Fully Processed candidates: an analytical solution to the Post-Awarding Drift problem

Robert Hales

Abstract

The number of candidates for whom mark data is available at the time of an award is less than the total entry in the vast majority of cases — the relative difference between these is referred to as the Percentage Fully Processed (PFP). At some point between award and publication of results, the mark data for the additional candidates will become available and will have some effect on the percentage outcomes at each grade — the outcomes will drift, known as Post-Awarding Drift (PAD). This effect has been the focus of previous studies, where empirical and simulation analyses were undertaken. These studies found some relationship between the PFP and PAD, which further resulted in a recommendation of lowering the minimally required PFP (MRPFP) at award from 85% to 70% across all specifications in order to avoid PAD (although 75% was ultimately chosen). The aim of this study is to investigate PAD from an analytical perspective in order that this MRPFP might be chosen in a more substantiated way. The results show that the PAD as observed empirically or by simulation can be analytically explained, at the expense of making some practical assumptions, and that the MRPFP will vary from unit to unit. This is mainly driven by the size of the total entry so that in the extreme cases of very large-entry units (e.g. total entry of 400,000) the MRPFP is less than 1%. Additionally, the methodology was explored during the summer 2012 awarding series with the primary aim of facilitating decisions being made about units with low PFPs. Discussion is also provided around the use of additional information that is available during award preparation that can better inform the decisions made around the MRPFP.

Introduction

At the time of an award, not all candidates' marks will be available. Some marks are late as the scripts are still awaiting marking whilst other marks may be unavailable due to the candidate having been absent. The percentage of marks that is available at award is known as the 'Percentage Fully Processed' (PFP). The distributions provided at an award are therefore an incomplete representation of the full cohort's performance. The aim of an award meeting is to set grade boundaries for each judgemental grade[1] in light of the judgemental and statistical evidence such that standards remain comparable across series and awarding bodies. Before an award, unit-level 'Statistically Recommended Boundaries' (SRBs) are derived from a statistical prediction. At the award meeting, recommended boundaries are chosen with an appreciation of how closely the outcome adheres to this prediction. Problems can arise when the distributions from which the boundaries were chosen are significantly different to those where all marks have been accounted for — the Fully Processed (FP) distribution — as this may result in a change in outcomes that may have affected the choice of grade boundaries. At the time of publication of results, there may therefore be a drift in the outcome at a judgemental grade boundary compared to at award. This is termed 'Post-Awarding Drift' (PAD)

[1] A judgemental grade boundary is one for which examiners scrutinise candidates' work at award meetings, for example A, C and F at GCSE. Other boundaries are calculated once these judgemental ones have been determined.

— the difference between outcome at award and outcome at provisional results publication[2]. Due to the importance of the statistical evidence in supporting examiners' judgements, the guidance it provides should be as reliable as possible and hence it is desirable to have an appreciation of the PAD resulting from partially available data.

Due to technological advances, the marking for many units is now performed on-line and captured electronically via Computer Marking from Images (CMI+). This system allows examiners to mark random electronically-captured items from candidates' scripts. This has the potential to reduce the impact of severity or leniency on any individual or centre as demonstrated by an individual marker since the distributed item marking will act to cancel any such overall tendency (Pinot de Moira, 2011). This also means that the marks for all papers marked in this way have greater potential to arrive randomly making a study of PAD easier because the systematic restriction of examiners marking by centre does not ostensibly exist. The actuality of realising this random allocation of items to markers is discussed later. Furthermore, more awards are now moving towards OnLine Awarding (OLA), which is a method of awarding that can only be used by utilising CMI+. OLA is intended to be more flexible than the traditional face-to-face meetings because awarding committee members are able to scrutinise candidates' scripts during a longer time frame of several days and at their convenience. However, on the other hand, current OLA procedures require a much shorter time frame in which to derive SRBs — potentially as short as only several hours. This is because the SRBs usually need to be uploaded on the same day that the distributions become available, which is two days before script scrutiny starts. OLA was piloted and trialled in several A-level subjects in June 2011 and January 2012 and rolled out to more subjects, including selected GCSEs, in June 2012. The preparation window is restricted by the current procedure, which currently means that distributions can only be provided for award once a Minimally Required PFP threshold (MRPFP) is reached — currently 75% of candidates' marks being available. Therefore, if this MRPFP could be justifiably reduced then this would not only provide a general opportunity to run awarding data for CMI+ units sooner but also to widen the window for the derivation of SRBs.

This issue has been the focus of a few previous studies. Up until 1998, the definition of the MRPFP had been 75%. As part of a study on the use of reference statistics used in awards (Baird, 1999), it was found that the PFP had the largest impact on distributions of all the events occurring between award and results publication. Examples of other events are: late examiner adjustments, re-marks, marking review and changes to coursework moderation strategies. Thirty six A-levels and thirty GCSEs from summer 1999 offered by AQA's predecessors (AEB, NEAB, SEG) were included in this study, all of whose PFPs were greater than 70% at award. The results of this paper suggested that the PFP did not predict the changes in outcomes between award and results publication for any judgemental grade boundary. However, it was found that if the PFP is low in a particular year, then the PAD itself would differ from that of the previous year. Following this study, but not as a consequence of it, the definition of the MRPFP was increased to 85% in 1999.

In a 2000 report (Baird et al., 2000) it was noted that the summer 1999 results statistics suggested that the new 85% threshold had had a positive impact on controlling differences in distributions from year to year. However, in many subjects this percentage had been difficult to achieve in time which consequently resulted in narrow windows for awarding meeting preparation. All summer 1999 A-levels and GCSEs were investigated for PAD as measured by change in percentage outcome between award and provisional results. It was found that between 75% and 88% of subjects resulted in no significant PAD as defined by the guidance limits at the time. Between 3% and 8% had PAD resulting in a move from within statistical guidance limits at award to without these limits at provisional

[2]To assist in managing the risk introduced by such an effect, CERP implements a Pre-Results Checking Procedure seeking to identify any instances where PAD may have compromised the outcomes potentially suggesting that a change to grade boundary position(s) may be appropriate. For details of this, see the AQA internal document 'Pre-Results Checking Procedures'.

results. Between 1% and 3% had PAD resulting in the opposite — a move from without guidance to within guidance.

Later in 2004, empirical and simulated PAD were investigated, measured by changes in percentage outcome (Dhillon et al., 2004). The empirical findings for thirty eight A-level and GCSE subjects in summer 2003 found that there was no significant relationship between PAD and PFP. Even though in the extreme case only 45% of candidates had been FP at award, it is noted that the 85% MRPFP of 2003 had artificially restricted the data used which may have led to this non-significant result. The majority of subjects had a PAD of less than 1% although it was clear that factors other than PFP for some subjects had contributed to the observed changes, with marking review having the next largest contribution.

The simulated findings of the 2004 study for thirteen A-level and GCSE subjects were produced by excluding various proportions of centres from the complete FP summer 2003 distributions. This was done either contiguously or randomly, depending on whether the mark data was originally allocated geographically or randomly, respectively. For each judgemental grade boundary and each PFP, many thousands of samples were taken from the full dataset. The mean of the cumulative percentage outcomes at the grade boundary, along with a 99% confidence interval, was compared to the outcome on the final FP distribution. Unlike the empirical findings, there was found to be a highly significant relationship between PAD and PFP in the form of an asymptotic relationship as illustrated by the example curve in Fig. 1. This significant result was likely to have been a consequence of being able to simulate all PFP values unlike with the empirical data. In all cases, it was found that in order to achieve a PAD no greater than 1%, the MRPFP was always less than 67%. Moreover, in extreme subjects the MRPFP was as small as 2% in order to achieve a PAD of no greater than 1%. There was therefore considerable variation across subjects although the simulations demonstrated a surprising level of stability as measured by PAD. The results of this study suggested that the MRPFP of 85% in 2003 was overly conservative. As a consequence of this work, the authors recommended 70% as the new MRPFP and ultimately 75% was chosen which is the value in use today.

Another motivation for this research was the lack of a rigorously substantiated PFP threshold. The simulations discussed above required the knowledge of the complete FP distribution, whereas in reality only a smaller sample is available at the time of an award. This raised the question of whether anything useful could be inferred about the full distribution from the sample distribution. Since the investigation of 2004 (Dhillon et al., 2004), there has been the introduction of CMI+, which, as mentioned earlier, means that the systematic restriction of examiners marking by centre does not ostensibly exist. However, even though this random allocation of items is not always present, which is discussed later, this investigatiion does provide a framework which can potentially be built upon to address such subtleties. Additionally, it has to be considered whether the investigation of PAD at subject level is possible. When a specification is linear, as most were in 2004, the subject mark distribution is known before any grade boundaries have been determined. When a specification is modular, as most are now, the unit marks for candidates can come from several different exam series which makes the subject mark distribution more complex as it is dependent on the grade boundaries and the conversion to the uniform mark scale (UMS)[3]. In either case, given that an investigation of PAD on a single unit is complex in itself, at this stage it makes more sense to consider just the PAD as defined on individual units rather than at the subject level. Even if controlling the subject level outcome is the focus of the award, minimising the PAD on the units is likely to also minimise the PAD on the subject also. For example, it might be intuitively expected that the contribution of the unit level PAD from one unit of an n-unit specification to the subject level PAD is proportional to $1/n$. Although this is clearly an over simplification, this may potentially be more applicable to the subject level mark

[3] For details of the UMS scheme, see the AQA booklet (2011).

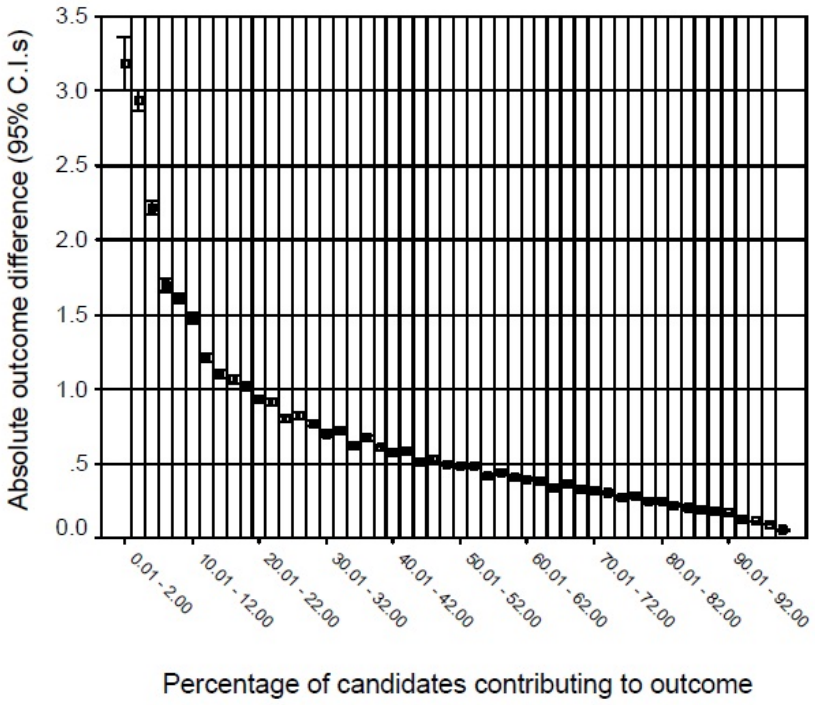

Figure 1: Example from (Dhillon et al., 2004) of randomly simulated PAD for English Literature A from 2003 with a total entry of 162,405.

distributions of future linear specifications, although in this instance there may be the opportunity to evaluate PAD on the subject level mark distribution itself.

In this paper, an analytical representation of PAD at the unit level is described. To facilitate this representation, the FP unit level marks are assumed to be modeled by a normal distribution with mean μ and variance σ^2. A method based on the use of simultaneous confidence bands (Cheng and Isles, 1983) is used to estimate this FP cumulative mark distribution given a sample distribution with PFP $= f\%$, mean $\bar{x}$ and standard deviation (SD) s. This theoretical background is explained in the next section. The method requires a total entry (the population), the PFP on distribution at award (the sample), the sample mean and sample SD and a chosen level of confidence (say 90%). The magnitude of PAD is defined as the difference between the grade boundary mark at award and the mark at results publication, which is allowed to be continuous. The method results in analytically being able to determine the MRPFP such that a statistically defined grade boundary mark would not change upon provisional results, with a certain confidence. An extension of this approach is discussed to illustrate the impact that additional information, which becomes available when the sample distribution is produced in the lead up to an award, may have on the confidence with which one may have in grade boundary marks that are determined from incomplete data. This may be useful when considering whether it is appropriate to proceed with awarding preparation when the PFP is less than the identified MRPFP. The intention of this work is to outline an analytical framework by which it is possible to clearly identify the impact of different factors on potential PAD. Applications of this current model and possible extensions to it are also discussed.

Theoretical background

A unit has a total entry which is also called here the population size N. The mean mark of the total entry is μ and the variance is σ^2 and these will be unknown at the time of an award. These two parameters form the model parameter set $\theta = (\mu, \sigma^2)$. The method of evaluating the potential extent of PAD consists of three stages:

1. construct a Confidence Region (CR) within which the model parameter set θ is likely to fall based on a sample's distribution and total entry
2. construct a Confidence Band (CB) around the sample cumulative percentage mark distribution within which the cumulative percentage mark distribution is likely to falll, where both distributions are represented as continuous probability density functions (CDFs)
3. for a chosen cumulative percentage, construct a plot of PFP versus potential boundary mark change due to PAD to inform the MRPFP.

The detailed derivations for these are presented in Appendices A, B and C, respectively, but a more general overview of them is presented below.

The Confidence Region (CR)

The CR, as computed in a similar way to the procedure set out in a previous study (Arnold and Shavelle, 1998), is a region in the parameter space θ where the population mean and SD are likely to belong with confidence γ, which is guaranteed by the underlying construction. This is dependent on the mean and SD of the sample, the total size of the population, the PFP and the specified level of confidence. As shown in Fig. 2, it has the shape of a horizontal slice of a (positive) parabola surrounding the sample mean $\bar{x}$ and sample variance s^2. With decreasing sample variance, increasing population size or increasing PFP, the region becomes asymptotically rectangular as the area decreases. In Fig. 2 the CR and its bounding curves are shown for total entry $N = 200$, PFP $f = 0.7$, sample mean $\bar{x} = 40$ and sample SD[4] $s = 16$ with confidence $100(1 - \gamma) = 90\%$. This value of entry size is unlikely to be of interest because, for such small units, the PFP usually nears 100% at award. Furthermore, for such entry sizes the theory begins to break down due to the assumptions made within it. However, it is shown here because for smaller sample sizes the shape of the CR is more apparent.

The Confidence Band (CB)

The marks available at time of sampling are formed into the sample cumulative mark distribution. These are used when deriving 'Statistically Equivalent Boundaries' (SEBs) for unit level boundaries when preparing for an award. Through simple scaling, this distribution forms the sample CDF, $F(x)$. The CB, which is directly computed from the CR, is a region bounded between two 'S'-shaped curves which totally enclose the sample CDF $F(x)$[5]. It is the band in which the population CDF is likely to lie with confidence γ. An example directly corresponding to the CR in Fig. 2 is shown in Fig. 3. Horizontally, it is thinner at the sample mean $\bar{x}$ and wider at the extremes of the mark range. With decreasing sample variance, increasing population size or increasing PFP the band more tightly wraps the sample CDF. This directly corresponds to the behaviour of the CR in these limits. It should be noted that

[4] The values of SD, rather than the variance, are given throughout because the SD is the usual measure of scale used in awarding procedures.

[5] It should be noted that the cumulative mark distributions generally used in the industry quote percentages of candidates achieving a given mark or higher. For coherence with the founding mathematical theory on which the approaches applied here are based, the conventional definition of percentiles is applied — the percentage of candidates achieving below a given mark.

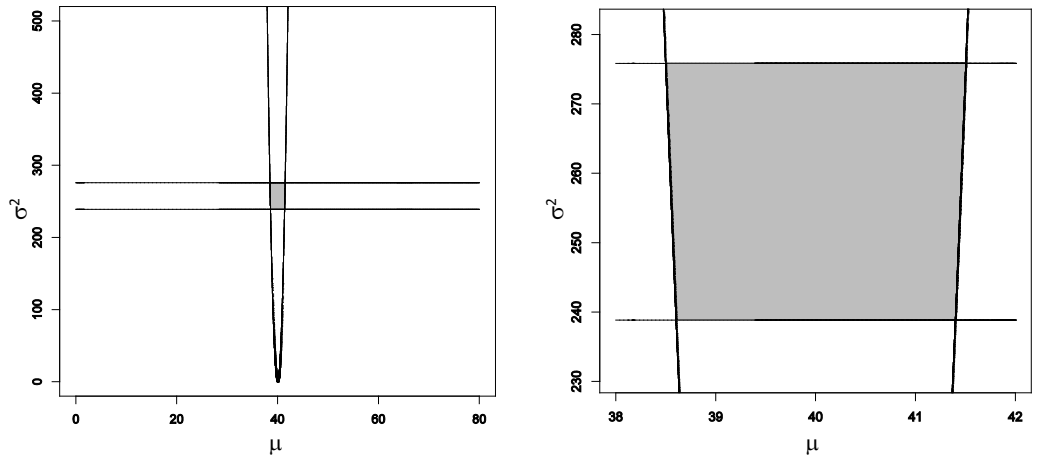

Figure 2: The Confidence Region (shaded) in (μ, σ^2) space for $N = 200$, $f = 0.7$, $\bar{x} = 40$, $s = 16$ and confidence $100(1 - \gamma) = 90\%$. *Left:* The region is bounded by three curves: a quadratic curve and two horizontal lines. *Right:* The CR zoomed in.

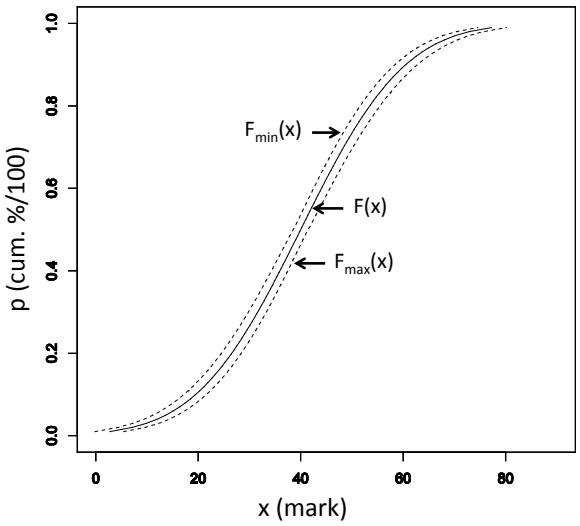

Figure 3: The confidence band for $N = 200$, $f = 0.7$, $\bar{x} = 40$, $s = 16$ and confidence $100(1-\gamma) = 90\%$. The band is the region between the two curves F_{min} and F_{max}.

the width of the CB around the sample CDF is slightly asymmetric ($F(x) - F_{\text{min}} \neq F_{\text{max}} - F(x)$) . This asymmetry is minimal and its presence is considered below. It should also be noted that the joint confidence level γ does not give a probability that the total entry CDF will lie in this band. The parameter set θ is fixed but the confidence region from which the band was generated is random as it depends upon the (random) sample. However, it does give a frequentist result stating that, should a sample of equal size be taken many times, then in $100(1-\gamma)\%$ of the samples the total entry CDF would lie in the confidence band.

The Post-Awarding Drift

Prior to an award, the percentage of candidates expected to obtain a particular grade may be known via a prediction for that grade or some other statistical reference. The chosen mark for that grade boundary is determined from the sample CDF at the time of an award. From the CB located around this sample CDF, a plot can be produced of the maximum possible deviation (up to the confidence level) from this chosen mark, for a range of PFP values. This variation in mark is used as a measure of potential PAD in boundary mark.

At a particular value p in the CDF percentile range, PAD $\Delta_p(f)$ is defined as the difference between the corresponding marks on the sample CDF F and one of the CB boundaries, F_{min} or F_{max}. In the appendix, this is derived as

$$\Delta_p(f) = \Psi^{-1}(p)\sigma + \sigma_+(f,N)\left(\frac{z(\alpha_1)c_\mu(f,N)}{\sqrt{fN}} - \Psi^{-1}(p)\right), \tag{1}$$

where $\Psi^{-1}(\cdot)$ is the inverse CDF of the standard normal distribution, $z(\alpha)$ is an upper percentile of the standard normal distribution, $\sigma_+(f,N)$ is the lower horizontal boundary of the CR and c_μ is a finite population correction factor — these terms and their relevance are explained in the appendix. An example of the estimates of potential PAD is shown in Fig. 4, using the same parameters as the CR and CB shown in Figs. 2 and 3, respectively, except for $N = 2000$. The use of the small population size ($N = 200$) in the earlier part of this section was for illustrative purposes only. This is not appropriate here due to the instability of the PAD plot as N approaches 100.

The PAD curves for $p \in (0, 0.5]$ lie almost coincident with the the PAD curves for $p \in [0.5, 1)$. Fig. 4 shows the PAD across the percentile range from $p = 0.01$ to $p = 0.5$. This (somewhat arbitrary) choice of percentile range has a negligible effect on the PAD plot because the asymmetry directly arises from the asymmetrical shape of the CR and, therefore, asymmetry in the CB described above, which will tend to a small rectangle (and hence symmetry) as n or N increases.

The plot shows that in the limit of the sample size going to the total entry ($f \to 1$), the PAD goes to zero across the whole CDF, as would be expected since all the marks are known. For decreasing $f \to 0$, the figure shows that the PAD increases, but also that it increases more in the tails of the CDF as $p \to 0$. It might be intuitively expected that the PAD would be more stable in at least one of the tails because 100% of candidates would always be expected to gain a mark of 0 or higher regardless of the PFP. The apparent discrepancy results from a continuous mark scale modeling a discrete one. However, in the other tail the instability is to always be expected because it is the unknown population variance that determines the number of candidates achieving the maximum mark. The behaviour of PAD as $f \to 0$ is $\Delta_p(f) \sim 1/\sqrt{f}$ — inverse square root behaviour. This appears to explain the asymptotic behaviour observed (Fig. 1) in the simulated findings of the previous investigation (Dhillon et al., 2004). Although the y-axis in Fig. 1 measures PAD by percentile and the y-axis in Fig. 4 measures PAD by quantile, these are related and amount to the same concept.

The model described works with a continuous mark scale. Clearly, in practice, these mark scales

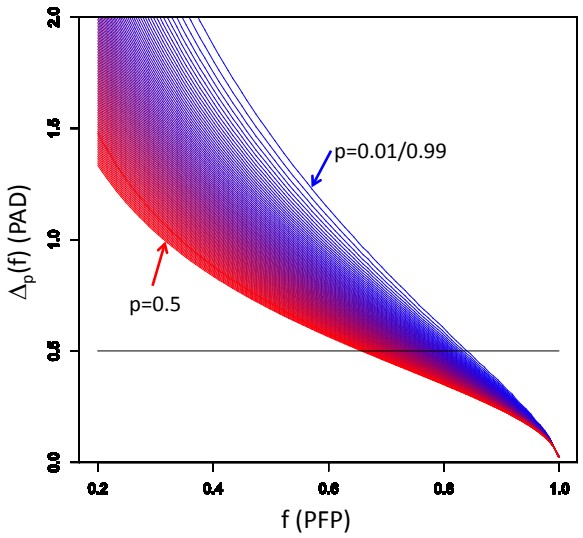

Figure 4: The PAD for $N = 2000$, $\bar{x} = 40$, $s = 16$, $\gamma = 0.1$, $p \in (0, 0.5)$. As p increases from 0 to 0.5, the PAD decreases in magnitude. As p increases from 0.5 to 1, the PAD would increase again almost retracing the curves for $p \in (0, 0.5)$.

use discrete whole number marks. When using statistical guidance to establish the position of a unit level SEB, the selected integer mark is the one that has the percentage outcome that is closest to the statistical recommendation. This is equivalent to linearly interpolating between two consecutive marks exactly to achieve a prediction/reference and then rounding the resulting decimal valued mark to the nearest mark. Therefore, in the general continuous sense, it is initially assumed that the maximum possible deviation in mark x due to PAD, that would not have resulted in a different SEB being selected, is half a mark. A deviation greater than the deviation criterium $\delta x = 0.5$ at a particular PFP value indicates a greater likelihood of a mark change due to PAD and hence that the chosen mark is more likely to change once all marks have been processed ($f = 100\%$). Fig. 4 shows that the intersections of the individual p-curves with the line $\Delta_p(f) = 1/2$ varies. Hence the MRPFP suggesting stability of the mark varies across the grade range since it varies with p. In Fig. 5 the same plot is produced for a single value of $p = 0.5$. This shows that should a grade boundary have a prediction or reference outcome of 50%, then the MRPFP needed in order that the (continuously defined) boundary will, with 90% confidence, not drift by more than ± 0.5 marks post-award is the intersection point $f_{\text{min}} = 67\%$.

When the potential PAD at one grade boundary, say grade A, is of interest the PAD and MRPFP can be evaluated for the percentile corresponding to the prediction for that grade, $p = p_A$. Due to the variation in potential PAD across the percentiles, in practice it would be necessary to select a single threshold percentile $p = p_{\text{thresh}}$ on which to evaluate an MRPFP. Ideally, this value would be chosen such that the potential PAD is maximised, therefore, providing a slightly conservative value of MRPFP for most grade boundaries. The point at which this occurs is where the difference between the sample CDF and the confidence band is greatest in terms of difference in mark (i.e. horizontally in Figure 3). However, in the limits as $p \to 1$ and $p \to 0$, the asymptotic nature of the relationship tends

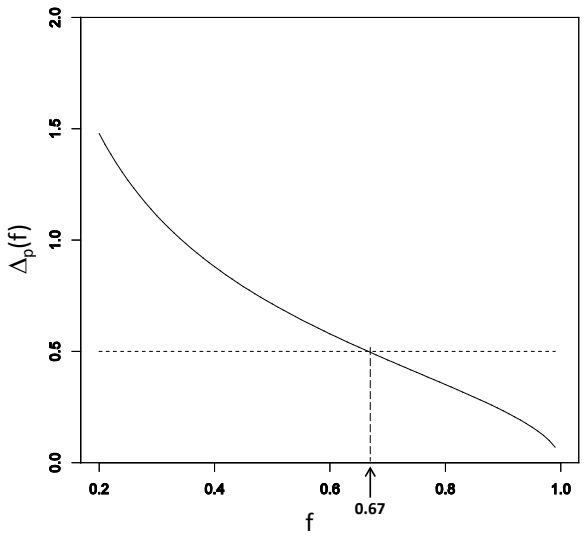

Figure 5: The PAD for $N = 2000$, $\bar{x} = 40$, $s = 16$, $\gamma = 0.1$, $p = 0.5$. This is one of the curves in Fig. 4.

to infinity[6]. The threshold percentile $0 < p_{\text{thresh}} < 0.5$ should be chosen such that all the percentiles of interest (that is, the percentiles that correspond to all the grade boundary outcomes of interest) fall between p_{thresh} and $1-p_{\text{thresh}}$. Since the potential PAD in Fig. 4 is monotonically increasing in the tails of the distribution, this means that the estimation of the MRPFP is maximised at p_{thresh} and $1-p_{\text{thresh}}$. Furthermore, the threshold removes issues with the unstable behaviour in the tails of the distribution, i.e. when sampling, it removes the unrepresentative effect of small numbers of candidates at the extremes of the mark range. Incidentally, for $p_{\text{thresh}} = 0.5$, used for demonstration purposes above, the value of the corresponding MRPFP is at its lowest across the range of p. Practically, therefore, it is important to evaluate the MRPFP between the limits defined by p_{thresh}.

Due to this approach acting on a continuous mark distribution is it not sensitive to the proximity of the predicted/reference outcome relative to outcome at the SEB. This information becomes available when preparing for an award once the specifics of the sample CDF are known and is potentially of great interest when evaluating the likelihood of the SEB moving due to PAD. However, when making general decisions about the MRPFP for a unit without this level of detail, the assumptions made above are necessary and provide sufficient approximation. This theory will later be used to create many sample distributions of a known total entry population. Each one of these samples represents a potential awarding sample. For the population and each sample, the continuous version of the mark for a predicted/reference outcome, as calculated by linear interpolation, should be determined. The distribution of these sample marks would indicate that in $100(1-\gamma)\%$ of cases the mark is no further than $1/2$ from the population SEB. However, because the SEB is actually an integer, the position of the (continuous) SEB from the awarding sample should be considered with respect to integer marks. If the population SEB is near to an integer mark, then the $\delta x = 1/2$ deviation criterium set in the theory accurately models the likelihood of an SEB moving due to PAD. However, if the population SEB is close to a half-integer mark, then the $\delta x = 1/2$ deviation criterium is less robust because there is

[6]Fig. 4 should technically show the PAD for all values of $p \in [0, 1]$. The reason this cannot, and should not, be done is because it is meaningless to measure the PAD at such extreme percentiles, where also the numerical procedure would break down. In the limits of $p \to 0$ or $p \to 1$ the PAD is a vertical line at $f = 1$, indicating that no matter what is done, $(1-\gamma)\%$ confidence can only be achieved when 100% of marks are processed. This is tautological because with 100% FP marks there would always be 100% confidence. The apparent inconsistency is actually due to the fact that a continuous mark scale model is being used to model a discrete one.

essentially a 50% chance that the SEB will be rounded to the neighbouring integer due to PAD. With this information, reflection should be made on whether the potential PAD may impact on the position of an SEB as this would then affect the preparation for an award.

Results

Analytical results

By way of demonstration, the method was applied to several simulated units. The only variable changed was size of total entry N as this is the variable that changes most from unit to unit across specifications. The fixed variables were the PFP, sample mean and SD and confidence for the CRs and CBs and the sample mean and SD, confidence and percentile for the PAD plots. Figure 6 shows the CRs for several values of N from 200 to 200,000. This covers everything from the smallest units that use statistical guidance (e.g any of the GCSE Panjabi units) to the largest units, in particular the GCSE English Unit 1 tiers ENG1F and ENG1H. It can be seen that the CR shrinks and becomes more rectangular with increasing N. The corresponding CBs for the same values of N are shown in Figure 7, which shows how the bands shrink accordingly around the sample CDF indicating that the FP CDF is increasingly likely to be contained near the sample CDF with increasing N. Finally, the corresponding PAD plots, again for the same values of N, are shown in Figure 8 for $p_{\text{thresh}} = 0.5$[7] and all values of the PFP. The intersection of these curves with the line $\Delta_p(f) = 1/2$ then gives the key value f_{min} — the MRPFP needed in order that with $100(1-\gamma)\% = 90\%$ confidence the mark corresponding to the 50% percentile on the FP CDF will not deviate by more than half a mark from the same mark on the sample CDF. The plots show that f_{min} decreases with increasing N.

Equation (1) can be used to analytically calculate the values of f_{min} for various scenarios which can then be tested by simulation here, by way of a validation. Simulations were performed for $\mu = 40$, $\sigma = 16$ and $p_{\text{thresh}} = 0.01, 0.1, 0.5$ with confidence $100(1-\gamma) = 90\%$. The value of $p_{\text{thresh}} = 0.1$ was chosen low enough such that it would encompass, for example, the judgemental grades of A and E of a typical A-level unit, but also not so low as to be too close to the maximum resolution of p, which is related to the standard deviation in a Poisson process: $1/\sqrt{N}$ (for $N = 200$, the maximum resolution is $1/\sqrt{200} = 0.07$). Population sizes ranging from $N = 1,000$ to $N = 400,000$ were considered. A plot of the corresponding f_{min} values is shown in Fig. 9. This confirms that the dependence of f_{min} on N is exponential as $N \to \infty$. Also shown in Fig. 9 are the corresponding sample size $n = N f_{\text{min}}$. This indicates that for the value of $p_{\text{thresh}} = 0.1$, there is a sample size of $n \approx 12,000$ beyond which little more information can be gained. As p decreases, this limiting value of the sample size increases so that it is $n \approx 25,000$ for $p_{\text{thresh}} = 0.01$. The values of $p_{\text{thresh}} = 0.01, 0.5$ were chosen to show how f_{min} is affected by p_{thresh}. These show that a very small choice of p_{thresh} increases f_{min} but not excessively so — for $N = 400,000$ for example, f_{min} only doubles from 3% to 6% when p_{thresh} is reduced by a factor of 10 from 0.1 to 0.01. For $p_{\text{thresh}} = 0.1$, and for each value of N, 10,000 simulations were performed and the number of simulations for which the PAD was no greater than $\Delta = 1/2$ was between 92% and 97%. These values clearly exceed the chosen 90% confidence value, but this can be attributed to the conservativeness of the CRs. They are likely to be conservative given that no optimisation of these regions has been considered. Therefore, the resulting CBs are likely to be wider than necessary and hence the derived values of the MRPFP are also likely to be generous.

[7] The choice of $p_{\text{thresh}} = 0.5$ here is arbitrary and only for purposes of demonstration.

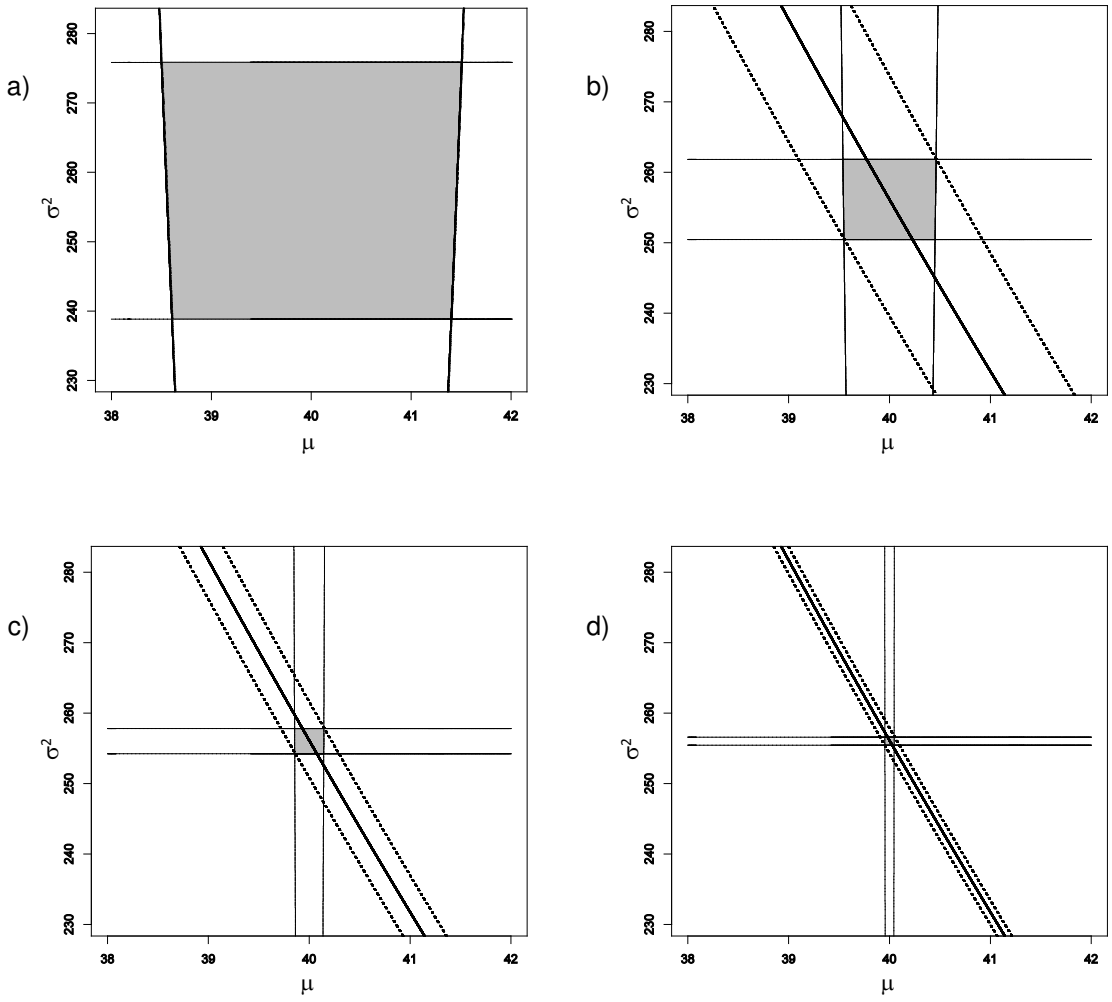

Figure 6: The confidence regions for $N = 200, 2000, 20000, 200000$ (a to d), $f = 0.7$, $\bar{x} = 40$, $s = 16$, $\gamma = 0.1$. The quadratic curves, which appear as straight lines on three of the plots, are used to calculated the CBs, for $p = 0.1$ in this case (see Appendix).

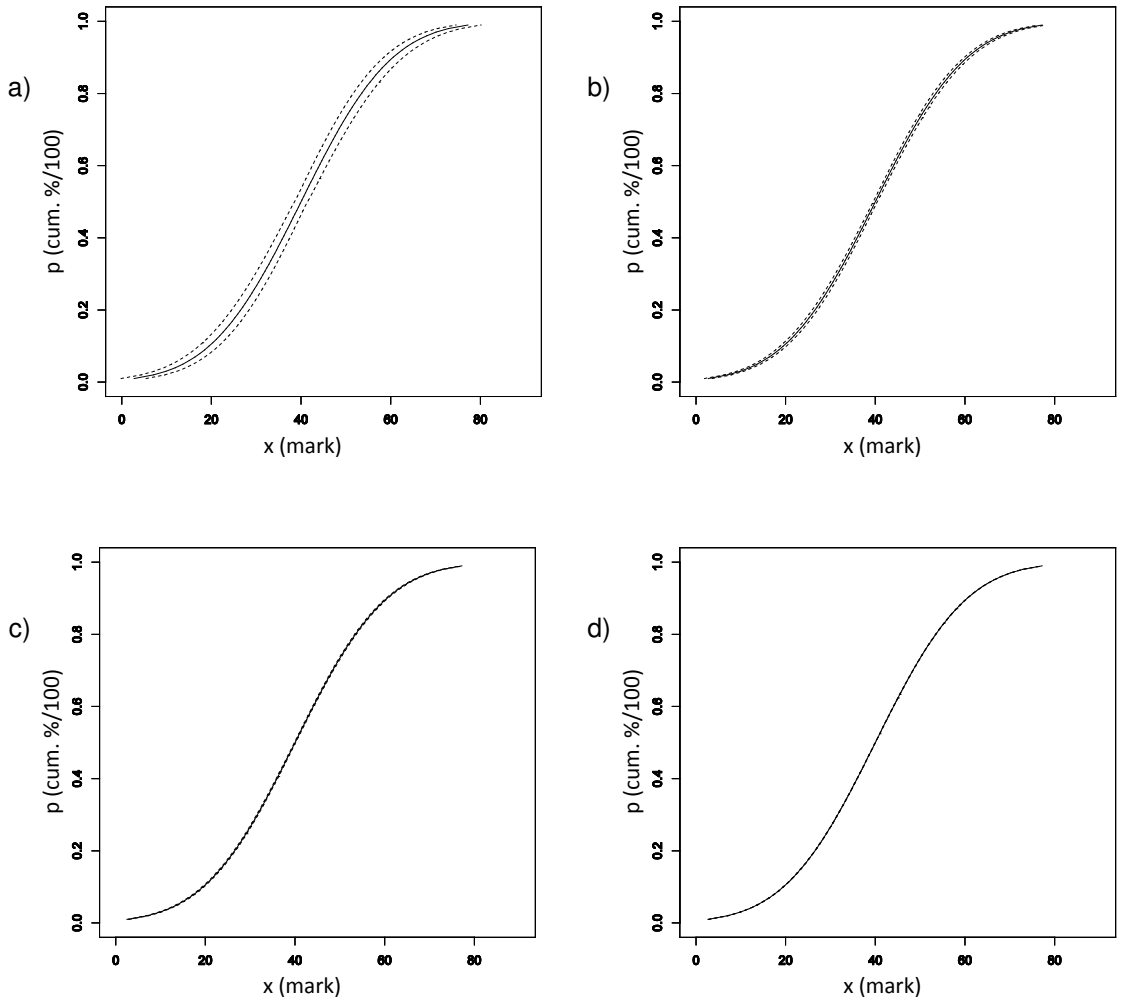

Figure 7: The confidence bands for $N = 200, 2000, 20000, 200000$ (a to d), $f = 0.7$, $\bar{x} = 40$, $s = 16$, $\gamma = 0.1$.

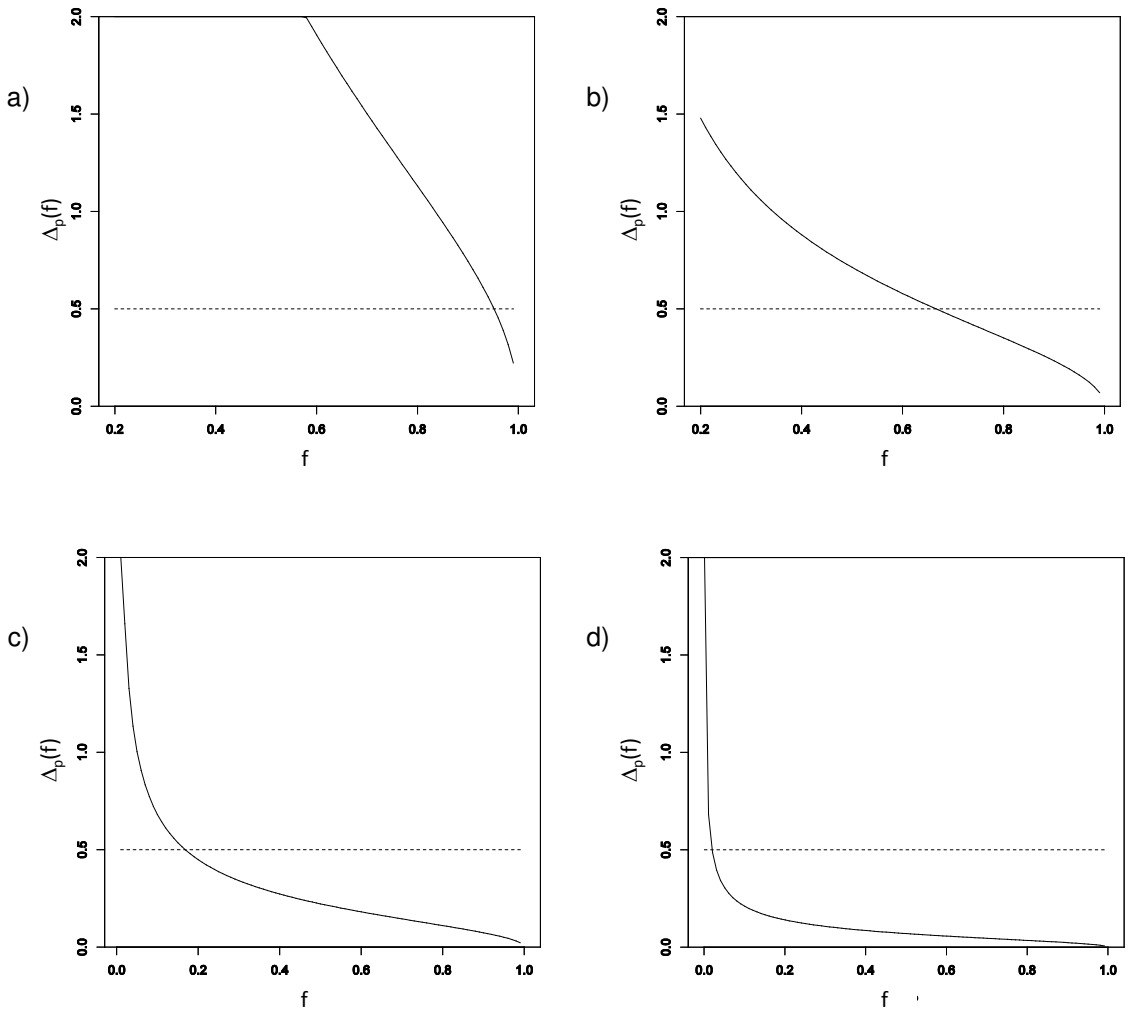

Figure 8: The PAD for $N = 200, 2000, 20000, 200000$ (a to d), $\bar{x} = 40$, $s = 16$, $\gamma = 0.1$, $p = 0.5$.

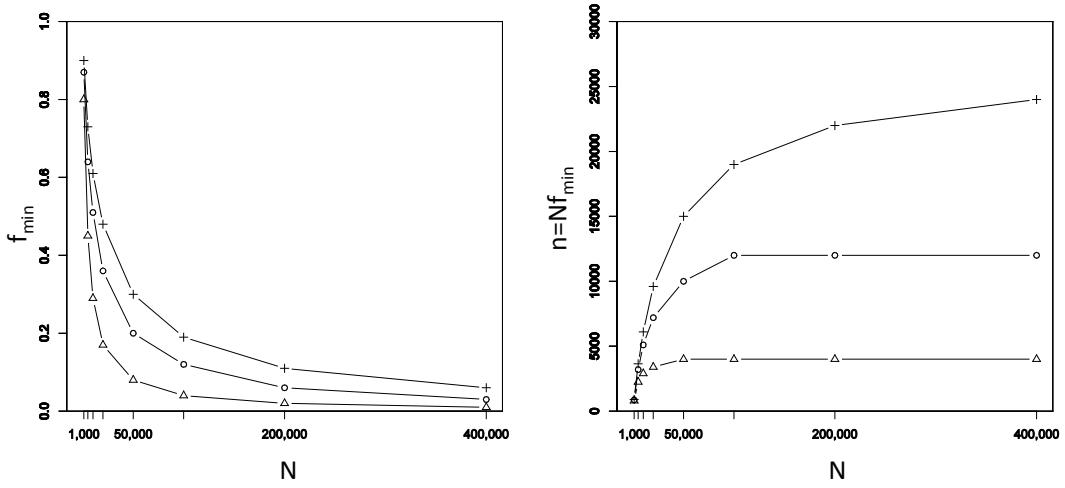

Figure 9: Left: The MRPFP values for the populations used in the simulation testing for $p_{\text{thresh}} = 0.5$ (Δ), $p_{\text{thresh}} = 0.1$ ($\circ$) and $p_{\text{thresh}} = 0.01$ (+). Right: The corresponding sample sizes n for the same p_{thresh}-values.

Empirical results

The theory was applied to actual exam data for three units. A small unit, a large unit and a unit that was rerun after award were considered. Additionally, the method was explored during the summer 2012 awarding series for several units where a problem was foreseen. These are discussed in the following subsections. It should be noted that, here, the assumption is that it does not matter exactly where a particular percentile sits between whole number marks. As a continuous distribution is being used to model a discrete one, all that is important is the percentile value itself. In all of the below, the empirical sample distributions are tested for whether they deviate by more than $1/2$ a mark from the population distribution. These are not then 'snapped' to the nearest whole mark because that is not how the theory was, or could be, defined. A further analysis is then conduced where this assumption is not made and the sample distribution of continuous marks and where it lies with respect to the integer marks is considered.

A small unit

Unit 1 of A-level Philosophy (PHIL1) in summer 2010, with a maximum mark of 90, had a total entry of $N = 5,657$ with a FP mean of $\mu = 40.20$ and a FP SD of $\sigma = 15.57$. The award was run with the PFP at just under 100%, but the idea here is to illustrate how the award could theoretically have been run at a lower PFP. The A grade boundary prediction for this unit was 16.15%, corresponding to a mark of 55. This translates to a mark of $x_A = 54.94$ on the continuous mark scale for the population. Following through the theory based on these figures gives the plot shown in Fig. 10. This shows the the MRPFP is $f_{\text{min}} = 57\%$. This indicates that if many different samples of size $n = 3,224$ of the PHIL1 total entry were taken, then the FP CDF would lie totally within in the sample CB on 90% of occasions. This is equivalent to saying that on 90% of occasions the PAD from x_A would differ by no more than $1/2$. A simulation was run for 100,000 samples of this size and in 94,447 cases (94.5%) the A boundary would not have drifted by more than half a mark from x_A.

In addition, an evaluation of the location of the distribution of continuous marks with respect the integer mark scale was made. For the above samples, the distribution of the continuous marks for

the x_A percentile $p = 0.1615$[8], is also presented in Fig. 10. This shows that 92.4% of the marks lie in the range $[54.5, 55.5]$, indicating that it is still very unlikely that the SEB would have changed from 55 as a result of PAD. This is due to the proximity of x_A to the whole integer mark.

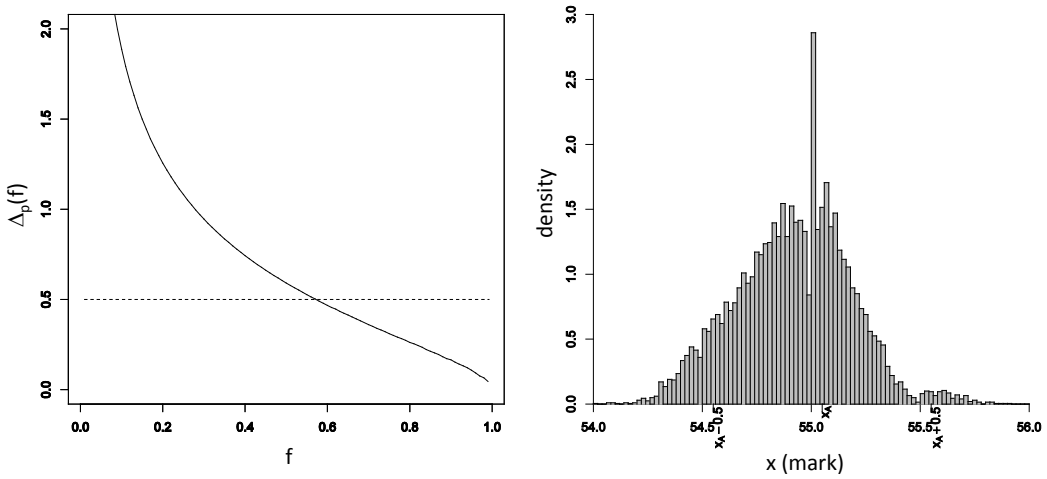

Figure 10: *Left:* The PAD for the grade A prediction on PHIL1 in June 2010. The intersection with the line $\Delta_P(f) = 1/2$ is at $f_{\text{min}} = 0.57$ ($100(1-\gamma) = 90\%$, $p = 0.165$). *Right:* The sample continuous mark distribution for 100,000 samples of size 57% of the total entry of PHIL1 at grade A ($p = 0.165$).

A large unit

The coursework component 3702/CS of legacy GCSE English A was one of the biggest components/units provided by AQA. It had a maximum mark of 54 and a total entry of $N = 377,800$ in summer 2011 (this component has now been replaced by unit ENG02, with a similar total entry for summer 2012). The FP mean was $\mu = 37.86$ and the FP SD was $\sigma = 8.26$. This component was split across two tiers of entry, but the marks were combined for the purposes of this analysis. The combined C grade outcome was 87.2%, corresponding to a mark of $x_C = 29.74$ on the combined continuous mark scale for the population (which would have been rounded to 30 had this been an SEB). Following through the theory based on these figures gives the plot shown in Fig. 11. This shows that the MRPFP is $f_{\text{min}} = 0.93\%$. This indicates that if many different samples of size $n = 3,514$ of the 3702/CS total entry were taken, then the FP CDF would lie totally within in the sample CB on 90% of occasions. This is equivalent to saying that on 90% of occasions the PAD from x_C would differ by no more than $1/2$. A simulation was run for 10,000 samples of this size and in 9,540 cases (95.4%) the C boundary would not have drifted by more than half a mark from x_C.

As above, for these samples the distribution of the continuous marks for the percentile $p = 0.872$ is also presented in Fig. 11. This shows that 88.2% of the marks lie in the range $[29.5, 30.5]$, indicating that, unlike above, it is more likely that the SEB would have changed from 30 as a result of PAD than the confidence would suggest. This is due to the closer proximity of x_A to the half-integer mark. In this scenario it is more likely that there would be a change to SEB which would consequently have an effect on the rank order of candidates. Therefore, some level of caution should be adopted in this case, perhaps by increasing the confidence used in the calculation of the MRPFP.

[8]Note that because the CDFs used in the industry are calculated from the highest to lowest mark, the values used numerically for these empirical tests are $1 - p$ in order to obtain agreeing boundary marks.

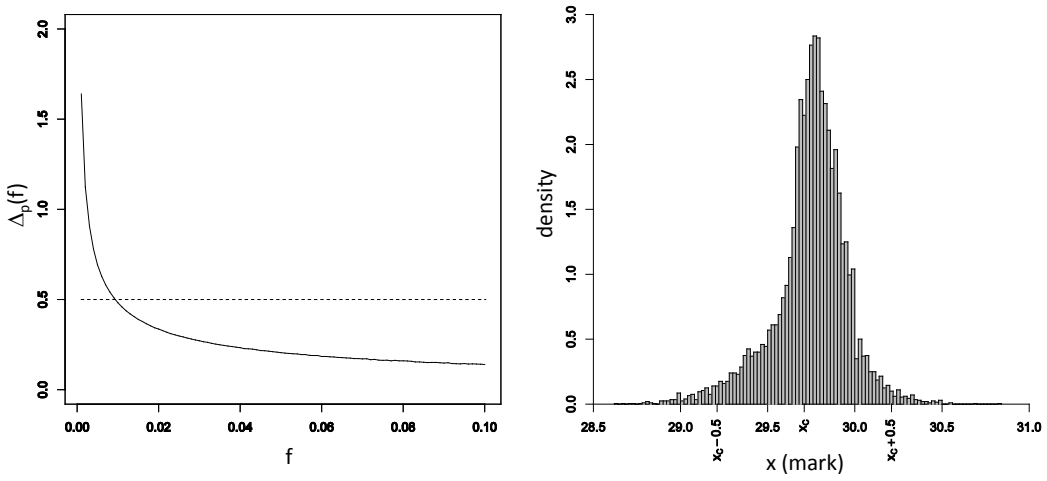

Figure 11: *Left:* The PAD for the grade C prediction on tier-combined 3702/CS in June 2011. The intersection with the line $\Delta_p(f) = 1/2$ is at $f_{\text{min}} = 0.0093$ ($100(1-\gamma) = 90\%$, $p = 0.872$). *Right:* The sample continuous mark distribution for 10,000 samples of size 0.93% of the total entry of 3702/CS at grade C ($p = 0.872$).

A unit requiring a rerun

Unit 405007 of GCSE Religious Studies in summer 2011, with a maximum mark of 72, had a total entry of $N = 3,051$. Only 62.2% of marks were FP at the time of the award and hence a rerun was required. This was performed ten days after the award with 99.6%, which at grade A (a mark of 58) resulted in a 2.26% downwards PAD. The mean and SD at award were $\bar{x} = 48.6$ and $s = 16.7$, respectively. The Grade A boundary prediction was 32.51% and resulted in the boundary mark of 58 at award. The PAD plot for this unit showed showed that $f_{\text{min}} = 79\%$. This indicates that if many different samples of size $n = 2,410$ of the total entry were taken, then the FP CDF would lie totally within in the sample CB on 90% of occasions. This is equivalent to saying that on 90% of occasions the PAD from x_A would differ by no more than $1/2$. A simulation was run for 10,000 samples of this size and in 9,269 cases (93.7%) the A boundary would not have drifted by more than half a mark from x_A. It is clear that on this occasion a rerun was warranted, although it could possibly have been performed earlier.

Summer 2012 exploration

During the summer 2012 awarding series, five units were considered using the methodology: GCSE Classical Civilisation Units 3F and 3H (CIV3F, CIV3H), A-level Geography Unit 1 (GEOG1), A-level Biology Unit 5 (BIOL5) and A-level Philosophy Unit 3 (PHIL3). These units were chosen because there was concern that the PFP would not hit the MRPFP of 75% by award. In contrast to using historical examinations as above, the population means and SDs were not known during this exploration. The mean does not appear in the PAD equation (1), but in this case the sample SD will substitute for the unknown population SD. Below is a summary of these five units:

- Unit CIV3F had a PFP of $f = 15\%$ but a total entry of only 132 candidates for which it was too small to draw solid conclusions, although the model suggested that a MRPFP of $f_{\text{min}} = 95\%$ was required for 90% confidence of avoiding PAD.
- Unit CIV3H, with a total entry of 1,451, had a PFP of $f = 46\%$ but the model suggested

$f_{\text{min}} = 64\%/75\%$ was required for $90\%/99\%$ confidence of avoiding PAD.

- Unit GEOG1 was running at PFP $f = 58\%$ with a total entry of 21,184, and the model suggested that $f_{\text{min}} = 12\%/22\%$ was required for $90\%/99\%$ confidence of avoiding PAD.
- Unit BIOL5, with a total entry of 25,015, had $f = 47\%$ at the initial run and $f = 70\%$ at the updated run three days later, which was still not at the current acceptable threshold of 75%. The model suggested that $f_{\text{min}} = 9\%/18\%$ was required for $90\%/99\%$ confidence of avoiding PAD.
- Finally, unit PHIL3 with total entry 3,264 had $f = 52\%$ and the model suggested that $f_{\text{min}} = 53\%/61\%$ was required for $95\%/99\%$ confidence of avoiding PAD.

This exploration was not intended to guide absolutely the decision on whether a unit was ready to have its mark data run for an award. It was used as a rough guide for evaluating the current 'state of play' of a unit. Most importantly, what it did highlight however was the need to consider other information when considering problematic units and their PFPs, specifically the total entry. The simulation results showed that the greater the total entry on a unit the less important it is to require a high PFP. On the other hand, when the total entry on a unit is small, a small PFP is less meaningful and fewer marking resources should be required in order to attempt to improve the situation (e.g. for the $f = 15\%$ on CIV3F, only 112 more candidates were waiting to be marked).

Discussion

The effect of potential PAD has here been evaluated by providing an SD, the total entry (population size), a confidence level and a percentile threshold to Equation 1. For the simulations, the SD was the population SD, but for live data the SD should be the awarding sample SD. For simulations, the percentile threshold was chosen such that it encompassed all the judgemental grades for the unit under consideration, but also not so low as to be too close to the maximum resolution of the percentiles. From the potential PAD, the theoretical value for the MRPFP was then calculated. With an awarding sample to hand, further consideration was made about where the continuous grade boundaries sit with respect to integer marks. This showed that more caution should be taken with the interpretation of the MRPFP value when grade boundaries sit near half-integer marks. However, because an SEB at a half-integer is very much a borderline case, which could change as the result of just one additional candidate on the distribution, the impact of an SEB change in this scenario is less critical and does not devalue the importance of the MRPFP as estimated by the methodology presented here.

As was noted earlier, no indication has been given about the conservativeness of the CRs within the theory, but they are likely to be conservative given that no optimisation of these regions has been considered. Therefore, the resulting CBs are likely to be wider than necessary and hence the derived values of the MRPFP are also likely to be generous. This was seen in the simulation results where the confidence level was always exceeded. Such optimisation could be performed by calculating approximating CRs and minimising the area of them, which then decreases the conservativeness (Arnold and Shavelle, 1998).

An assumption of the theory is that the sample at award is random, however this is not necessarily always the case. An example issue that was highlighted during the summer 2012 awarding series was the effect of 'candidates with additional sheets'. Such candidates will have used one or more sheets during their examination that were additional to the main writing space provided on the examination paper. These additional sheets cannot be processed for use within OLA and hence the corresponding candidates are discounted from the distributions. Given that such candidates may be

the higher performers, this adds a possible systematic bias to the distributions during the awarding process. Cases have even been discovered where the PFP was above the MRPFP of 75% but a large proportion of the remaining candidates were those with additional sheets. Because the extent of this is still unknown, further research should be undertaken to investigate this. A theoretical model that takes into account this type of bias via a skew parameter could also be developed as an extension of this current work.

A further assumption of the theory, even if the candidates with additional sheets could be accounted for, is that all of the remaining candidates' marks arrive randomly. This then implies that the effect of examiners marking by centre is assumed not to exist. In reality, however, the original candidate scripts are scanned into the CMI+ system in batch jobs and it is unclear how this is exactly performed. It is likely that this is done on a centre-by-centre basis, as and when the packs of scripts arrive. This means that in the earlier stages of marking, the scripts that are available to the examiners for marking do not form a true random sample but are from only a handful of centres. As time progresses, the sample of marks would become more and more representative of all the centres. Because the exact methods used by the company DRS who manage this process are unknown, some investigation should be undertaken to gain a better appreciation of the methods used. This is a separate issue to that discussed above of introducing skew into the population and the samples drawn from it. This issue concerns samples being systematically unrepresentative of the population. A theoretical model that takes into account this type of bias via additional centre variables could also be developed as an extension of this current work.

Conclusion

This paper has demonstrated that it is theoretically possible to run awards with fewer than the currently recommended MRPFP 75%, at unit level. There is not a single solution to the PAD problem, and it has been highlighted that it should be considered on a unit by unit basis. The main driver of the MRPFP is the total entry of the unit, but it has also been shown that information from the awarding sample should also be taken into consideration. To appropriately manage the risk surrounding the consequences of PAD, values of the MRPFP as calculated from the theory could be increased by, say, 10%. Furthermore, the confidence level could be increased from that used above to 95% or 99% to further minimise the risk of SEB changes due to PAD. As has been highlighted, there is also further research that could be undertaken to extend the approach presented here.

If the limitations are taken into consideration, the results of this work could be used in the following way: because in the PAD equation the most important aspect of the sample SD is just its order, rather than its exact value, for each unit a value for the 'representative' FP SD could be used to calculate a value of the MRPFP, across all judgemental grade boundaries within p_{thresh} and $1-p_{\text{thresh}}$. Such a representative value of the FP SD could be acquired from historical data or an 'ideal' value. This could be done in advance of any candidate scripts being marked. During awarding preparation, the actual awarding SD and the positions of continuous SEBs with respect to the integer marks could be used to further aid the decision about whether the current PFP is high enough to avoid risking grade boundary changes resulting from PAD. In the simulations it was revealed that as the population size doubled, the MRPFP eventually halved, meaning that the sample size corresponding to the MRPFP approached a fixed value (this was 12,000 for $p_{\text{thresh}} = 0.1$). This key result demonstrates that, more often than not, it is important to consider PFP values in conjunction with the corresponding awarding sample sizes.

Robert Hales
13 November 2012

References

Arnold, B. C. and Shavelle, R. M. (1998). Joint confidence sets for the mean and variance of a normal distribution. *The American Statistician*, *52*(2), 133–140.

Baird, J. (1999). *Reference Statistics for Award Meetings*. Internal Report RAC/804, Manchester UK: Associated Examining Board.

Baird, J., Eason, S., and Morrissy, M. (2000). *Post-Awarding Drift in Examination Statistics - Summer 1999*. Internal Report RC71, Manchester UK: AQA Centre for Education Research and Policy.

Cheng, R. C. H. and Isles, T. C. (1983). Confidence bands for cumulative distribution functions of continuous random variables. *Technometrics*, *25*(1), 77–86.

Cho, E. and Cho, J. C. (2008). Variance of sample variance. *Journal of the American Statistical Association*.

Dhillon, D., Eason, S., and Pascoe, J. (2004). *Percentage of marks on file at awarding: consequences for 'post-awarding drift' in cumulative grade distributions*. Internal Report RC258, Manchester UK: AQA Centre for Education Research and Policy.

Pinot de Moira, A. (2011). *Why item mark? The advantages and disadvantages of e-marking*. Internal Report CERP_11_APM_WP_003, Manchester UK: AQA Centre for Education Research and Policy.

Rice, J. A. (2007). *Mathematics Statistics and Data Analysis*. Belmont, CA, USA: Brooks/Cole.

Appendix A: The Confidence Region

For a sample of size n from a $N(\mu, \sigma^2)$ distribution, the Central Limit Theorem says that

$$\frac{\bar{X} - \mu}{\sigma/n} \sim N(0,1)\,, \tag{2}$$

where $\bar{X}$ is the sample mean

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3}$$

and that

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}\,, \tag{4}$$

where s^2 is the sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\,. \tag{5}$$

Stating (2) using probability notation means that

$$P\left(-z(\alpha_1/2) < Z < z(\alpha_1/2)\right) = 1 - \alpha_1\,, \tag{6}$$

where $z(\alpha_1/2)$ is the upper $\alpha_1/2$-th percentile of the standard normal distribution function $\phi(z)$. This results in the usual $100(1-\alpha_1)\%$ confidence interval (CI) for the population mean μ as

$$C_\mu = \left[\bar{X} - \sigma_{\bar{X}} z(\alpha_1/2), \bar{X} + \sigma_{\bar{X}} z(\alpha_1/2)\right]\,, \tag{7}$$

where $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ is the standard error (SE) of the sampling mean distribution. Similarly for the variance in (4), the relationship

$$P(-z(\alpha_2/2) < Z < z(\alpha_3/2)) = \alpha_2 - \alpha_3 \tag{8}$$

yields the CI for the population variance σ^2 as

$$C_{\sigma^2} = \left[\frac{(n-1)s^2}{z(\alpha_3)}, \frac{(n-1)s^2}{z(\alpha_2)} \right] , \tag{9}$$

where $z(\alpha_2)$ and $z(\alpha_3)$ are the lower α_2-th and upper α_3-th percentile of the χ^2 distribution, respectively. It is usual to use equal tail precisions in the χ^2 distribution, such that $\alpha_2 = \alpha_3$. For small sample sizes, $n < 50$, this would not be appropriate as it has been shown by trial and error that for suitably large sample sizes, $n > 50$, it is optimal to choose $\alpha_2 = \alpha_3$ (Arnold and Shavelle, 1998). This is because the χ^2_n distribution is standard normally distributed as $n \to \infty$. Because the χ^2_n distribution has mean n and variance $2n$, this means that

$$\frac{\chi^2_{n-1} - (n-1)}{\sqrt{2(n-1)}} \sim N(0,1) . \tag{10}$$

for large n. In probability notation this implies that

$$P\left(-z(\alpha_2/2) < \frac{\frac{(n-1)s^2}{\sigma^2} - (n-1)}{\sqrt{2(n-1)}} < z(\alpha_2/2) \right) = 1 - \alpha_2 . \tag{11}$$

Solving this gives an alternative CI for the population variance for large $n > 50$ as

$$C_{\sigma^2} = \left[\frac{s^4}{s^2 + z(\alpha_2)\sigma_{s^2}}, \frac{s^4}{s^2 - z(\alpha_2)\sigma_{s^2}} \right] , \tag{12}$$

where $\sigma_{s^2} = s^2\sqrt{2/(n-1)}$ is the SE of the sampling variance distribution. This form of CI for σ^2 is useful because it explicitly contains the SE, σ_{s^2}.

For a finite population N, which corresponds to the total entry on a unit, a finite population correction (PFC) has to be applied to the SEs appearing in the CIs in (7) and (12). For μ the correction is (Rice, 2007)

$$c_\mu = \frac{N-n}{N-1} , \tag{13}$$

and for σ^2 the correction takes the more complex form (Cho and Cho, 2008)

$$c_{\sigma^2} = \frac{N(N-n)(N^2n - 3Nn + 3n - 3N + 3)}{n(N-3)(N-2)(N-1)^2} . \tag{14}$$

Therefore the scalings $\sigma_\mu \to c_\mu\sigma_\mu$ and $\sigma_{s^2} \to c_{s^2}\sigma_{s^2}$ are used.

Because the normal distribution and χ^2 distribution appearing in (2) and (4) are independent, the joint CR for the parameter set $\theta = (\mu, \sigma^2)$ is given by

$$R(\theta) = \left\{ \theta : \mu \in C_\mu, \sigma^2 \in C_{\sigma^2} \right\} , \tag{15}$$

with an associated joint confidence level of $\gamma = 1 - (1-\alpha_1)(1-\alpha_2)$. The area of such a region is a constant multiple of s^3, which is itself only dependent on the population size N, the sample size n and the confidence level γ. This area is large for small samples due to the difficulty in estimating σ^2 in these cases. In general, it is preferable to optimally proportion the confidence level α_2 between the two tails of the χ^2 distribution of the sample variance SE. This means dividing α_2 as to minimise the CR area. For large sample sizes, as would be the case for feasible PFP values, this means choosing an equal division $\alpha_2 = \alpha_3$ as described above and also allowing $\alpha_1 = \alpha_2$. This means that all confidence parameters are determined by the joint confidence level — $\alpha_1 = 1 - \sqrt{1-\gamma}$.

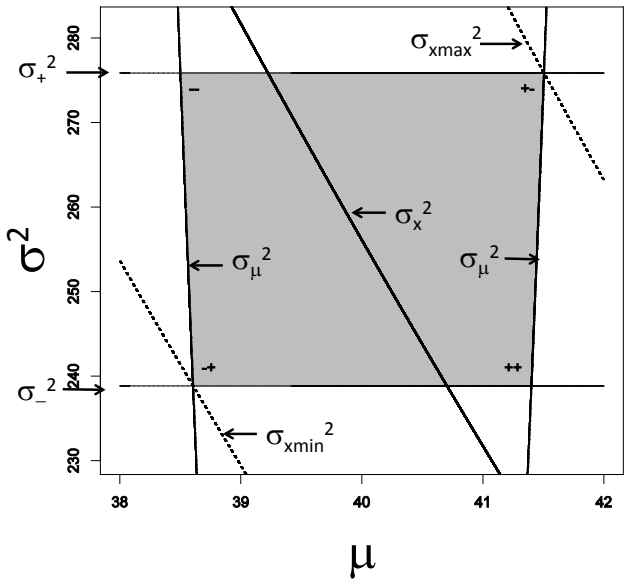

Figure 12: The Confidence Region in (μ, σ^2) space for $N = 200$, $f = 0.7$, $\bar{x} = 40$, $s = 16$, $\gamma = 0.1$, $p = 0.9$. The region is bounded by the three curves $\sigma^2 = \sigma^2_+$, $\sigma^2 = \sigma^2_-$ and σ^2_μ. The intersection of σ^2_x with the μ-axis gives the pth quantile of the sample distribution and the intersections of $\sigma^2_{x_{\text{min}}}$ and $\sigma^2_{x_{\text{max}}}$ with the μ axis give x_{min} and x_{max}, respectively, which are the limits of the CB for the pth quantile. The four vertex labels are explained in the text.

Appendix B: The Confidence Band

For a given total entry, PFP, sample mean and sample SD, the CB gives the region in which the full cohort CDF is likely to lie with confidence $100(1-\gamma)\%$. It is constructed as in (Cheng and Isles, 1983) by considering how the quantiles of the distribution change as the parameter set θ varies in the confidence region $R(\theta)$. As for any location-scale parameter model, the pth quantile is found from the equation

$$\frac{x_p - \mu}{\sigma} = \Psi^{-1}(p) \equiv q\,, \tag{16}$$

where $\Psi^{-1}(p)$ is the inverse CDF for the standard normal distribution. This gives

$$\sigma^2 = \left(\frac{x_p - \mu}{q}\right)^2\,. \tag{17}$$

The values of $\theta = (\mu, \sigma^2)$ which give a constant x_p thus lie on a quadratic curve with intercept with the μ-axis at $\mu = x_p$. The pth quantile of the sample distribtuion CDF $F(x)$ is given by

$$x_p = \bar{x} + qs\,. \tag{18}$$

For a fixed p, the confidence band that encloses the sample distribution CDF is obtained by considering how x_p varies as θ is constrained to be within $R(\theta)$. This depends on the gradient of the curved part of the boundary of the CR which is

$$g(\mu) = \frac{d}{d\mu}\sigma^2(\mu) = \frac{2n}{z^2 c^2_\mu}(\mu - \bar{x})\,. \tag{19}$$

The coordinates of the four 'corners' of the CR as shown in Fig. 12 can be found as

$$(\mu_{--}, \sigma^2_-) \quad (\mu_{+-}, \sigma^2_-) \quad (\mu_{-+}, \sigma^2_+) \quad (\mu_{++}, \sigma^2_+)\,, \tag{20}$$

where

$$\mu_{\pi_1,\pi_2} = \bar{x} + \pi_1 \frac{z}{\sqrt{n}} \sigma \pi_2 \tag{21}$$

gives the four μ-axis values for $\pi_1, \pi_2 = \pm$.

From Eq. (16) this yields

$$x_{\text{min}} = \begin{cases} \mu_{--} + q\sigma_-, & \text{if } -\infty < -1/q < g(\mu_{--}) \\ g^{-1}(-1/q) + q\sigma(g^{-1}(-1/q)), & \text{if } g(\mu_{--}) < -1/q < g(\mu_{-+}) \\ \mu_{-+} + q\sigma_+, & \text{if } g(\mu_{-+}) < -1/q < 0 \\ \mu_{--} + q\sigma_-, & \text{if } 0 < -1/q < \infty \end{cases} \tag{22}$$

$$x_{\text{max}} = \begin{cases} \mu_{+-} + q\sigma_-, & \text{if } g(\mu_{+-}) < -1/q < \infty \\ g^{-1}(-1/q) + q\sigma(g^{-1}(-1/q)), & \text{if } g(\mu_{++}) < -1/q < g(\mu_{+-}) \\ \mu_{++} + q\sigma_+, & \text{if } 0 < -1/q < g(\mu_{++}) \\ \mu_{+-} + q\sigma_-, & \text{if } -\infty < -1/q < 0 \end{cases} \tag{23}$$

The two bounding curves F_{min} and F_{max} of the CB are then constructed by computing x_{min} and x_{max} for $p \in (0, 1)$, noting that the extremes of this interval are excluded as they cause problems numerically. The CR, as computed in a similar way to the procedure set out in a previous study (Arnold and Shavelle, 1998), has the shape of a horizontal slice of a parabola. The three bounding curves, as determined from the CIs (7) and (10), including the FPCs, are given by the equations

$$\sigma^2(\mu) = \frac{n(\mu - \bar{x})}{z^2 c_\mu^2} \tag{24}$$

$$\sigma_\pm^2 = \frac{s^4}{s^2 + z\sigma_{s^2} c_{\sigma^2}} \,. \tag{25}$$

Appendix C: The Post-Awarding Drift

From the CB located around a sample CDF, a plot of the maximum possible deviation (up to the confidence level) from a mark that would be chosen can be produced for a range of PFP values. This is used as a measure of PAD. At a particular value p in the percentile range, PAD $\Delta_p(f)$ is defined as the difference between the corresponding quantiles as calculated from the sample CDF and the CDF of one of the boundaries of the CB. The essentially symmetric nature of the CB means we can choose either. A choice of F_{min}, assuming that $p_{\text{thresh}} \ll 1/2$, results in

$$\Delta_p(f) = x_p - x_{\text{min}} = \Psi^{-1}(p)\sigma + \sigma_+(f, N) \left(\frac{z(\alpha_1) c_\mu(f, N)}{\sqrt{fN}} - \Psi^{-1}(p) \right) . \tag{26}$$

This relationship is derived under the usual assumption that the sample variances s^2 estimates the population variance σ^2 (with finite population correction). This is in fact the case as $f \to 1$ and $p \to 1/2$ where the SE of the sample variance is $\sigma_{\sigma^2} = 0$. This less valid for small sample sizes ($n < 50$) for which the original assumptions of the CR would be broken anyway. In this scenario, the PAD would be driven mainly by the PFP f and finite correction c_μ:

$$\Delta_p(f) \sim \frac{z(\gamma) c_\mu(f, N)\sigma}{\sqrt{f}} \,. \tag{27}$$

Therefore, Eq. (26) can be used to calculate the key value of interest — f_{min}. This is the PFP required in order that with $100(1-\gamma)\%$ confidence the pth quantile mark on the FP CDF will not deviate by more than half a mark from the same quantile mark on the sample CDF. This value is found as the intersection of Eq. (26) with the line $\Delta_p(f) = 1/2$.