are actually preferable within their niches, however unlikely these situations are to emerge in reality. If two assessment forms' distributions *only* vary by position (and spread for linear) then mean equating is ideal – not only does it follow the principle of Occam's razor, but it introduces less random error than equipercentile equating (Kolen and Brennan, 2014).

4.1.3.1. Circle-arc equating

Circle-arc equating is an unusual method of equating that does not quite fit into the structure of this paper. We discuss it here because it is a fairly simple equating method that is proposed as a potential alternative to equipercentile equating. Livingstone and Kim (2009) developed this method as an approach that, whilst lacking a theoretical underpinning like most approached, nonetheless produces results comparable to those of many other equating approaches.

Per the name, a circle's arc is drawn through three points on a graph of the marks on form A and form B. The first two points are the maximum and minimum marks of each assessment form, and

the third the midpoint of the assessment form (usually the average mark achieved[3]). The curve plotted serves to describe an equating relationship between marks on the two forms. A circle rather than a curve is used because it is computationally simpler and more closely represents the curves emerging from equipercentile equating with large groups. An example of this is shown in the Figure below.

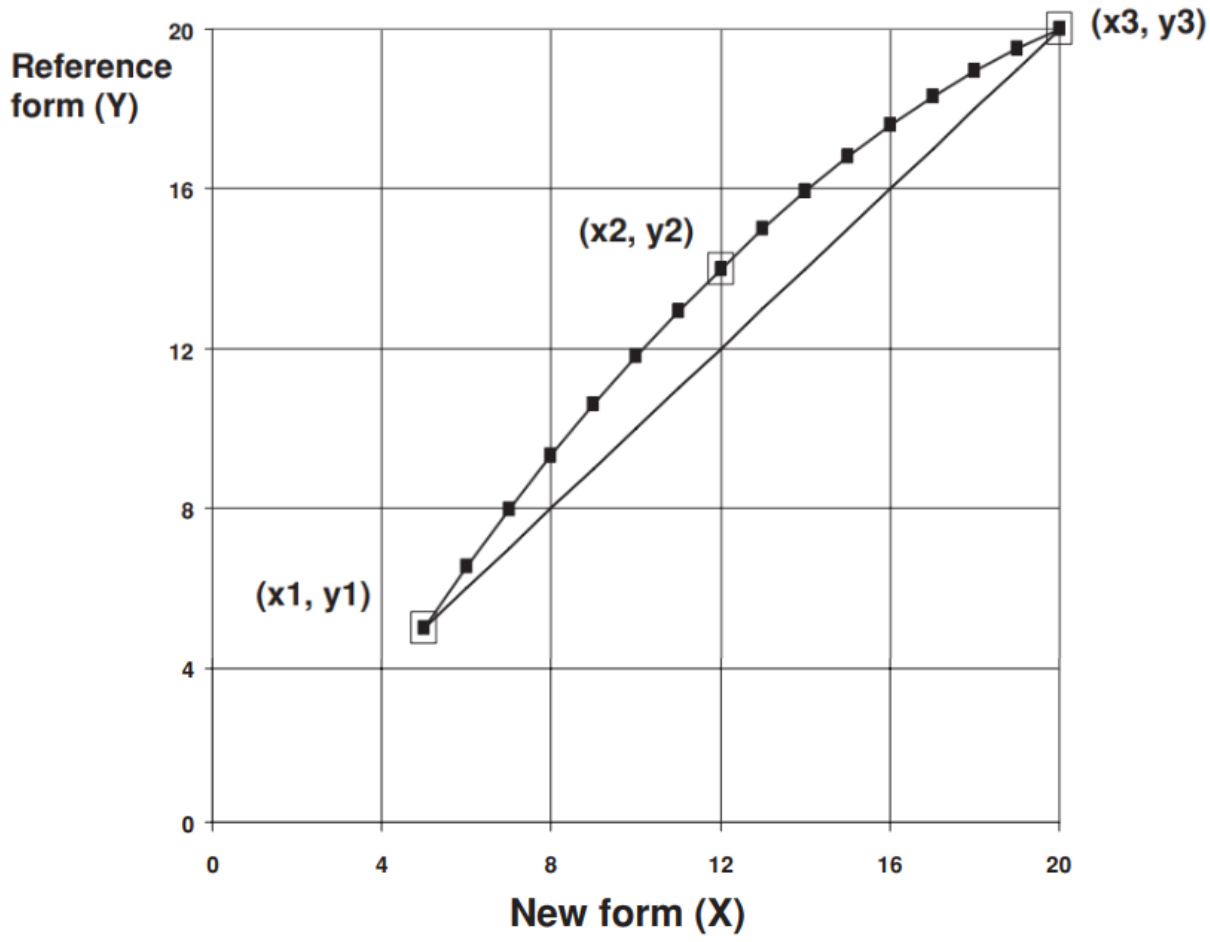

Figure 2: Example of circle-arc equating two assessment forms – from Livingstone & Kim (2009; pg 335)

This method has several key benefits. First, it is very robust with small samples, as the main datapoint underpinning it is the mean (like the mean equating method). However, unlike mean equating it is accurate in the upper and lower reaches of the mark distribution, a key advantage for assessments where accuracy across the range of marks is needed. And as stated above, it mimics the results of complex approaches like equipercentile equating but is much more straightforward to implement and explain; it has also performed well (and better than many other approaches) in operational settings (Livingston and Kim, 2010; LaFlair et al, 2017).

Its main downside is the aforementioned lack of theoretical underpinning. If one is asked to explain "why are we equating in this manner" there is not really a logical explanation to give other than "it happens that it works like other approaches, but it's easier". This does not detract from its accuracy, but does impact on its defensibility somewhat.

[3] In nonequivalent groups designs (see below) a chained equating approach must be used to equate the midpoints of the score distributions on each test form.

4.1.4. Visualisation of basic equating techniques

Understanding the nuances of various equating methods can be challenging via the medium of text. For this reason we reproduce below a figure from Albano (2016) which neatly visualises how each of the above methods links marks on one assessment form to those on another.

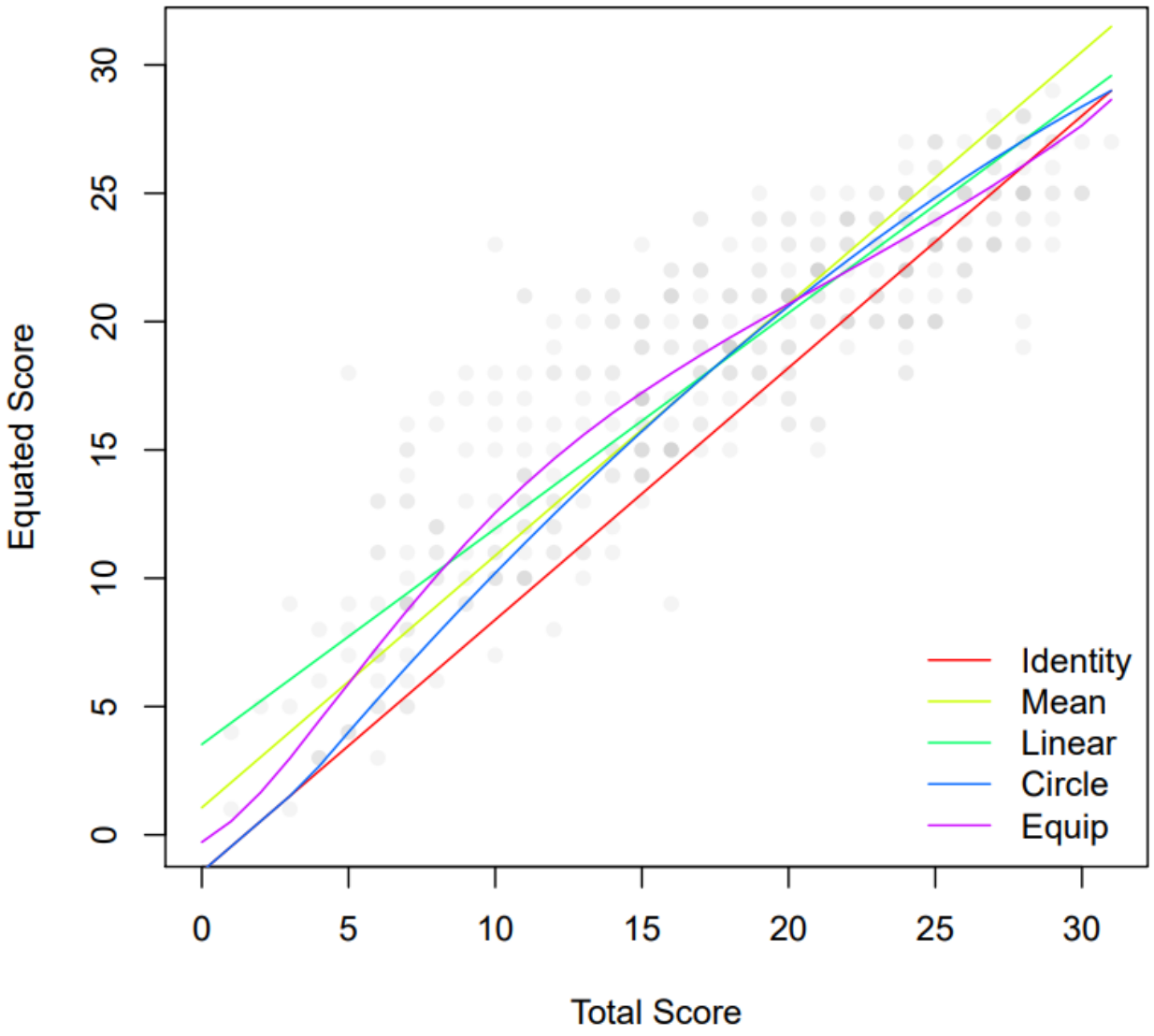

Figure 3: Example linking functions – from Albano (2016; pg. 22)

This graph can be read as showing a mark on one form (form A) on the x axis, and the mark on the second form (form B) on the y axis. The lines each show what different equating methods consider the marks on each form that are equivalent to one another.

The red 'identity' line simply represents the x = y line, i.e. considering one mark on form A equivalent to the same mark on form B. It is presented solely for the purposes of comparison with other approaches.

The lime 'mean' line simply adjusts the intercept of the identity line – shifting it up higher on the graph. The green 'linear' line however adjusts both the slope and intercept of the identity line.

The blue 'circle[-arc]' line, as outlined above, is a neat circle's arc. The purple 'equip[ercentile]' line however is a curve, allowing it to suggest that at the very highest marks on form A are no longer

equivalent to higher marks on form B (which can be observed where the purple line drops below the red one).

4.2. Smoothing techniques

One of the issues identified with equipercentile equating in the literature (Kolen & Brennan, 2014) is that even with large sample sizes (a few thousand), when the relationship between marks on form A and B is plotted, it can be quite jagged. This is attributable to only a sample of the population's data being available (even when working with assessments sat by full national cohorts we could consider "all past and future national cohorts" the population); presumably if we did have the full population's data, the relationship would smooth out in a similar manner to sampling increasing numbers of individuals from normal distributions.

Smoothing techniques are an attempt to mitigate this issue by using interpolation to 'even out' the unevenness in a distribution in an attempt to better approximate the population's statistics[4]. Of course, this is contingent on the assumption that the smoothed distribution does indeed better represent the population distribution – which we can never be certain of; in some cases maintaining the original uneven distribution may actually be better. An example of smoothing a mark distribution from Kolen and Brennan (2014) is presented below to provide visual reference.

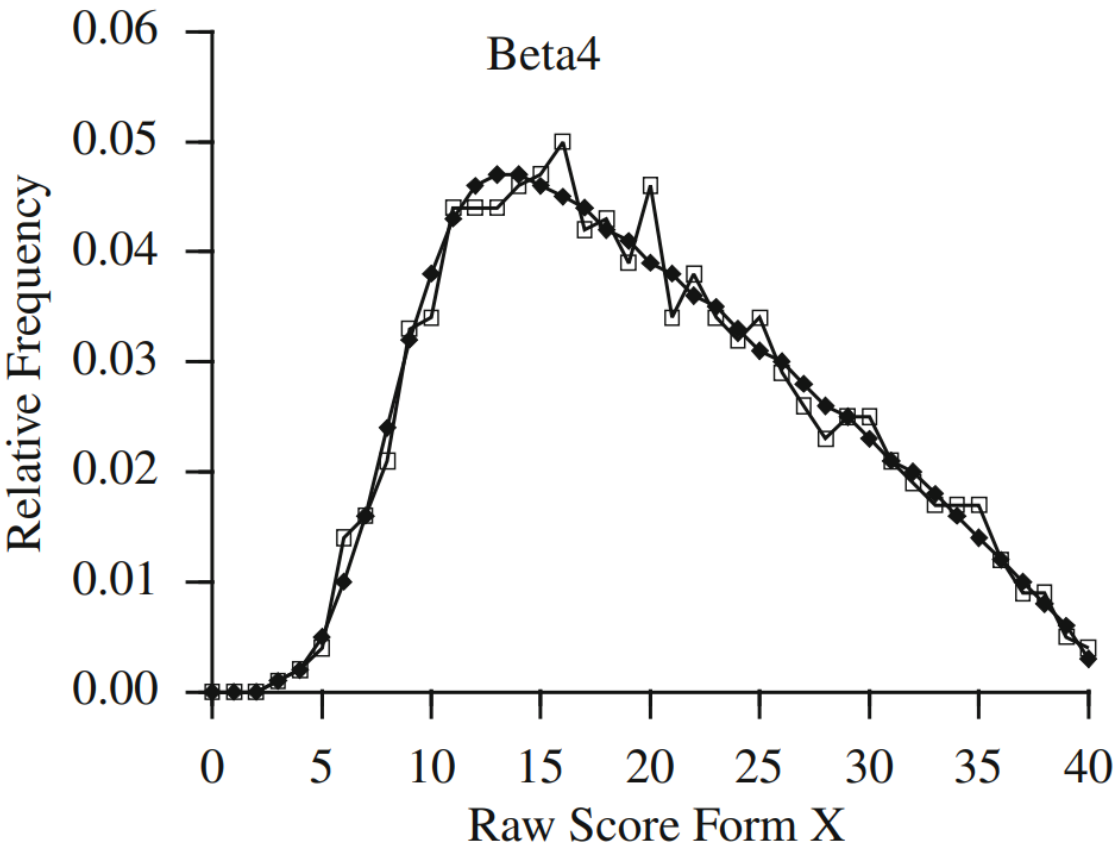

Figure 4: Example smoothing of a mark distribution – from Kolen and Brennan (2014, pg. 75)

[4] Note however that smoothing cannot resolve issues caused by equating based on an unrepresentative cohort; the smoothed distribution which emerges will always mimic the original unsmoothed one (Puhan, 2011).

In the figure, the actual mark distribution is shown with white squares, and appears quite uneven particularly in the 15-25 region. The black rhombuses show the smoothed mark distribution, which is much less 'jagged'.

Kolen and Brennan (2014) make the distinction between pre-smoothing and post-smoothing; smoothing the raw mark distributions, and smoothing the equipercentile distributions respectively. There are a range of different pre- and post-smoothing techniques described in detail in their book (chapter 3), which are broadly similar in outcome. One notable feature of pre-smoothing methods is that statistical tests emerge which aid the assessment of how useful the smoothing is. Broadly if smoothing is deemed of interest, it would be fruitful to model various approaches and assess which is best for the situation at hand.

We elect for the sake of brevity to summarise arguably the most common smoothing method in the literature, kernel equating, below.

4.2.1. Kernel equating

Kernel equating makes use of both pre-smoothing methods and post-smoothing ones (Holland and Thayer, 1989; von Davier et al, 2004). Notably, it smoothly interpolates between discrete mark totals, which helps in solving the issue of (for instance) a substantial proportion of the cohort stacking up on particular mark points, making it hard to utilise equipercentile methods to equate.

More technically, the approach initially uses a log-linear model of the "true" total mark distribution in the population to smooth out the unevenness in the sample data that is taken to be a result of sampling error. Next, it finds the probability (percentage values) of a random person drawn from the population having each mark. Then the discrete data is turned into continuous data, a process known as continuization, by drawing a gaussian filter over the data points to interpolate between them, allowing equipercentile equating to be used on this continuous dataset. As with most of the broad approaches detailed in this paper, there are many minor adjustments which can be made to a kernel approach that may improve its performance in some situations (Liang and von Davier, 2014).

This process is a very elegant solution to the discreteness of total marks, with the resulting outputs being quite intuitive (even if the method sounds complex). Arguably, most people would agree that it is better than a linear interpolation between marks. As outlined above, it is however difficult to say which of the various smoothing methods is 'best' – but kernel is certainly one of the most developed and popular in the literature, for what that is worth.

However, any form of smoothing (and particularly kernel) requires in-depth statistical knowledge to understand and specialist software to carry out, and is therefore unlikely to be scaleable to entire suites. Instead it seems like a potentially useful tool for a handful of subjects which would benefit from smoothing the most.

4.3. Nonequivalent groups

As should be apparent from our discussion of the three most basic equating approaches, mean, linear and equipercentile, all fall foul of the assumption that the cohorts sitting both forms are equivalent in ability. This is a major issue for many examinations, particularly those high-stakes ones which are used for onward selection where failing to maintain the correct standard is a fairness issue (Alberts, 2001). Nonequivalent groups designs, as the name suggests, use a range of techniques to eliminate this assumption, and as such are invaluable in situations where one cannot be sure of each cohort's ability being comparable (Kolen and Brennan, 2014). The distinction between 'linking' and 'equating' used in some literature comes into play here; linking is typically used to refer to equating situations where group A and B's abilities being comparable is

assumed (the approaches discussed in prior sections could therefore be termed linking approaches).

All are underpinned by the use of 'common items' – some set of items sat by both the distinct group of candidates who took form A and the group instead sitting form B. These can be either embedded within the paper itself or sat as an external anchor assessment. How differently cohort A and B perform on the common items is used as a proxy for differences in their overall ability, thus allowing nonequivalent groups approaches to address the assumption of equivalent ability by taking this into account during the equating process. How each method accomplishes this varies slightly, however.

The use of common items does however introduce a key assumption – that performance on the common items accurately reflects performance on the non-common items in each form being equated. In other words, if cohort A does worse on the common items than cohort B, it is presumed that cohort A is weaker and will score lower on the non-common items than cohort B would have. If however, cohort A is actually stronger overall but just has particular difficulty with the specific common items, then this assumption will be violated. The slightly concerning factor here is that there is no quantitative way to tell whether this is the case in a nonequivalent groups design (qualitative scrutiny might flag up unusual performance patterns).

As a result, it is of critical importance to try and mitigate the risk of this occurring. The best practice for this (Cook and Paterson, 1987) is essentially to ensure that the common items are as representative as possible a sample of the assessment as a whole; they must sample the same content areas, not just focus on one or two; they must sample the full breadth of item types included in the assessment; they should be spaced at intervals throughout the assessment(s) and in comparable positions in each of the two forms being equated; and there must be sufficient common items to reliably form a link between the two forms[5].

It is also important when using common items to review their performance and consider de-anchoring any which behave particularly differently between the two cohorts (i.e. treat them as if they were different items rather than as an anchor). For this reason it is useful to have a reasonable number of anchor items to manage any such attrition – Kolen and Brennan (2014) suggest 20% of the assessment length for assessments of 40 items or more, or a minimum of 30 items if this is lower than 20% for very long assessments. However this is a lower limit – if more items than this are needed to meet the above criteria of content and item type sampling within the anchors, then a higher proportion should be used.

Being able to use concurrent items is also not always a given, particularly in high-stakes assessments where any exposure of items is undesirable. In these situations one may need to fall back on a more simple equating design such as those outlined above, or utilise one of the prediction-based approaches discussed below.

4.3.1. Tucker, Levine and Braun/Holland

The Tucker, Levine, and Braun/Holland equating approaches are all discussed together in this section, as they can be used in similar circumstances and accomplish similar things. These three approaches are applicable to mean, linear and circle-arc equating (with linear being the situation they are most commonly applied) – but not equipercentile – and can reasonably be characterised

[5] Some recent work by Furter and Dwyer (2020) does however suggest that with IRT methods (see below) it might be preferable to have more anchor items at a similar same level of demand to the cut score – in order for IRT models to be have more 'information' about the items at that crucial region.

as 'variant approaches which facilitate nonequivalent groups designs'. Kolen and Brennan (2014) offer some guidance on the various micro reasons one might elect to pick one rather than another.

The Tucker approach (Gulliksen, 2013) is a regression-based approach, and as such assumes a linear relationship between marks on each assessment form; if this is violated it may not be appropriate and the Braun/Holland (Braun, 1982) approach is likely to be more valid. There are several sub-variants of Levine approaches (Levine, 1955), with Levine observed score equating notably being more suitable if there is a difference in ability between cohort A and B. Levine approaches however must see marks on both forms correlate well in order to remain valid. There are many further statistically convoluted tweaks that can be applied to these approaches (for instance, Chen and Holland, 2010, who integrated kernel smoothing with the Tucker and Levine approaches).

One notable feature of the Tucker approach is that it can accommodate very small samples of as few as 20-80 candidates with some modifications (though circle-arc equating in a nonequivalent groups setting performs similarly; Babcock, Albano and Raymond, 2012).

4.3.2. Chained equating

'Chained' is one of the most commonly used terms to describe equating, and can be applied to a few different approaches we have discussed already. In short, chained equating modifies other equating approaches by carrying it out in several steps, which is possible thanks to the use of common items in nonequivalent group designs. First marks on form A's unique items are linked to marks on form A's common items (or to the anchor assessment those sitting form A took). Then the common item marks for the population sitting form A and form B are equated[6]. Finally, these equated marks on the common items are again converted into marks on form B itself. Each equating step is visualised in Figure 2 below.

Figure 5: Visualisation of chained equating

The two equating approaches where chaining is most frequently applied are chained linear and chained equipercentile (often abbreviated to CEPE) – though chained mean and circle-arc are also possible (Peabody, 2020). As we have already discussed the basic versions of these approaches above, we will not repeat that explanation here; suffice to say that chained linear equating equates the means and SDs of marks on each of the four elements in Figure 2 above in turn, whilst CEPE uses the percentile ranks of candidates on each element as the core of its approach.

Notably, there are two key differences between CEPE and the other nonequivalent groups approaches (chained linear included) in terms of their outputs. Firstly, CEPE inherently has higher estimation error than the other approaches, which would generally mean it is less 'good' at finding the most valid equated set of marks between two assessment forms. However, the second factor

[6] Oh and Moses (2012) investigated whether equating the form B scores to the form B common items (rather than the inverse way round) had any impact on the result when using CEPE; it was established that inversing the direction of this final step made almost no difference.

is that the other approaches assume that the two cohorts being equated are similar in ability[7] (somewhat ironically, given that a main benefit of nonequivalent groups designs is to allow for accounting for changes in ability). When similar group abilities is not the case, CEPE tends to more than compensate for its higher native estimation error and render it the superior method.

Kolen and Brennan (2014) offer the following rules of thumb for judging when each group's abilities are different enough that CEPE becomes preferred. Mean differences between groups of .3-.5 SDs or more on the common items are troublesome for most approaches, as are cases where the ratio of group SDs on the common items is <.8 or >1.2.

As a result, the decision to use CEPE rather than other approaches comes down to the question "are the cohorts likely to differ markedly in ability?" If yes, then CEPE is the best option of the nonequivalent group designs. This becomes tricker when an entire suite would prefer a standardised approach – a factor the literature rarely accounts for.

Notably, CEPE is specifically utilised for cross-tier equating purposes in GCSEs in the UK, whilst the vast majority of standard setting is conducted via a different approach we will discuss at length later (Ofqual, 2017). This is both due to it being one of the few situations where there are common items between assessment forms in the GCSE suite, and tiered exams being a textbook example of cohorts with very different ability, rendering many other approaches inadvisable.

4.4. Item response theory

Item response theory (IRT) is an alternative framework for assessment statistics (as opposed to classical test theory which underpins the models discussed thus far). Its key feature is that it uses an iterative procedure to estimate both item difficulties and person abilities simultaneously, but both are placed on a common probabilistic (logit) scale rather than expressed with relation to raw marks. One of the reasons to utilise IRT is that placing items onto a common logit scale naturally aids in the equating of assessment forms to one another – or crucially, to an item bank, which is not possible with classical approaches.[8]

With all IRT analyses, the strong assumptions of IRT itself must be upheld by the data in order for the model to be valid[9] (Kolen and Brennan, 2014). Notably, the construct must be unidimensional; there must be a single underlying 'ability' which can be used to accurately model any candidate's likelihood of getting each item correct. The second key assumption is of local independence; candidates' responses to any one item should not be contingent on another – which is likely to be violated in for instance exams where multiple items relate to the same stimulus or prompt. Similar assumptions to those discussed in section 5.3 regarding item functioning remaining stable across both forms also apply; we do not repeat those here for the sake of brevity.

Another factor flagged in Wheadon and Evangelidou (2008) is that whilst IRT approaches are well established in many situations, one that does not lend itself well to IRT modelling is assessments featuring items with a high total number of marks (often essay-style). Whilst variant IRT approaches exist which can handle items worth multiple marks (i.e. the partial credit and graded response models), items with total number of marks approaching 10 or so are increasingly prone to 'disordered threshold' issues which violate the IRT model's assumption that achieving each mark on the item must be more difficult than achieving the last. This means that assessments with items

[7] With the exception of Levine observed score equating, which is more suitable than most other nonequivalent groups designs for groups with ability differences (Kolen and Brennan, 2014).
[8] Another slightly more fringe scenario IRT equating has a benefit in is, with small cohort sizes, when more than two forms are available. Babcock and Hodge (2020) found that in these cases (where data from multiple forms could be pooled using a concurrent calibration), Rasch equating outperformed the other alternatives the authors modelled.
[9] Note that the same could be said of many other approaches' assumptions!

with a high number of marks available are generally unsuitable for use in IRT modelling, and as such IRT equating approaches.

As with other schools of approaches, there are a range of equating approaches possible under IRT, which we will outline below.

4.4.1. Common item

Common item equating is similar in approach to nonequivalent group designs, as it is underpinned by a reliance on both cohorts having sat the same items (though due to IRT enabling item banking, one cohort can have sat the common items a long while ago and allowed the 'banking' of their item parameters).

There are a few sub-variants of common item equating in IRT. These are whether a concurrent or stepped process is used (Kolen and Brennan, 2014). A stepped process is analogous to chained methods above, where first form A is calibrated onto the IRT scale, then form B separately calibrated – but 'fixing' the common items' parameters to those emerging from the calibration of form A. In a concurrent process, both forms A and B are calibrated simultaneously. Concurrent calibration is more computationally intensive and was therefore not preferred in earlier IRT equating applications, but is increasingly viewed as the superior option as it uses all the data available for common item responses to calibrate them. Notably, when an item bank has been set up a stepped process is typically used; if an item bank has been calibrated some time ago, and we need to link a new form's items in, then it is generally more sensible not to re-estimate the whole bank's parameters as this is done.

The other major variation is the use of true or observed score equating, two approaches used to move back from IRT parameters into units of marks (Lord and Wingersky, 1984). 'True' scores in the IRT context are never known; they are the 'expected mark a candidate with a given ability would achieve'. Observed score equating uses the IRT model to approximate the distribution of marks on each form by effectively simulating distributions of marks for candidates at different ability levels. These observed mark distributions can then be equated using equipercentile approaches. True score equating is much more computationally simple, but has a theoretical flaw in that it assumes the observed marks are true scores (though this becomes less of an issue as cohort sizes grow larger). True score equating is also less robust at the extremes of the mark range due to a lack of data. However, Lord and Wingersky (1984) found that both approaches produce quite similar results, so in some situations the difference at the extremes may be minimal.

There are many other relatively 'under the bonnet' variants and tweaks which can be made to IRT models that we omit both for brevity and their relatively minimal material impact on IRT common item equating. One that is worth mentioning, however, is the application of Bayesian uncertainty reduction techniques when estimating an IRT model (Birnbaum, 1969). In short, this approach is valuable with small sample sizes where the model cannot be reliably estimated. It uses estimates of item difficulty and/or person ability based on prior knowledge or expert judgement to refine the model (essentially, create a compromise between the expectations and the empirical model). Whilst helpful to make an IRT approach function in small sample settings, it is obviously dependent on the accuracy of the estimates used – in cases where unexpected or unknown changes in item difficulty or person ability occur (as was the case in the Covid-19 pandemic) then it is likely to produce misleading results.

Ultimately, the IRT common item approach accomplishes something similar to a CEPE approach, and follows similar logic (each cohort's performance on the common items is used to infer what performance would have been on the non-common items, and thereby deduce each cohort's ability spread). Ultimately, the reasons one would generally pick one approach over the other are down to the theory being used to model the assessment outside of linking – it's a lot more trouble to link with CEPE if you are already using IRT, for instance. Obviously in situations where item banking is

a factor, IRT is necessitated, but more generally the choice of a classical or IRT-based common item approach does not have a clear 'better' option. It is worth noting however that the IRT approach essentially always includes a smoothing approach, whereas this is an additional consideration often necessitated in non-IRT nonequivalent groups designs (Lord and Wingersky, 1984).

4.4.1.1. Pseudo-anchor

One notable alternate approach to IRT equating is to use pseudo-anchor items rather than common anchor items (von Onna, Jongkamp and Lamprianou, 2021). This means identifying (via largely qualitative means given the new items will not have been calibrated yet) pairs of items across form A and B which are similar enough that we expect their psychometric properties to be extremely similar, and treating them as if they were the same item (i.e. giving the new item in B the properties of the paired item in A).[10]

This is an attractive approach in high-stakes assessments where using items multiple times is not desirable. However, it is quite risky to make the assumption that the paired items will perform identically – even "cloned" items with the exact same format but (for instance) changed quantities in a numeracy assessment often have quite different item parameters.

4.4.2. Common person

The notable alternative to common item equating which IRT facilitates is common person equating (Masters, 1985; Boone and Staver, 2020). In this scenario, in place of the analyst "fixing" the item parameters of anchor items in form B (because their difficulties and so forth have already been established when analysing form A), common person equating fixes the ability of candidates who sit both form A and B. This necessitates a change in design; instead of candidates sitting either form A or form B (with some common items between forms), candidates must sit both form A and form B (but no common items between the two are required).

The assumption of item stability above here changes to one of candidate stability; that candidates' abilities will be stable across sitting the two forms. This necessitates sitting both forms at a very close point in time (likely with counterbalancing so half the cohort gets form A and half form B first). This is arguably less likely to be the case than item stability; items generally do not change from paper to paper without outside intervention, but even if forms A and B are administered in quick succession fatigue may begin to affect candidates on the second paper, impacting their effective ability.

An IRT common person equating design, whilst an important novel approach to be aware of, is also arguably less likely to occur in assessment than many other designs. However, if for instance it was desired to calibrate all the various exams comprising for instance, a Mathematics GCSE onto one scale, a common persons design would be a reasonable approach. This could aid in identifying whether the standard of different papers were comparable, for instance, but would be of little use in helping to maintain a year-on-year standard.

[10] Note that in principle, a similar pseudo-anchor approach could be utilised without the use of IRT in chained equating approaches; in the literature the use of pseudo-anchors is most commonly coupled with IRT however.

5. Prediction-based

There are several key elements to any prediction-based approach. Firstly the method used to derive the prediction, and, in almost all cases, the external indicator of cohort differences used to derive said prediction. This section deals with, initially, the means of generating predictions, and then the various metrics commonly used to accomplish this.

5.1. Ways of deriving a prediction

5.1.1. Maintain prior outcome

In a **maintain prior outcome (MPO)** approach, a prior cohort's outcomes are directly used as the prediction for the current session's outcomes. Broadly there are several possible variations on the theme of MPO approaches:

1. Maintain the outcome achieved by the entire prior cohort
2. Maintain the outcome achieved by a subset of the entire prior cohort
3. Maintain the outcome achieved by several prior cohorts aggregated together

The first option is the IB's current SRB setting methodology and can be characterised as the default MPO approach; the whole prior year's cohort's outcomes are used as the prediction. It is underpinned by a very straightforward assumption, in that all other things being equal, one session's cohort should achieve similar outcomes as the next. This is certainly a necessary starting point for any standards maintenance approach, and underpins all the other approaches discussed in this paper. As outlined in the discussion of IB's current approach above however, the issue with the approach is that this is all it factors in – it does not utilise any evidence of cohort ability or assessment difficulty change. Whilst IB's grade award approach does qualitatively factor these elements into standard setting, there is no reason why an SRB could not also aim to account for these issues, easing the burden on experts and guiding their adjustments to SRBs based purely on maintaining the prior outcome.

5.1.1.1. Maintain a subset of the cohort's outcomes

This variant of the 'maintain prior outcome' option is generally driven by concerns about the stability of the cohort, in situations like substantial cohort growth from year to year. Typically a subset of the cohort perceived to be stable from one year to the next is selected, and their outcomes used as the prediction for the same subset of the current year's outcomes. For instance, if IB opened up a qualification to a new region, then it might be prudent to generate a prediction of this year's outcome for the regions that have historically been sitting the qualification, which would be used to maintain just the historic regions' outcomes. This approach means that the "non-stable" element of the cohort is entirely excluded from the process of setting grade boundaries, and therefore that if they are indeed considerably more or less able than the existing cohort, that the current standard is maintained.

The challenge with this second approach is in identifying which subset of the cohort is "stable" vs not stable. Without substantial change as per a new region's introduction above, it is extremely challenging to do this. The solution much of the literature has settled upon identifies **common centres** between one session and the next, and uses them as the subset (Pinot de Moira, 2019). In other words, just the centres who enter candidates in both sessions are used to generate predictions and set grade boundaries. This approach is therefore useful when there is change in the centres that are entering for a qualification over time, and there is concern that the cohort's ability might be shifting over time. Notably, Pinot de Moira (2019) found that common centres was

preferable to equipercentile approaches[11] for exactly this reason; it aims to control for changes in the cohort ability over time.

The key assumption of a **common centres** approach is therefore that the subsetted centres' candidates will, on aggregate, will be of similar ability from one session to the next. This is often quite a risky assumption to make, as research on intra-centre volatility in results demonstrates that any individual centre's cohort ability will likely not be particularly stable from one year to the next – after all, whilst facilities, practices and teachers persist, the students themselves will be completely different. Common centres (and indeed any other MPO approach) is therefore contingent on there being a substantial enough subset of common centres from session to session to smooth out the volatility in individual centres. However, by the very nature of subsetting the cohort reducing the sample size, common centres is at greater risk of falling foul of this volatility than an 'entire prior cohort' prediction.

A notable variant of common centres approaches is **stable common centres** (Pinot de Moira, 2019). In this approach, only centres with a similar number of candidates in the reference and current session are included in the subset. It aims to eliminate centres whose outcomes from one year to the next might be volatile (due to their entry changing substantially) – and this logic is certainly consistent with that the common centres approach is based upon, that of identifying a stable subset of the cohort in each year. However, it does result in further attrition of the data above and beyond that of common centres, and as such is at greater risk of the benefits of identifying a consistent subset of the cohort being outweighed by the increased error inherent in trying to maintain outcomes with a small and more volatile dataset. The below figure demonstrates the high attrition in centres which takes place when imposing even moderately stringent constraints on stability – note that this is from the English system, which may be more stable than the IB's entry.

[11] Specifically, ones not featuring common items a nonequivalent groups design, as there are no common items from session to session in GCSEs and A-levels.

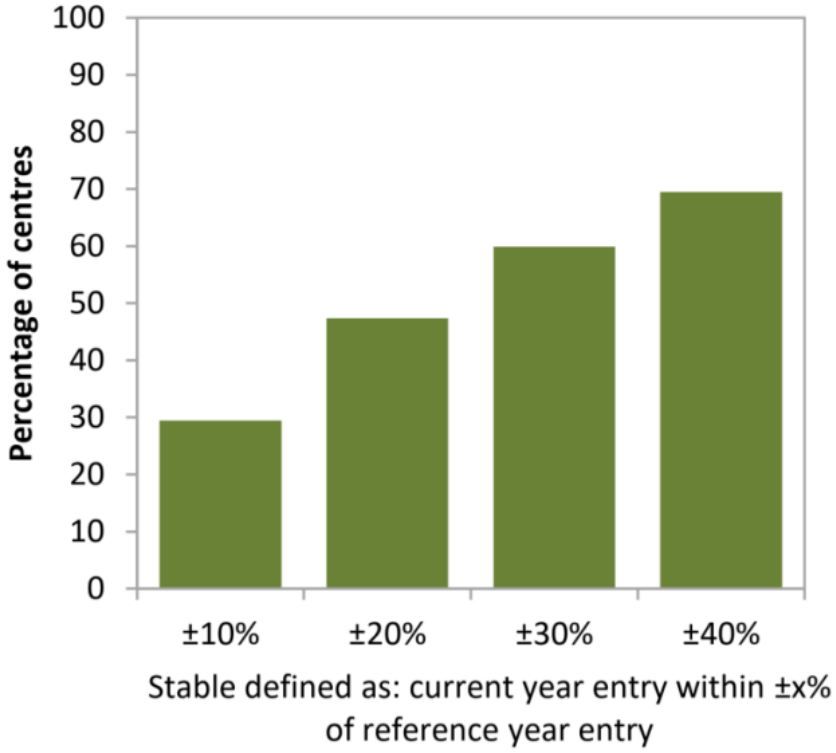

Figure 6: Proportion of English common centres treated as stable with varying stability constraints – from Pinot de Moira (2019; pg 2)

It is also worth noting that common centre approaches are not the only possible methods by which a stable subset of the cohort could be identified, just the most prominent. Given that any subsetting technique is reliant on the information about a cohort available in awarding contexts, it is perhaps unsurprising that common centres is the most common such approach; in non-IB contexts centre and perhaps a few other variables like date of birth and gender are all the information available at the time of a grade award. If more information such as demographic data were available about a cohort, then **matching techniques** (i.e. Mahalanobis distance or caliper matching; Baser, 2006) could be used to identify a subset of the cohort that were similar based on demographic data rather than (or in addition to) based on which centre they attended.

However, whilst IB does have more demographic data available about its candidates than many other awarding organisations tend to, we would suggest that such an approach, whilst very useful for research purposes, might fail the test of public acceptability if utilised to maintain a standard. A matching-based approach would effectively be saying "based on historic candidates with similar demographic characteristics, this is how the current cohort is expected to perform". Tying stability to protected or somewhat sensitive demographic characteristics rather than centre attendance is politically unpalatable (Priestly et al, 2020), even if the approach would be statistically defensible (and potentially 'better' than common centres).

5.1.1.2. Maintain an aggregate of several cohorts' outcomes

The third approach is the other logical alternative to the others; instead of maintaining outcomes for the entire cohort of a subset of it, multiple previous cohorts' outcomes are aggregated together to generate a prediction. As might be apparent, a key advantage of this approach is that it increases the volume of data used to generate predictions, and therefore issues of volatility are much reduced. It is therefore particularly appealing for contexts with low entry numbers each sitting, but with a relatively stable cohort otherwise – and obviously, several prior session's worth of data to draw upon.

Notably, it is entirely possible to combine this approach with subset approaches (we will use the example of common centres), which can help to tackle their key drawback of attrition of data. For example, centres which have been sitting the assessment for the past three years might be used to derive the prediction. Whilst in situations of relative stability this can be valuable, in cases where centres drop in and out of a subject over time, this approach could plausibly result in a smaller cohort than common centres, thereby defeating the entire purpose of bringing in older data.

The main drawback of this third approach is that older data is time-distal to the current session; if there has been change in the cohort over time (perhaps due to a sawtooth effect; Cuff, Meadows and Black, 2019; Newton, 2020) or to the assessment, then there is a danger than including older data in the prediction might compromise the validity of the approach. Ultimately this is why this approach is not widely used; whilst in theory it sounds promising, the situations it is most valid in are those where any number of other approaches are adequate. Nonetheless, for some small but stable IB contexts, it could prove a useful extension to an SRB-setting approach.

5.1.2. Adjust prior outcome

The approaches discussed in the above section deal with exactly maintaining a prior outcome, for varying groups of candidates, and is therefore contingent on the cohort's ability being stable over time. Whilst subgroup-based MPO approaches attempts to mitigate this substantial assumption by limiting the assumption to "a subset of the cohort's ability remaining stable", another possibility is to use an external indicator of cohort differences from the reference to the current session to adjust a MPO prediction.

Exactly what these external indicators can be is a substantial topic in its own right and is discussed in detail in later sections, but initially it is important to understand the varying methods by which this adjustment of prior outcomes can be carried out. For the sake of the following examples, we can consider the external indicator to be an external assessment that candidates in both the reference and current years have sat, and that we know their results from.

The two methods by which prior outcomes can be adjusted are **prediction matrices** (Ofqual, 2017) and **logistic regression** (Benton & Sutch, 2014), and are explained in the subsequent two sub-sections.

5.1.2.1. Prediction matrices

In a prediction matrix approach, the reference cohort is split into Xciles (i.e. quintiles or deciles) based on their performance on the external indicator – in this example, the external assessment. For the learners in each Xcile, the proportion of them achieving each grade is then recorded. You may see for instance, the top Xcile achieve 20 per cent grade 7s, and 40 per cent grade 6s, whilst the next Xcile down achieves 16 per cent 7s and 35 per cent 6s, for instance. This percentage outcome data is typically presented in the titular matrix, with rows being Xciles and columns grades.

This process is then repeated on the current year's cohort – the same Xcile cutoffs as in the prior year are used to allocate the candidates to Xciles based on their external assessment results (meaning that there will not be an even number of candidates in each Xcile). The approach is then governed by the principle that, of the people in each Xcile, we should see the same proportion achieve each grade as in the reference year (i.e. 20 per cent of the top Xcile should achieve grade 7s, and so on). To accomplish this, the change in N per Xcile from the reference to current year is used to determine how many candidates we would expect to achieve each grade in the current year. More technically, the N of candidates per Xcile is multiplied by the percentages in each cell of the matrix, which are then summed over each grade to derive a prediction for the number of candidates in the current cohort who could be expected to achieve each grade.

An example of such a matrix used for English GCSEs (with prior attainment at KS2 level as the external indicator) is presented below for reference (Benton and Sutch, 2014).

Prior attainment distribution for reference year

K2LevG	%
<3.00	2.00
3.00	1.63
3.33	3.27
3.67	7.85
4.00	19.64
4.33	18.35
4.67	20.51
5.00	26.76

Prediction matrix, showing GCSE grade distribution for candidates with each level of prior attainment in reference year

	% awarded GCSE grade in reference year								
K2LevG	U	G	F	E	D	C	B	A	A*
<3.00	11.9	23.0	23.8	20.1	13.0	5.9	1.6	0.7	0.0
3.00	6.5	19.4	27.4	21.2	16.2	7.4	1.4	0.4	0.1
3.33	3.5	11.1	21.5	25.8	23.1	11.7	2.6	0.7	0.0
3.67	1.7	5.2	13.0	22.9	28.9	20.4	6.3	1.5	0.1
4.00	0.9	2.1	6.6	15.3	27.0	29.3	13.8	4.5	0.4
4.33	0.5	0.9	2.8	8.4	21.5	31.5	22.2	10.4	1.9
4.67	0.2	0.3	1.0	3.9	13.5	26.3	29.0	20.4	5.3
5.00	0.0	0.1	0.2	1.2	4.7	14.0	26.0	34.4	19.3

Prior attainment distribution for outcome year, by AO

K2LevG	AO1	AO2	AO3	AO4
<3.00	1.67	2.07	1.94	1.43
3.00	1.48	1.72	1.73	1.40
3.33	3.10	3.46	3.48	2.95
3.67	7.27	8.15	8.26	7.49
4.00	22.91	23.05	23.31	22.41
4.33	20.05	19.63	20.07	19.79
4.67	21.19	20.14	20.30	21.31
5.00	22.33	21.78	20.90	23.23

Predicted grade distribution for outcome year, by AO

	Predicted GCSE grade (%)								
	U	G	F	E	D	C	B	A	A*
AO1	0.9	2.1	4.8	9.4	17.7	23.8	20.2	15.3	5.9
AO2	1.0	2.4	5.1	9.7	17.9	23.6	19.7	14.8	5.8
AO3	0.9	2.4	5.1	9.8	18.1	23.7	19.7	14.6	5.6
AO4	0.8	2.1	4.6	9.2	17.5	23.7	20.3	15.6	6.1

Figure 7: Example prediction matrix approach – from Benton and Sutch (2014, pg. 11)

Working through this example, in the top left we note the proportion of candidates in each (in this case) octile of prior KS2 performance in the reference year. This is then used to complete the initial matrix in the top right – for example, 11.9 per cent of candidates who achieved less than a 3.00 on their KS2 go on to achieve a U at GCSE. We then (in the bottom left) repeat the first step to establish the proportions of candidates in each octile in the current year (outcome year in the figure)[12]. These are then mapped across to each cell in the prediction matrix; if there are 1,000

[12] Note that Benton and Sutch's (2014) work is complicated by accounting for the UK's multiple awarding organisations. In a single provider scenario, there would just be one column in the bottom left table and one in the bottom right.

candidates who fall into the first octile in the current cohort, we would predict 119 of them to achieve only a grade U in the current session, 230 a grade F, and so on. This process is repeated for the candidates in each octile, then the N of candidates allocated to each cell summed in each column and converted back into a percentage to derive the overall prediction in the bottom right.

The key element of this approach is that changes in the cohort's ability as quantified by the external indicator (here external assessment results) are accounted for in the prediction. In other words, if the cohort appears to be getting stronger over time (i.e. their external assessment results rise) then we would allow more high grades to be awarded to acknowledge this increase in cohort ability – rather than (arguably unfairly) limiting this year's more able cohort to only the amount of top grades achieved last year (Ofqual, 2017).

There are a few key considerations in this approach. Firstly, how many Xciles to use. This is largely governed by a) the available sample size and b) the granularity of the external indicator. If cohort sizes are small, then splitting the cohort into many Xciles is likely to introduce instability into the approach (though arguably, the entire prediction matrix approach is not optimal with small cohorts as the relationship with an external indicator is likely to be unstable regardless of the number of Xciles). If the indicator is a set of discrete grades, then it only makes sense to utilise as many Xciles as there are grades; if it is more fine-grained marks then using a higher number is feasible.

The second and more important consideration is what the external indicator is. As mentioned above, the many options here will be discussed at length in a later section, but the key point to be aware of is that this approach's potential is entirely contingent on how strong and consistent of a predictor the external indicator is of performance on the assessment at hand. If the predictive relationship is weak, then changes in the external indicator will not necessarily reflect changes in performance on the assessment, undermining its usage. If the relationship is not consistent and stable over time, then changes in the external indicator might be under- or over-compensated for in the method's assembly of predictions, introducing bias (which may occur if for instance, performance on the external assessment rises over time whilst performance on the assessment at hand does not). As such, when selecting an external indicator we must be confident that is it a strong and stable predictor of outcomes on the assessment at hand.

Though it muddies the clear definition of equating and prediction-based approaches in this paper, it is worth mentioning the work of Bramley and Vidal Rodeiro (2014). They established that technically speaking, the prediction matrix approach (with prior attainment as the external indicator) employed by the English awarding organisations for GCSE and A-level standards maintaining is functionally a frequency estimation equipercentile equating approach, using prior attainment as if it were an anchor assessment. The main difference is that the equating is only applied at specific grade boundaries rather than throughout the mark range. Though it has not come to pass, their work questions whether, given this similarity, it might be logical to switch to the equating approach to allow for smoothing techniques to be applied, the whole mark range to be equated and its assumptions to be thoroughly checked.

5.1.2.2. Logistic regression

In a logistic regression approach, the reference cohort is used to model the chance of achieving each grade on the assessment based on the external indicator. Technically this is termed a 'multinomial' logistic regression as our outcome variable (grade on the assessment) has more than two levels.

With this models created, we can then use the fitted regression line to predict how likely every candidate in the current cohort is to achieve each grade, based on their external indicator. For example a candidate who scores 20/50 on the external assessment might have a 1 per cent chance of achieving a grade 7, a 4 per cent chance of achieving a grade 6, and so on. Once this

has been done for all candidates, these probabilities of achieving each grade can be averaged across the entire current cohort to provide a prediction for the proportion of the cohort that we expect to achieve every grade, just as a prediction matrix approach results in.

An example of fitted logistic regression lines for certain key grades can be seen below. Note that this example incorporates some elements of an ANCOVA as outlined above; it also factors in school type as a background variable.

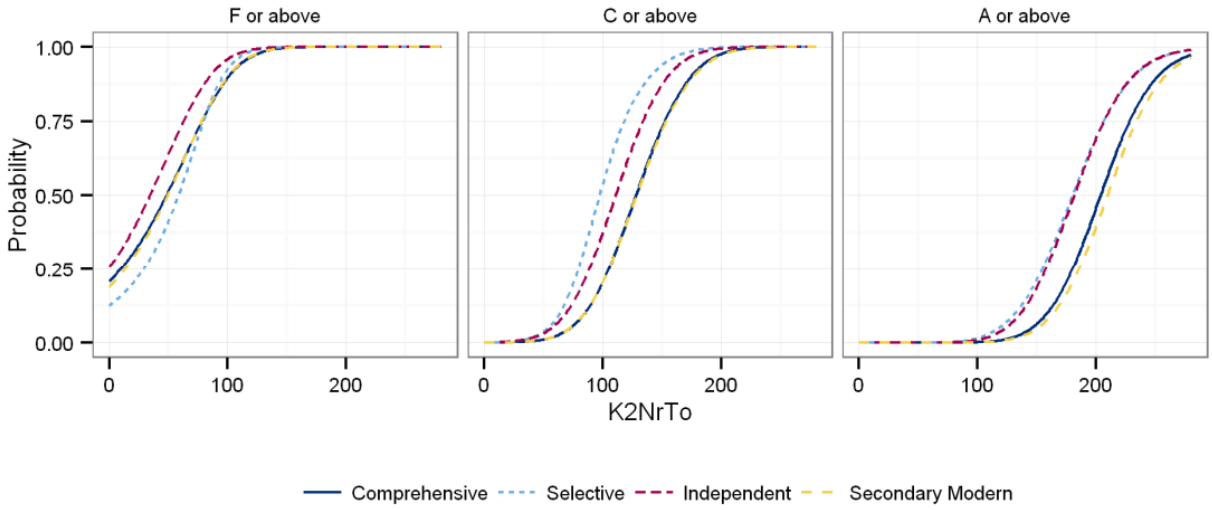

Figure 8: Example of a logistic regression model for predictions – from Benton and Sutch (2014; pg 79)

If we ignore the ANCOVA element and consider just a Comprehensive school (solid blue line) we can see that a candidate achieving 100 on the x axis (the predictor variable selected; here normalised KS2 marks) would have around an 90 per cent chance of achieving a grade F or above, around a 20 per cent chance of achieving a grade C or above, and around a 0 per cent chance of achieving an A or above. These probabilities can be integrated to derive the probabilistic prediction for candidates on 100 as outlined above: they have a ~10 per cent chance to achieve lower than an F, a 70 per cent chance of achieving between a D and an F, a 20 per cent chance of achieving a B or a C, and no chance of achieving an A or above.

Just as with prediction matrices, the same assumptions are made in logistic regression – the relationship between the external indicator and assessment needs to be strong and consistent in order for the approach to produce valid predictions, meaning the choice of external indicator is just as important. However, whilst the number of Xciles is not a decision needed, there are others required.

The main choice needed is the regression methodology utilised. The most straightforward is a linear regression which will work well in many situations – but some external indicators may not have a linear relationship with performance on the assessment at hand. Pre-live usage modelling would be necessary to establish whether a linear relationship does bear out, and in live settings the fit of the regression model (potentially compared to non-linear alternatives) should be checked to verify that the specific type of regression chosen remains appropriate.

Compared to prediction matrices, the main advantage of this approach is that it is less coarse than a prediction matrix (particularly where the number of Xciles selected is low), and therefore retains more fine-grained information from the external indicator than binning candidates into Xciles does. In Benton and Sutch (2014) a comparison of the two approaches in the UK general qualifications setting found that logistic regression resulted in more accurate predictions – but only marginally so.

Whilst their evaluation of it is somewhat conflated with use of a normalised external indicator to combat its inflation over time, the key benefits of a logistic regression approach were improvement in predicting the extent of differences between awarding organisations, and being more accurate for subjects with a large number of high ability candidates.

Whilst IB is a sole practitioner and does not have other awarding organisations to worry about, this might be applicable to (for instance) differences between countries which have different value added relationships between the external indicator and performance on the assessment at hand. This raises another potential benefit of logistic regression; it is relatively straightforward to add additional control variables into the regression to account for this type of differential relationship with the external indicator amongst subsets of the cohort.

The main downside of logistic regression is its complexity – not in terms of implementation; there are many robust software packages which can be used to implement the approach technically – but in terms of explaining it to stakeholders. Whilst it is a relatively straightforward and well known statistical technique, any regression method is somewhat of a black box in terms of how the predictive relationship between variables is established. There is a risk that an audience would need to have how regression as a concept explained in order to understand how a logistic regression functions, whereas prediction matrices are straightforward enough to be computed with pencil and paper and lend themselves to much easier exemplification. Explaining regressions to a lay person almost necessitates the use of figures and graphs to show the relationship between variables and the regression line.

5.2. External indicators of cohort differences

As noted above, the strengths and weaknesses of the adjust prior outcome approaches depend on which external indicator of cohort differences is used. Ultimately any external metric could be used for this purpose, but in this section we discuss a few of the most commonly utilised.

5.2.1. Prior attainment

The most commonly utilised external indicator of differences between a reference and the current cohort is a measure of both cohorts' prior attainment on another assessment. This is the chief method utilised in England's general qualifications; for GCSEs key stage (KS) 2 national curriculum tests (NCTS)[13] are used to derive a metric of prior attainment, and for A-levels GCSEs are used to derive a metric of prior attainment (Ofqual, 2017). In both cases, results across the suite of KS2 tests/GCSEs are averaged in an attempt to approximate a "general ability" measure for all candidates, and this average of performance in the suite of assessments is used as the prior attainment measure for standards maintenance purposes.

The derivation of a general ability type measure is key to the success of this approach in the UK context; any individual GCSE result would be a very poor predictor of outcomes on a given A-level[14], as there is a substantial amount of error inherent in any individual subject's grade (not to mention, the grades used here are fairly coarse measures of performance in the first place[15]). Similarly, another major factor in its success is the sheer volume of data available; nearly all

[13] Commonly, although falsely, referred to as 'SATs'.
[14] Whilst a given GCSE predicting an A-level in the same subject might be assumed to be quite strongly predictive, there are several issues with this. Firstly, some A-level subjects have no clear comparator at GCSE, or several, which renders gathering prior attainment information problematic. Secondly, some subjects are much smaller at GCSE than A-level (i.e. Psychology or Philosophy), which leads to only a fraction of the cohort having prior attainment data. So whilst using (for instance) English Language GCSE results to predict English Language A-level results might work well, this would fall down for many other lower entry subjects.
[15] Though it is possible to normalise KS2

students in England will have KS2/GCSE results to inform the method[16]. In combination, these two factors combine to render prior attainment a powerful approach (for instance it has been found to exceed the predictive power of common centres approaches; Taylor, 2014).

However, the primary issue with prior attainment is that it is time-distal from the assessment at hand, and there is no guarantee that the candidates have all progressed equally since the time the prior assessment was sat. For instance, if there were changes to the teaching of students after their prior assessment that have increased the 'value added' to their learning and resulted in them performing better on the assessment at hand, then this would be completely missed by a prior attainment methodology, which can lead to grad inflation/deflation (Bramley, 2013). In other words, the core assumption is that any shift in prior attainment will be perfectly reflected in achievement in the assessment at hand.

A related issue faced by the English setting in using this approach is that there are different awarding organisations (AOs) each of which deliver the same qualification and thus need to generate separate predictions. Benton and Sutch (2014) note that this results in an issue when the centres which select to use each AO have different value added between the prior assessment and the current ones – leading to some AOs' predictions being over-generous and others' under-generous. IB, despite having different countries and centre types which are likely to experience similar differential value added, is insulated against this issue because it is the sole provider of its programmes. As long as the value added relationship across the whole cohort remains consistent year-to-year, a sole provider can be confident in prior attainment as external indicator.

5.2.1.1. Reference test

One of the pitfalls of England's prior attainment approach to standards maintenance is that it fails to account for improvements in performance in particular subjects over time (Bramley, 2013). Because whether outcomes are allowed to improve is pinned to an approximation of general ability, if there has been progress in a specific subject but not others, this will be 'glossed over' by the approach. The solution to this was to implement a reference test – get a small but representative sample of students from across the country to sit a "GCSE-like" assessment at a similar time to their GCSEs that is graded in a similar manner (Ofqual, 2019). Crucially, the reference test remains almost entirely the same over time – the same questions are posed year-on-year.

If there had been specific absolute improvements in performance, this use of the same paper year-on-year readily identifies this. This information is then fed into the prediction based on prior attainment (as outlined above) – but with an element of judgement involved. For instance, if the reference test indicated that 3 per cent more candidates were at the A* level, then this would not automatically mean that the prediction for A* was increased by 3 per cent – an increase of only 1 or 2 per cent might be applied.

It's also possible to use a reference test in isolation as a source of prior attainment[17] information which can then be fed into a prediction matrix or logistic regression model. However, doing this loses many of the advantages of the cohort-level prior attainment information outlined above – chiefly that of sheer volume of data. Whilst a reference test's cohort should be selected to be representative, there is still a greater risk of cohort effects than with a near-system wide source of

[16] Eason (2006) suggests that the approach works with cohorts of over 100 candidates, but Ofqual (2015)'s information on tolerances applied to the approach in GCSEs shows that it is most stable with 3,000 or more candidates.
[17] Though arguably, whilst a reference test must be sat prior to the assessment at hand, it could be deemed 'concurrent attainment' as it is much more time-proximal to the assessment at hand than prior attainment usually is.

prior information. However, in situations where such a wide-reaching source of prior attainment is not available, this ceases to be a concern.

Whilst reference tests seem like a valuable tool to mitigate a key weakness of using prior attainment that is time-distal to the current assessment, there are some substantial obstacles. Firstly, practically operating a reference test is extremely challenging – it will be time consuming to develop, administer, mark and analyse, especially given the stringent security requirements which need to be in place to prevent it becoming exposed over time. There is also an issue of school recruitment – given the need for the sample of students sitting it to be representative, unless participation is compulsory it's likely that some schools will be unwilling to take part and will cause representativeness to be compromised.

Secondly and perhaps more problematically, a reference test is only really useful relative to the standard of its own subject – a mathematics reference test can tell us whether there have been gains in the prior assessment to current assessment value added for mathematics, but not for music or history. A reference test in each subject would be required in order to adjust that subject in this manner, which rapidly becomes completely infeasible.

Arguably the logical conclusion to the question of when reference tests are most valuable is "when we have a stable cohort but expect a shift in absolute performance to manifest due to extraneous factors like a new method of teaching being rolled out". For example, if a new computer science curriculum is deployed, then a reference test in this subject might be very valuable in order to detect and factor in absolute improvements in performance filtering through with consecutive cohorts. However, because a reference test needs to be in place before improvements begin, and the lead time on developing and implementing them is substantial, the practicality of such an approach is questionable at best. In England, reference tests are only used in the two most high-stakes and largest entry GCSEs; English Language and Mathematics (Ofqual, 2019).

Another locale which makes substantial use of reference tests is Hong Kong (Burdett et al, 2013), but again only referencing in the "core subjects" is implemented, again speaking to the practicalities of reference testing an entire suite.

5.2.2. Concurrent attainment

An alternative to utilising a measure of candidates' prior attainment to inform predicted outcomes, is to instead utilise a measure of their concurrent performance on the suite of assessments currently being sat. That is, in GCSEs, instead of the average KS2 score being used to predict GCSE outcomes, the average GCSE points score could be instead (the IB analogues being MYP and DP respectively). The problem with this approach is fairly clear – it is cyclical, because until we have awarded all subjects, we will not know any given candidate's mean grade.

This limits the usefulness of concurrent attainment in a live awarding setting, with it typically being used to validate grade boundary setting and awarding after the fact (England carries out a cross-AO screening exercise using mean GCSE grade immediately prior to releasing results to candidates; Taylor, 2014).

That said, it is worth noting that research in the English context has shown that mean GCSE grade is a better predictor of outcomes on a given subject than average KS2 score, or any of the various common centre approaches (Eason, 2012; Taylor, 2014). This is unsurprising; the time-proximal nature of concurrent attainment means that concerns about differential value-added since a prior time point are nullified, and unlike common centre-type approaches, it can utilise the entire cohort's data. If an approach could bypass the practical issues of utilising concurrent attainment, it is likely it would be a strong and robust external indicator of performance to input into prediction-based methods.

5.2.2.1. Instant summary of achievement without grades (ISAWG)

The key problem with typical concurrent attainment approaches outlined above is actually two sub-issues, the first being that as more and more subjects are awarded, any given candidate's mean grade will change, and the second being that grades are not known until after award meetings and grade boundary setting – causing a significant delay in any one subject's data contributing to the mean grade metric.

The former problem is fairly insurmountable, though it is possible to address it to some degree by shifting the award of the largest subjects to earlier in the session. The latter however can be addressed by moving away from using grades as a measure of concurrent attainment; this is the aim of the 'Instant summary of achievement without grades' or ISAWG method developed by Benton (2017). If we are not contingent on knowing candidates' grades to input a subject's data into a concurrent attainment approach, then we can utilise much more data at any given point in an awarding session.

The below figure demonstrates how the additional data available to ISAWG means that a fairly stable estimate is available a relatively short way through an awarding session; each pane represents a point further through the awarding session and how closely the ISAWG measures generated at that point correlate to the measures which are derived at the close of the session.

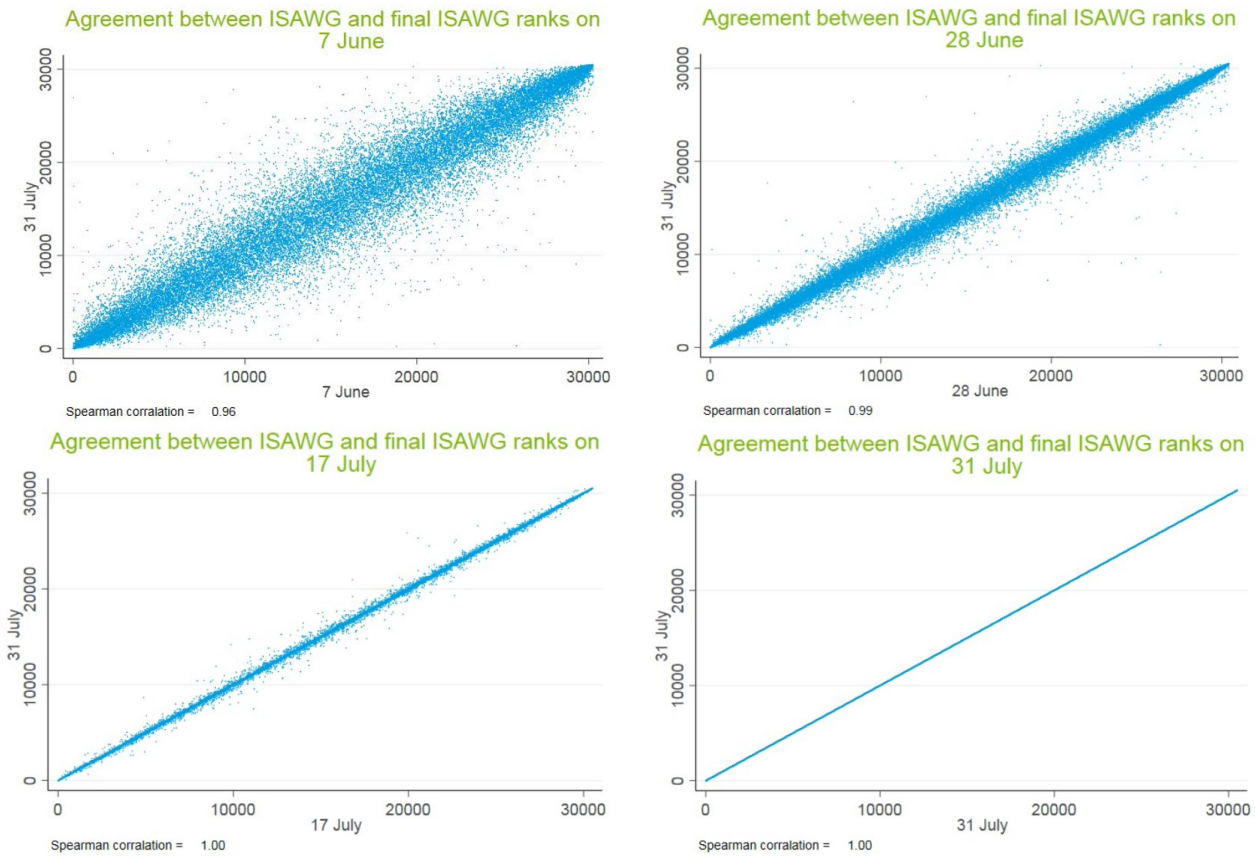

Figure 9: ISAWG convergence over a WJEC awarding series – from Johns & Evans, 2019, slide 6)

Benton summarises the ISAWG approach as follows:

1. Within [the current and reference years] create a single measure of ability for all candidates that is on the same scale regardless of which assessments any given pupil has taken.
2. Having created this measure, by looking at achievement across all assessments together, we can find sufficient centres with stable entries over time to [put it on a common scale] between years.
3. Now that we have a calibrated measure of achievement that is comparable over time, we can use it to inform the positioning of grade boundaries in all qualifications.

More technically, the approach uses standardised scores[18] on each assessment (notably, not subject as a whole, but rather each individual component within each subject) to derive an ISAWG score for each candidate in each session (i.e. the reference and current session). This is accomplished using an iterative procedure called alternating regression to "home in" on the best estimate of ISAWG score for each candidate, until error is reduced to an acceptable level (or the maximum number of iterations is reached). The full details are given in Benton (2017), but in essence, the approach is effectively a form of Principal Components Analysis; reducing many estimates of candidates' ability to a single numeric metric (or component).

Once this has been done, a number of equating methods outlined above can be used to calibrate the reference year and current year's ISAWG scores onto the same scale. Benton (2017) utilised weighted unsmoothed equipercentile and chained equating, finding that chained equating resulted in a much lower equating error – but that this might be attributable to the simulated data used for the research. They conclude that both frequency and chained approaches should be tested when using an ISAWG approach.

After the calibration of each year's ISAWG has been completed, the metric is ready to be used as an external indicator of performance in any prediction-based approach; much like with prior attainment, it can be used as an input to prediction matrix or logistic regression methods to establish a prediction for the proportion of candidates that should achieve each grade.

A key factor to raise about why ISAWG 'works' is that it relies on overlap in entry between subjects within the suite being used to derive the ISAWG metric. If, for example, for A-levels in the UK, candidates sitting a particular group of subjects (for example French, Spanish and German) only sit those subjects and not any others, then they would be isolated from the rest of the suite and it would be impossible to work out a measure of language subject candidates' 'general ability' that is comparable to such a measure for the rest of the subjects in the suite. However, this is not the case in reality (though some subjects are always "less connected" than others) – particularly in the IB's case[19] since the 'subject area' approach means that candidates are guaranteed to have sat a wide range of subjects, and that there are strong links for all candidates through high entry subjects like Mathematics and so forth. Because the whole suite is linked by candidates sitting many different combinations of multiple subjects, a general ability metric can be derived across the suite.

The slight issue this introduces is that the ISAWG measure itself will be more influenced by some subjects than others. If more candidates do certain subjects which in turn act as key links in the network of entries, then those subjects might have considerable influence on the overall ISAWG score (Johns & Evans, 2019). This is a similar reason to why in assessment statistics we remove marks on the item at hand from the total mark used in discrimination indices; items with a very high

[18] Z-scores with a mean of zero and an SD of one.
[19] Or indeed many other national assessment systems worldwide; A-levels are if anything slightly outlying in narrowing the scope of study to so few subjects.

total number of marks available otherwise have a substantial impact on candidates' overall rank order in a paper.

This issue can be visualised by creating connectivity plots between subjects, such as the example below. Each circle represents a subject, with its size the number of entrants. Lines represent common entrants between subjects.

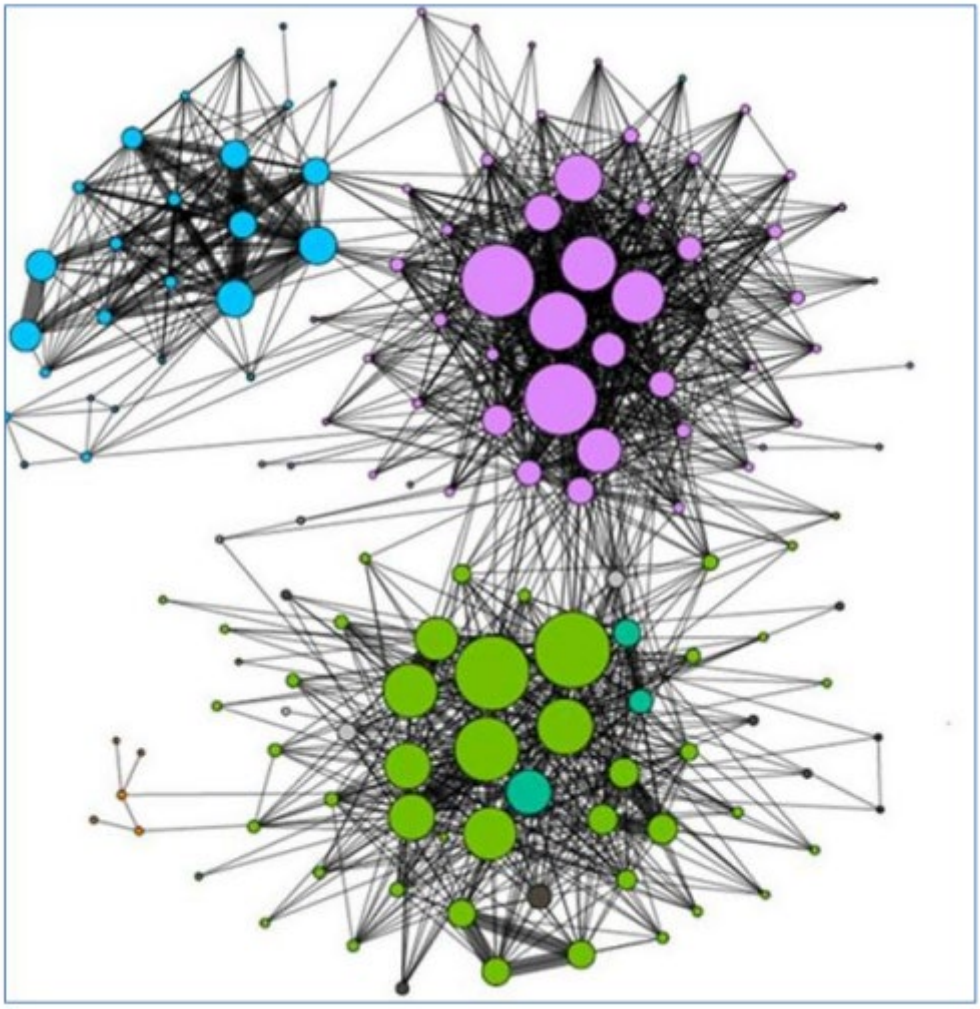

Figure 10: Example of subject connectivity for Cambridge Assessment qualifications – from Benton (2017; pg 5)

In this graph, there are a number of large highly connected subjects with many connections – these will be influential on any ISAWG measure derived. The small extraneous subjects are much less connected and will have minimal impact on an ISAWG measure. Notably, the suite of assessments in this example has three somewhat discrete sets of qualifications (in each colour) and so there might be an argument for deriving separate ISAWGs for each.

Similarly, if there are more assessments for a given subject (or subject area) than for others, these will have more impact on the ISAWG scale derived. Think about a suite with 10 mathematics exams and 3 reading ones; the mathematics exams will contribute significantly more to the ISAWG than the handful of reading ones, meaning that the resulting ISAWG score ceases to reflect general ability and more just 'mathematics ability, with a side of reading'. Essentially the point here is that the ISAWG is the sum of its parts, and whilst how it is computed is somewhat of a 'black box' which obfuscates how it is arrived at, it will always be the case that some subjects or subject areas will carry greater weight than others.

There are a few key benefits that make ISAWG unique amongst the possible approaches to maintaining standards statistically. Firstly, it is arguably the "best" concurrent attainment approach insofar as it allows for more data to be included in derivation of the general ability measure than one relying on grades – and as discussed above, concurrent attainment has been proven to be a better predictor of attainment than prior attainment. So, ISAWG means using a better predictor of attainment than other options – likely due to the time-proximal nature of concurrent attainment.

Secondly, ISAWG is unique amongst all the approaches discussed in this paper in that it is the only approach that can be reliably used for completely new specifications. Because most approaches rely on maintaining the prior outcome/equating to another assessment form *in the same subject*, if a subject is new then there is nothing to maintain.[20] However, ISAWG instead generates an expected outcome based on performance *across the whole suite*, so can be used even in a subject's first session.

Thirdly, ISAWG can be used quite reliably for very small subjects' awards, again because it draws on data from across the entire suite. Whilst small subject predictions can never fully avoid issues of unreliable predictions, it is considerably better to use (for instance) a 100 strong cohort's data in six other subjects than to use last year's 80 strong cohort's data in just this one small subject to generate a statistical prediction. Indeed, this was a key reason for Johns and Evans' (2019) investigation of the approach as an alternative to common centres.

However, whilst these benefits are real and very substantial, there are also real drawbacks to this approach. As outlined above, one relates to the ISAWG's nature as dependent on "what goes in", and inevitably more influenced by some subjects than others – which is exacerbated by its being somewhat of a 'black box', as it is not possible to easily tell what subjects have greatest influence without further analyses. That said, there is a possible way to mitigate this – by carrying out ISAWG on a subset of data. For example, a 'Science ISAWG' and 'Arts ISAWG' could be independently computed and used for the relevant subjects' predictions, allowing more control over which subjects influence which predictions and potentially increasing their validity (though analysis would be needed to verify this). The downside is that this reduces the amount of data included in each ISAWG, which might undermine the robustness of the approach in a similar way to subsetting common centres to just stable common centres. In summary, much further investigation would be needed to establish whether such a subsetting approach could help mitigate this issue – and even then, the ISAWG statistic itself remains non-transparent and inevitably will be more impacted by some subjects than others; with subsetting we could just control which subjects those are to some degree.

Any concurrent attainment approach is also, practically speaking, non-deterministic, in that because more data is fed into the ISAWG as the session progresses, a different prediction for a component would emerge at the end vs the start of the session (Johns and Evans, 2019). At its extreme, the earlier subjects may not be awardable using ISAWG because no concurrent achievement data is yet available. This can however be mitigated by ensuring that the largest and/or most influential subjects in the measure are awarded early on in the session, thereby stabilising the measure as quickly as possible. Nonetheless, any use of ISAWG must be content with this feature and plan the session around it.

Finally, an additional major factor in ISAWG's development is the complexity surrounding the approach, and how lengthy this makes modelling of its use. ISAWG can be used with prediction matrices or logistic regression, to inform the whole cohort or a subset's prediction (common

[20] Notably, this represents a deviation from the standards maintenance scenario this paper focuses on – but new subjects are a scenario of interest, so a brief discussion of standard setting is useful in this context.

centres), can be carried out on subsets of subjects, needs checking in order to establish which subjects are most influential, ISAWG could be used for just some subjects' awards and not others, and the equating forming step two above can be one of the various score equating ones explored previously in this paper. In short, there are innumerable possible variants to an ISAWG approach which would need to be investigated to determine the optimal one – as the length of this section attests to. Whilst, once an approach is agreed, it is not too computationally intensive (though it still needs specialist statistical software to carry out), actually settling on the precise approach would require substantial investigation up front.

6. Combining multiple sources

The above two sections have dealt with the two main schools of statistical standard setting methods; score equating techniques and prediction-based ones. However, if anything has been made apparent in this paper, it is hopefully that the myriad of different approaches have different benefits, detriments, and applications – and give different results. Carrying out several methods, as much of the literature does for purposes of comparison, will give several (often very) different answers as to what marks on one assessment form are equivalent to those on another.

This is, fairly intuitively, generally considered a problem – how can an analyst tell which approach is the most valid for their particular situation? This section covers a handful of pieces of work which have investigated ways in which the myriad of different approaches can be part of a solution instead.

Von Davier (2011) discusses the averaging of equating functions to achieve a compromise between them. Weighting the two (or more) functions averaged is possible in order to favour one or several. A number of different ways to achieve this aim are discussed, including the angle bisector method and the 'swave' approach. If using equating techniques and uncertain of which is 'best', for instance when all possible approaches have some of their assumptions violated, then this is a viable approach – but it naturally takes additional time and adds complexity. It might also be difficult to explain why the selected average is more valid than any one equating function.

Bimpeh (2018) used a Bayesian approach to integrate examiner judgement and statistical predictions about where grade boundaries should fall in UK general qualifications. The main challenge here was selecting weights to apply to each source of information – the number of candidates the predictions were based off was used for the statistical information, and the number of candidates on marks scrutinised by examiners was utilised for the judgemental. Ultimately, the result is not too dissimilar to a weighted average; the main benefit of the Bayesian approach being the easy derivation of confidence intervals around said average.

Whilst a neat approach to integrating generally discrete information together in a manner that provides valuable certainty metrics, ultimately the choice of what values to use as weights determines which is given more credence in the results – arguably it could be reasonable to set rules on what weights to use governed by the analysts' perception of the reliability of various evidence sources. For instance, the number of matched candidates could always be the weight applied to the statistical data, and the number of expert judges could be multipled by a flat value (say 1,000) to ensure that it was given more weight in cases where there was little statistical information. The other main factor to consider with such an approach is its complexity, both in terms of explaining it to laypersons, and implementation.

Von Onna, Jongkamp and Lamprianou (2021) utilised a combination of ANCOVA, pseudo-anchor and a judgemental standard setting approach (termed 3DC), together with a resit analysis to provide a lower bound for the exercise. This was combined in a complex manner governed by a 'flowchart', with rules dictating the course to take if there was a discrepancy between different sources of information (i.e. by eliminating one approach). One method was selected as the baseline, and then others were used to validate it. If passing through the specified rules (i.e. not completely discrepant from other approaches), they would be added into a weighted average of the pass marks suggested by each approach, using either the number of candidates or the standard error as the weight as appropriate.

This provides a template for a potential means to integrate multiple standard setting approaches via a combination of a rules-based and statistical system, rather than a purely statistical one. Different contexts would of course need to develop their own set of workable approaches based on the data available and assessment form design, but in principle in any situation several equating

approaches could be performed and an integrative exercise conducted. The main downside is the significant additional complexity and effort involved – and arguably, introducing many equating methods might compromise perception of the equated marks as valid if the nature of statistical error is not explained carefully. Nonetheless, especially in high-stakes situations it is hard to argue that combining several sources of information would not increase the validity of an equating exercise – if all such approaches had been carefully considered and modelled beforehand.

7. Initial review of suitability for the IB

Whilst the approaches taken forward for modelling will be agreed with IB personnel, in this section we present our initial reflections on the approaches which may or may not be suitable for the IB. It is important to outline the contexts IB's assessments encompass to enable this:

1. Large stable subjects
2. Small subjects
3. Growing subjects
4. Changing curriculum or assessment models
5. New subjects
6. Cohorts with varied strengths (due to inconsistent syllabus coverage in different regions)
7. (Usually small) subjects with no cohort overlap from year to year
8. Verification model (coursework where fixed boundaries are reviewed each session)

It is also worth noting some important factors about the data available to the IB and their assessments in general:

- The IB MYP provides some prior attainment information for DP subjects, but a large proportion of the DP candidature does not take the MYP.
- The IB does collect some demographic information about its candidates beyond (for instance) gender and age.
- The IB's assessments are generally high-stakes and as such security is a key concern, meaning re-use of items is not conducted from session to session. However, there are some limited commonalities between sessions:
    - Coursework is common between sessions, and currently retains the same boundaries each sitting.
    - In some cases there are commonalities between papers for different timezones.

It is perhaps easiest to begin by ruling out some approaches. Despite its heavy use in other systems, prior attainment is unlikely to prove a fruitful approach for the IB, as it is reliant on having substantial volumes of prior attainment data in order to be such a powerful way to generate predicted outcomes. Similarly, implementing reference test(s) is unlikely to prove practical.

Due to the lack of common items between sessions, we can also broadly rule out nonequivalent groups designs as a whole, along with IRT-based approaches[21]. There is one notable caveat, however; in subjects where coursework features, it might be possible to use this as an anchor assessment – though of course this assumes that changes in performance on the coursework would be reflected by those on the other assessments. This is not necessarily the case; often with static coursework tasks outcomes rise over time due to increasing teacher familiarity, which if coursework is used as an anchor assessment would lead to equating models assuming a constant rise in the ability of the cohort, and commensurate overall creep in outcomes over time. As such a review of whether this is the case for IB coursework generally would be needed to assess whether any nonequivalent groups design is likely to be advisable for subjects with coursework.

This leaves, to generalise, three broad approaches:

d. Basic equating techniques

[21] Common person equating is unlikely to prove fruitful in maintaining standards from one session to the next given the time gap between them – resit analysis would be more appropriate. Further, context 6, cohorts with varied strengths, clearly violates the unidimensionality assumption key to IRT.

e. Concurrent attainment approaches
f. Approaches seeking to maintain the prior outcome (i.e. via common centres)

Basic equating techniques, as a whole, are suitable in situations where the two cohorts are comparable in ability. However, considering the IB's list of contexts above, only around half of the scenarios meet this assumption (stable subjects in particular, but the approach is also likely to work when curriculum or assessment models change or in new subjects, if a suitable comparator assessment can be found and the cohort is stable). Ultimately these cases where basic techniques can work are those where the cohort is stable and consistent over time (it being reasonably large is important too, as this impacts how likely the cohort is to be stable). In these scenarios basic techniques are likely to be viable due to their relative simplicity – equipercentile and circle-arc equating would seem, based on the literature, to be the preferred choices. Whilst ANCOVA approaches are useable given the IB's wealth of demographic data, they are fraught with potential ethical debates so should be weighed up extremely carefully before adoption.

It is worth noting that basic equating approaches *can* be applied in almost any circumstance (they need only a small sample size), which might mean that in some cases they are the only viable option. The question is whether it is advisable to do so (i.e. if cohorts are likely to be dissimilar), or whether it would be preferable to rely on judgemental approaches alone. Another key issue with adopting basic approaches is that they do not flag when the cohort begins to destabilise – this would need to be screened for to avoid a situation where basic approaches are inappropriately persisted with even as a subject's cohort begins to change.

Concurrent equating approaches largely boils down to ISAWG approaches; as outlined in our discussion of that approach, it moving away from grades is a major practical benefit which means more data is available to inform the ISAWG score at every point in an awarding session. ISAWG is undoubtedly an extremely powerful approach, and one that appears suitable for IB's programmes due to their featuring a broad range of assessments with some (i.e. Mathematics) that form strong links throughout the dataset. Further, it offers (by some margin) the most convincing equating approach for some of the most awkward contexts, including very small subjects, those with complete cohort change, and completely new subjects. Completely new subjects in particular are completely unmanageable by any other approach.

However, ISAWG approaches are extremely complex, with a huge wealth of available options and modifications (even when compared to the other approaches in this paper). It is worthy of a full literature review and investigation alone, such is its complexity. It seems likely that ISAWG would be a method that *can* offer solutions for IB's most challenging contexts, but would require a substantial amount of effort to adequately trial and implement it – effort which might be disproportionate to the benefits it offers. The approach also has other drawbacks, being tricky to implement and a black box in terms of ease of explanation to laypersons.

Approaches seeking to maintain the prior outcome can likewise be characterised as either the current IB SRB approach, or the common centres group of approaches. The current approach shares the flaws of the basic equating approaches (being strictly norm-referenced it is similar to an equipercentile approach), so refer to the above discussion for its strengths and weaknesses. The common centres approach however is a well-established means of attempting to account for cohort changes that is viable as long as there is a large enough cohort, and sufficient centres taking the subject from one year to the next. Whilst found to not be as strong of a method as prior attainment for maintaining outcomes, it is still superior to many other approaches as it aims to account for any change in cohort ability over time. It is also appropriate in just about all of IB's contexts, with the exception of very small cohorts and completely new subjects (though there is the possibility of using common centres to link to a similar existing subject, dubious as this may be).

As such, common centres seems worthy of consideration as somewhat of a ‘default’ option for SRB-setting, given its fairly broad useability in the IB’s contexts. However it will be important to investigate how many subjects have sufficiently large entries to use it, as many of IB’s subjects have fairly small entry sizes which would render the data attrition common centres leads to highly problematic. It may also emerge that common centres prove unstable and therefore the approach does not work as well as in a single-country system.

8. References

Albano, A. D. (2016). equate: An R package for observed-score linking and equating. *Journal of Statistical Software, 74*(8), 1–36.

Alberts, R. V. J. (2001). Equating Exams as a Prerequisite for Maintaining Standards: Experience with Dutch centralised secondary examinations. *Assessment in Education: Principles, Policy & Practice*, 8(3), 353–367, DOI: 10.1080/09695940120089143

Babcock, B., Albano, A., & Raymond, M. (2012). Nominal Weights Mean Equating: A Method for Very Small Samples. *Educational and Psychological Measurement*, 72(4), 608-628. DOI: 10.1177/0013164411428609

Babcock, B., & Hodge, K. J. (2020). Rasch Versus Classical Equating in the Context of Small Sample Sizes. *Educational and Psychological Measurement*, 80(3), 499-521. DOI: 10.1177/0013164419878483

Baird, J. A., & Scharaschkin, A. (2002). Is the Whole Worth More than the Sum of the Parts? Studies of Examiners' Grading of Individual Papers and Candidates' Whole A-level Examination Performances. *Educational Studies*, 28(2), 143–162. DOI: 10.1080/03055690220124588

Baser, O. (2006). Too much ado about propensity score models? Comparing methods of propensity score matching. *Value in Health, 9*(6), 377-385.

Benton, T. (2017, November). Pooling the totality of our data resources to maintain standards in the face of changing cohorts. Paper presented at the 18th annual AEA-Europe conference, Prague, Czech Republic.

Benton, T., & Elliott, G. (2016). The reliability of setting grade boundaries using comparative judgement. *Research Papers in Education*, 31(3), 352–376. DOI: 10.1080/02671522.2015.1027723

Benton, T., & Sutch, T. (2014). *Analysis of use of Key Stage 2 data in GCSE predictions.* Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.

Bimpeh, Y. (2018). Intelligent integration. *Inside Assessment, 1,* 23-26.

Birnbaum, A. (1969). Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology, 6*(2), 258-276.

Black, B., & Bramley, T. (2008). Investigating a judgemental rank-ordering method for maintaining standards in UK examinations. *Research Papers in Education*, 23(3), 357–373. DOI: 10.1080/02671520701755440

Bond, L. A. (1996). Norm-and criterion-referenced testing. *Practical Assessment, Research, and Evaluation, 5*(1), 2.

Boone, W. J., & Staver, J. R. (2020). *Advances in Rasch Analyses in the Human Sciences.* Springer. https://doi.org/10.1007/978-3-030-43420-5

Bramley, T. (2013). *Prediction matrices, choice and grade inflation*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.

Bramley, T., & Gill, T. (2010). Evaluating the rank-ordering method for standard maintaining. *Research Papers in Education*, 25(3), 293–317. DOI: 10.1080/02671522.2010.498147

Bramley, T., & Vidal Rodeiro, C.L. (2014). *Using statistical equating for standard maintaining in GCSEs and A levels.* Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.

Braun, H. I. (1982). Observed-score test equating: A mathematical anaysis of some ETS equating procedures. *Test equating*.

Burdett, N., Houghton, E., Sargent, C., & Tisi, J. (2013). Maintaining Qualification and Assessment Standards: Summary of International Practice. Slough: NFER.

Chen, H., & Holland, P. (2010). NEW EQUATING METHODS AND THEIR RELATIONSHIPS WITH LEVINE OBSERVED SCORE LINEAR EQUATING UNDER THE KERNEL EQUATING FRAMEWORK. *PSYCHOMETRIKA*, 75(3), 542–557. DOI: 10.1007/S11336-010-9171-7

Cook, L. L., & Paterson, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11*(3), 225-244.

Cuff, B. M. P., Meadows, M., & Black, B. (2019). An investigation into the Sawtooth Effect in secondary school assessments in England. *Assessment in Education: Principles, Policy & Practice*, 26(3), 321-339, DOI: 10.1080/0969594X.2018.1513907

Curcin, M., Howard, E., Sully, K., & Black, B. (2019). *Improving awarding: 2018/2019 pilots.* Coventry: Ofqual.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York: Springer.

von Davier, A. A. (Ed.). (2011). *Statistical Models for Test Equating, Scaling, and Linking*. Springer. DOI 10.1007/978-0-387-98138-3

Eason, S. (2006). *Predicted Outcomes: Some Issues Related to Small Numbers of Candidates.* Guildford: AQA.

Eason, S. (2012) *Alternative key stage 2 models for predicting GCSE outcomes*. Research report CERP_TR_SE_13092012. Guildford: AQA.

Furter, R. T., & Dwyer, A. C. (2020). Investigating the Classification Accuracy of Rasch and Nominal Weights Mean Equating with Very Small Samples. *Applied Measurement in Education*, 33(1), 44-53. DOI: 10.1080/08957347.2019.1674307

Gulliksen, H. (2013). *Theory of mental tests*. Routledge.

Holland, P. W., & Thayer, D. T. (1989). The kernel method of equating score distributions. *ETS Research Report Series, 1989*(1), i-45.

IBO (2018). *Assessment principles and practices—Quality assessments in a digital age.* Cardiff, U.K.: International Baccalaureate Organisation

Johns, D., & Evans, A. (2019). *IMPLEMENTING ISAWG IN AWARDING.* [Conference presentation]. Ofqual Educational Assessment Seminar.

Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling, and Linking: Methods and Practices*. (3rd ed.). Springer. DOI 10.1007/978-1-4939-0317-7

LaFlair, G. T, Isbell, D, Nicholas May, L. D., Gutierrez Arvizu, M. N., & Jamieson, J. (2017). Equating in small-scale language testing programs. *Language Testing*, 34(1), 127-124. DOI: 10.1177/0265532215620825

Liang, T., & von Davier, A. A. (2014). Cross-Validation: An Alternative Bandwidth-Selection Method in Kernel Equating. *Applied Psychological Measurement*, 38(4), 281-295. DOI: 10.1177/0146621613518094

Livingston, S. A., & Kim, S. (2010). Random-Groups Equating with Samples of 50 to 400 Test Takers. *Journal of Educational Measurement*, 47(2), 175-185.F

Livingston, S. A., & Kim, S. (2009). The Circle-Arc Method for Equating in Small Samples. *Journal of Educational Measurement*, 46(3), 330-343.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Welsley Publishing Company

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement, 8*(4), 453-461.

Masters, G. N. (1985). Common-person equating with the Rasch model. *Applied Psychological Measurement, 9*(1), 73-82.

Newton, P. (2020). *What is the Sawtooth Effect? The nature and management of impacts from syllabus, assessment and curriculum transitions in England.* Coventry: Ofqual

Ofqual (2015). Inter-board comparability of grade standards in GCSEs, AS and A levels: Summer 2015. Retrieved 13/01/2022 from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/488432/Inter-board_comparability_summer_report_2015_23_Dec.pdf

Ofqual. (2017). Awarding and comparable outcomes. The Ofqual blog. Retrieved 13/01/2022 from: https://ofqual.blog.gov.uk/wp-content/uploads/sites/137/2017/03/Awarding-and-Comparable-Outcomes-maths-meeting-2017-03-07.pdf

Ofqual. (2019). NRT Annual Statement 2019. Retrieved 13/01/2022 from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/826570/NRT_annual_statement_2019__-_FINAL196527.pdf

Oh, H., & Moses, T. (2012). Comparison of the One- and Bi-Direction Chained Equipercentile Equating. *Journal of Educational Measurement*, 49(4),399-418.

van Onna, M., Jongkamp, C., & Lamprianou, I. (2021, November 2-November 5). *Equating Cyprus teacher admission exams with multiple methods.* [Conference presentation]. 22nd Annual Meeting of the Association for Educational Assessment – Europe.

Peabody, M. R. (2020). Some methods and evaluation for linking and equating with small samples. *Applied Measurement in Education, 33*(1), 3-9.

Pinot de Moira, A. (2019). *Common Centres: In the context of maintenance of standards for the GCSE.* Unpublished report for the WJEC and CCEA.

Priestley, M., Shapira, M., Priestley, A., Ritchie, M., & Barnett, C. (2020). Rapid review of National Qualifications experience 2020: final report, September 2020.

Puhan, G. (2011). Futility of Log-Linear Smoothing when Equating with Unrepresentative Small Samples. *Journal of Educational Measurement*, 48(3), 274-292.

Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests*. MESA Press, 5835 S. Kimbark Ave., Chicago, IL.

Taylor, R. (2014). *Comparing statistical approaches for maintaining standards.* Centre for Education Research and Policy Research Report. Manchester, UK: AQA Education.

Wheadon, C., & Evangelidou, L. (2008). *IS IRT TEST-EQUATING BETWEEN TIERS ROBUST? A study of the context effects on the common items between tiers.*

Unit 109 Albert Mill
10 Hulme Hall Road
Castlefield
Manchester
M15 4LY

www.alphaplus.co.uk
john.winkley@alphaplus.co.uk