Using pupils' grade obtained in national examinations as an outcome measure in evaluations: Some considerations for the design of randomised controlled trials

Ben Smith,

AlphaPlus Consultancy, Manchester, UK

Stephen P. Morris, (ORCID: 0000-0001-6869-8933),

Department of Sociology, Manchester Metropolitan University, Manchester UK

Harry Armitage,

Department of Sociology, Manchester Metropolitan University, Manchester UK

Corresponding author: Ben Smith, Senior Statistician, AlphaPlus Consultancy, Unit 109 Albert Mill, 50 Ellesmere Street, Castlefield, Manchester. M15 4JY Email: ben.smith@alphaplus.co.uk

Ben Smith is Senior Statistician at AlphaPlus Consultancy, Manchester, UK

Stephen Morris is Professor of Evaluation, at the Policy Evaluation and Research Unit, Department of Sociology, Manchester Metropolitan University, UK

Harry Armitage is Research Assistant, at the Policy Evaluation and Research Unit, Department of Sociology, Manchester Metropolitan University, UK

Using pupils' grade obtained in national examinations as an outcome measure in evaluations: Some considerations for the design of randomised controlled trials

Abstract: It is not uncommon for randomised trials in education to have the performance of sample members in national examinations as their primary outcome. In many cases examination results are available as summary measures only. Taking the example of GCSE examination results in England, this paper shows that using summary measures of an underlying score or mark, such as exam grade, complicates the design of trials and can lead to under-powered studies. Simple simulations are used to explore the consequences of powering trials to detect a difference assuming grade or summary measures are the only outcome metric available, where the effects of an intervention is primarily captured in the unknown mark or score distribution. The analysis draws on data that relate the entire distribution of marks in English language and mathematics examinations to grades. Recommendations are made in order to address this problem.

Keywords: randomized controlled trials, outcome measurement, statistical power, sample size, effect size, examination results

Introduction

This paper is concerned with the design of randomised experiments in education. In the not so distant past, a paper on this topic might have been considered niche or fringe, too remote from day-to-day research practice to be of interest to a general audience of education researchers. This is no longer the case. Over the past ten to 15 years the research landscape in education has changed. This transformation has taken place in England as well as arguably elsewhere in Europe (Pontoppidan et al., 2018), and has been underway in North America for some time (Hedges & Schauer, 2018). Randomised controlled trials (RCTs) have moved from the margins to become part of mainstream research practice.

Until the advent of the Education Endowment Foundation (EEF), very few empirical studies in England were RCTs. A fact bemoaned by a number of researchers (Davies, 1999; Oakley, 2006; Torgerson & Torgerson, 2001, 2007). The EEF was established in 2011 with an endowment of £125million from the English Department for Education (Edovald & Nevill, 2020). Its remit was to address under-achievement among disadvantaged children and in particular, close the gap in attainment between the disadvantaged and those from better-off backgrounds (Dawson et al., 2018; Edovald & Nevill, 2020). Taking its inspiration from evidence-based medicine, pioneering work in international development and education research in North America, the EEF has from its inception maintained a strong commitment to RCTs (Dawson et al., 2018; Edovald & Nevill, 2020; Norwich & Koutsouris, 2019). At times this commitment has provoked the ire of the educational research establishment, who continued to pronounce on the perceived shortcomings or inadmissibility of evidence from randomised designs (Biesta, 2007; Thomas, 2020). Debates and controversies have rumbled on, as all the while the number of RCTs in education has risen rapidly. At the time of conducting this research,

we counted 108 RCTs with published results (since 2011) on the EEF website, and a further 46 studies that had published protocols but had not yet reached the stage where their findings had been reported[1] [2]. This situation mirrors that in the United States, where the activities of the Institute of Education Sciences and the National Centre for Education Evaluation and Regional Assistance have had a similar impact since their work began in 2002 (Herrington & Maynard, 2019).

We believe evidence from randomised studies has an important role to play in improving teaching practice and the quality of instruction in schools. But, at the same time, randomisation is not a silver bullet. Alone, RCTs cannot fully meet the aspiration of a more evidence-based approach to education. Other forms of evidence are necessary. Furthermore, a study with a randomised design is not invariably a good study. Randomisation is not a guarantee of quality, nor are RCTs destined to provide actionable evidence for policy and practice.

One of the pervasive difficulties encountered in the design of field experiments is education is the frequency with which studies are under-powered statistically (Cheung & Slavin, 2016; Dawson et al., 2018; Lortie-Forgues & Inglis, 2019; Spybrook & Raudenbush, 2009; Torgerson et al., 2005). Trial samples are often insufficient to

---

[1] This is the total number of studies involving the randomisation of pupils, or classes or whole schools to intervention and control conditions and that had reported findings prior to or by Friday 29th May, 2020.

[2] A full list of studies identified can be found at: https://mmuperu.co.uk/blog/education/the-effects-of-using-examination-grade-as-a-primary-outcome-in-education-trials-to-evaluate-school-based-interventions/

have a reasonable chance of detecting a relevant, or meaningful effect size, given standard thresholds for statistical significance. This leads to trial findings that are frequently inconclusive. A potential contributory factor to the underpowering of trials comes from the choice of primary outcome. Previous research has pointed to the selection of outcomes that are too distal to the intervention or treatment, leading to an effective loss of power (Allen-Platt et al., 2021; Jacob et al., 2019). In this paper, we focus on the nature of a particular outcome indicator frequently used in education trials – summary examination grade – that can act as a contributory factor to the under-powering of studies, with particular reference to the situation in England. Our findings, however, generalise to any graded qualification used as an outcome measure in RCTs. As we will show, the use of a grade, which is a summary or coarse measure of an underlying score (in the case of grade a summary of the underlying mark achieved by a student), is just one example of a practice of using summary measures that may be widespread in education research, with attendant limitations to those we discuss here in relation to the use of grades.

At the outset, it is worth noting that of the 154 randomised studies commissioned by the EEF since 2011, 40 have used examination results as a primary outcome and 27 of these specifically summary grade awarded at GCSE (General Certificate of Secondary Education; national examinations taken at age 16, and the end of Key Stage 4 in the English education system).

What is an examination grade and why choose it as primary outcome?

The focus of this paper is the use of national examination results as an outcome in randomised trials; specifically, the use of the summary grade achieved by pupils as opposed to raw marks or scores. In this paper we consider the example of the grade pupils achieve in their Key Stage 4 GCSE examinations in English language and

mathematics..

A significant number of randomised trials in England, and elsewhere, test interventions targeted at pupils in the years running up to formal or national examinations, with the aim that such interventions will raise attainment in these examinations. In such circumstances, a typical RCT involves the creation, at random, of intervention and control groups (of either schools, classes or individual pupils); the exposure to the intervention for units assigned to the intervention group, whilst those in the control group remain unexposed and are usually subject to business as usual conditions; and then at some point subsequent to exposure, the trial sample typically sit their national examinations. In the English context, the grade achieved by all pupils sitting these examinations, whether the pupil is in the trial sample or not, is recorded in the English national pupil database (NPD), an administrative database which the UK government's Department for Education maintains. Grades achieved by pupils in the trial sample can be extracted from the NPD and converted into outcome indicators, with analysis proceeding on the basis that the grade obtained is a relevant measure of performance in both intervention and control arms of the study. The average performance of pupils in each arm of the study can be compared, with inferences made concerning the causal effects of the intervention on the grade obtained.

Using national examination results obtained from administrative data systems such as the English NPD is an attractive option. Firstly, in England and elsewhere, considerable resources are devoted to the design of examination questions and the maintenance of standards. In England, examination results cannot be linked with a particular level of mastery of the construct assessed (i.e. it is difficult to define what level of performance a grade 3 in mathematics reflects), but they do measure relative

performance consistently over time in a way that is publicly recognised, transparent and open to scrutiny (i.e. a grade 3 one year is comparable to a grade 3 the next).

Second, those administering and scoring instruments that measure attainment outcomes of study participants should ideally be blind to a participant's intervention/control group status (Boutron et al., 2017; Torgerson et al., 2005). Where attainment outcomes are obtained through national examination systems, it is almost invariably the case that those marking examination scripts are indeed "blind" to a pupil's participation in a particular study, and this would no doubt be the case elsewhere as well as in England.

Third, the derivation of outcome measures from national examinations avoids the problem of "treatment inherency". A treatment inherent measure is one derived from a test of constructs not ordinarily taught and "inherent" to the treatment or intervention (Slavin & Madden, 2011). Measures derived by researchers in the context of a specific study tend to err toward being treatment inherent whereas standardised assessment measures are in most cases treatment independent. Due to the purposes for which national examinations are developed, performance as measured in examinations is guaranteed to be general and independent of treatment[3]. Moreover, studies that rely on researcher derived and treatment inherent measures are also associated with the reporting of inflated effect sizes (Cheung & Slavin, 2016; Slavin & Madden, 2011). Thus outcomes derived from national examinations possess the advantage of being

[3] This is in the sense that examinations are not designed with knowledge of the treatment or intervention under-investigation in mind. It is entirely possible that interventions are chosen with performance in examinations directly in mind.

fundamentally treatment-independent in nature and as such can contribute to the avoidance of inflated effect sizes.

One recurring challenge faced in randomised studies is sample non-response. Of particular concern are sample attrition processes (loss to follow-up) that differ in the intervention and control arms of studies. Statistical estimates derived from data affected by differential non-response suffer from a heightened risk of bias. Early EEF-funded RCTs saw high levels of non-response and considerable problems with missing data (Dawson et al., 2018). If a study is reliant on either researcher-developed or standardised instruments administered to study samples through primary data collection, there is a heightened chance that not all sample members will complete an assessment or provide a response across every survey/questionnaire/assessment item from which an outcome measure is derived. If the rate at which this occurs differs in intervention and control arms then analyses based upon such data can suffer from bias. Particularly in the case of English and mathematics at GCSE, almost all pupils across England sit these examinations. This means that outcomes derived from examination results are available for more or less all pupils. As a result, reliance on outcome measures derived from national examinations provides a close-to-universal measure of attainment and by their nature, a much reduced threat from missing data.

A fifth advantage of deriving outcome measures from national examinations is that the costs of administering examination papers, scoring or marking, processing and derivation of results are not borne by the evaluation. In many trials the costs of primary data collection are a substantial fraction of overall budgets. This is a particular concern where there is downward pressure on budgets and a commitment to contain costs but simultaneously concern that studies are systematically under-powered and that study samples are too small (Lortie-Forgues & Inglis, 2019, 2020). The use of results from

national examinations avoids many of these budgetary consequences and their use can either help contain costs or permit larger trial samples for a fixed budget.

Finally, deriving measures of attainment from national examinations has one further benefit; that of policy relevance. The EEF alongside many other education policy-bodies has a commitment to closing the "attainment gap" (Dawson et al., 2018; Edovald & Nevill, 2020). In the English context, and in practice, this is taken to mean closing the gap in performance in national examinations between pupils that qualify for free-school meals and those that do not, where qualification for free school meals is understood as a proxy measure for general disadvantage. Moreover, it is understood that attainment in national examinations is an important determinant of advancement for pupils, playing a central role in facilitating access to further education, training opportunities as well as entry to higher quality jobs and apprenticeships. For these reasons, RCTs that have performance in national examinations as a primary or secondary outcome enable the consequences of interventions to be assessed for their likely contribution to important policy goals.

What is the problem with using examination grades as an outcome?

Examination grades are among the more accessible performance outcome measures available to researchers in England and elsewhere. Details of students' GCSE grades can be obtained through accessing the English NPD under controlled conditions through the UK Office for National Statistics Secure Research Service (ONS SRS)[4]. Trial sample data can be imported into the ONS SRS and linked to NPD student-level

[4] Whilst we focus on GCSE grades as the main example in this paper, performance in other national examinations such as KS2 SATs and A-levels is also available via the NPD.

records. Given this and the wider attractions of using national examination discussed above, what then is the problem with using grades as measure of performance?

Broadly there are two separate problems with using grades as an outcome variable. The first relates to the fact that by design such outcomes measure a particular construct – reading, literacy, or numeracy and so forth. Some interventions target on a specific aspect of one of these domains; conversely, others are much broader interventions (for example, interventions focused on mentoring programmes) which only indirectly act on the examined construct. In such cases, it is likely that a very sizeable impact from the intervention would be required in order to be able to detect an effect when examination grades are the outcome variable. Whilst many commercial standardised instruments mirror that of national assessments and thus are subject to this same limitation, there are also instruments which focus specifically on one aspect of a wider domain and others which directly measure constructs broader interventions may target, and may therefore be more appropriate outcome measures.

Whilst the suitability of examination grades as an outcome measure is an issue for many interventions, the second problem is the focus of this paper. It relates to the fact that in the English context, as elsewhere, the grade a student achieves is a summary measure of an underlying score or mark. The effect of an intervention as captured through performance in an examination will be reflected initially in the mark a student is awarded; we term this the primary metric. The effect of an intervention in terms of grades is only seen if there is a difference in the average mark awarded to pupils in the intervention arm of a study, relative to the average mark received by pupils in the control arm. But, changes in the underlying marks awarded to an individual that reflect the effects of an intervention will not necessarily be reflected in the grade awarded. To see this consider Figure 1.

[figure 1 near here]

In the English system, pupils are awarded a grade based on the mark they have obtained in examinations for a particular award, for example in English language, or mathematics. In Figure 1, the percentage of candidates achieving a given mark in one awarding body's English language GCSE in 2019 is shown. The number of candidates achieving each mark in English language and mathematics GCSE in 2019 was provided for this research by the Joint Council for Qualifications (JCQ), the inter-awarding body group for the largest such bodies in the UK. Where these marks sit in relation to the grade boundaries at GCSE is also indicated, by over-laying the point in the mark distribution where students transition from being awarded a given grade to being awarded either the next grade up or down the distribution[5]. What the Figure shows is that the width between grade boundaries measured in marks varies. Also, that if subsequent to being exposed to an intervention a student improves their mark relative to what they would have achieved had they remained unexposed, this improvement does not necessarily lead to an improvement in the grade obtained. Whether an improvement is seen in a student's grade will depend on: 1) where in the mark distribution the student would have sat had they remained unexposed; 2) the size of the effect of the intervention on the mark obtained for the student concerned; and 3) the width of the grade interval in which the mark they would have received, had they remained unexposed, was located. For example, imagine a student's mark in the absence of exposure to an intervention placed them just above the boundary between grade 2 and 3 - that is with an award of grade 3 - but at the lower end of the mark distribution for this

[5] GCSE grades transitioned from an A*-G to a 9-1 model in a phased manner, with the first 9-1 subjects awarded in Summer 2017. Grade 9 is the highest available grade in the new model.

grade. It is apparent that an improvement in the student's performance resulting from exposure would have to lead to a substantial change in the mark awarded for their grade to improve it from a grade 3 to grade 4. As a result of these considerations, grades should be seen as a coarse measure of performance and this coarseness has implications for trials that only have access to grade data and plan their sample sizes on this basis.

To see this, imagine the situation in which a researcher is planning a simple RCT in which individual students are allocated at random to intervention and control groups on a 1:1 basis. Pupils in the intervention group are exposed to an intervention that has as its goal the improvement of examination performance. Imagine, the researcher selects a sample sufficient in size such that a difference in performance might be detectable at the 95 per cent level of statistical significance, with 80 per cent power, equivalent to 0.25 of a standardised mean difference (the effect size or minimum detectable effect size). The researcher only has access to examination results in the form of the grade awarded to students and uses this as the trial's primary outcome measure. The researcher might use an equation similar to equation [1] below to calculate the size of sample required under these conditions (Dong & Maynard, 2013):

$$n = \left(\frac{M_{n-k-2}}{\Delta}\right)^2 \left(\frac{1}{P(1-P)}\right) \text{.....[1]}$$

Here n is the total number of students to be allocated to intervention and control groups; Δ the chosen effect size to which the trial is powered (in this case $\Delta = 0.25$); and M a multiplier derived from values drawn from the t-distribution consistent with the chosen level of statistical significance and power, with n-2 degrees of freedom (in this

case M_{n-k-2} ~ = 2.50 for samples where n >= 40)[6]. Under the assumption of 1:1 allocation to intervention and control groups the quantity in the first set of brackets is simply multiplied by 4. So in this hypothetical case the total required sample is roughly 400 pupils.

These calculations show that to detect a difference between intervention and control groups in the grade awarded at GCSE for the trial sample, equivalent to a standardised mean difference of 0.25, the researcher would need to recruit around 400 students in total to their study. From published statistics we can see that the standard deviation for grades in English language at GCSE, is typically around 1.9 grades. This means that a trial of 400 students powered to detect an effect size of 0.25, would be able to detect an average improvement in the grade among those exposed to the intervention of around 0.47 of a grade. This is based on the assumption that grades are analysed as a continuous response.

Bearing all this in mind, it might seem entirely plausible to the researcher – based on their prior knowledge, theory and previous research – that the intervention under consideration might lead to pupils in the intervention group scoring 0.47 of one grade higher on average than those in the control group. The crucial issue though is whether the change in marks between those same pupils, that would be necessary to yield an effect size of 0.25 in standard deviations and 0.47 in grades, would also be plausible when viewed from the perspective of existing evidence and theory? Given the data we have on the full distribution of marks for GCSE examinations taken in 2018/19 (discussed in detail below) we can see that the standard deviation of the mark

---

[6] Here we assume a one-tailed test for statistical significance, 95 per cent statistical significance and 80 per cent power.

distribution is typically about 23.3 marks. If we convert an effect size of 0.25 into marks we obtain an effect in marks consistent with this effect size of around 5.83 marks. The problem is that the grade/mark relationship is non-linear. The average width between grade boundaries expressed in marks for GCSE English language is about 14 marks, but it varies. Thus for a number of students an effect size of 0.25, equivalent to a 0.47 improvement in the grade, would not change their grade because the equivalent change in marks (5.83 marks) is quite a bit smaller than the average width of the grade boundaries expressed in marks. This non-linearity or jumpiness of the response measured in grades to changes in underlying marks is at the root of the problem. In effect there is a loss of power to detect an effect measured in grades relative to marks. The situation is shown more clearly in the Figure 2 below.

[figure 2 near here]

Figure 2 shows that for many points on the mark distribution an improvement of 5.83 marks would not translate into a change in the grade awarded. Further, that students already achieving a grade 9 under control conditions cannot see an improvement at all in terms of grades, even if their underlying marks do improve.

Complicating this, the distribution of pupils is not uniform across all marks; Figure 1 shows that the distribution of marks is relatively normal. This means that students are more likely to achieve marks and grades around the peak of this distribution, between grades 4 and 6. In turn, the width of these more frequently achieved grades will thus have a greater impact on how likely students are to see a change in grade as a result of the intervention.

The purpose of the rest of this paper is to illustrate this problem further. Essentially, we answer the question: to what extent does using examination grade as an

outcome measure, in contrast to using marks, lead to a decrease in statistical power, for GCSE English and mathematics?

Materials and methods

Approach

Our approach to illustrating the problem, and addressing the question we have set ourselves, is based on simulating a number of simple RCT designs using data from the actual mark distributions for GCSE English and mathematics sat by pupils at English examination centres in the summer of 2019 (that is prior to the outbreak of Covid-19). We use these simulations to examine whether statistical power is lost or gained when we convert simulated results from marks to grades based on the observed relationship between the two.

We produce 5,000 simulated trials for five different trial designs. These different designs are identical but for the fact that each of the five designs is powered to detect a different effect size[7]. What this means is that each design is associated with a different sample size. Table 1 below illustrates.

[table 1 near here]

The table shows that in the case of design 1, powered to detect an effect size of 0.25 of one standard deviation, the researcher will need to recruit and randomise roughly 400 subjects to intervention and control conditions. We run 5,000 simulations of such a trial. As we have complete information (discussed further below) on the full distribution of marks for GCSE examinations in English and mathematics in 2019, we can convert each of these simulated results into marks. Thus we can show how

[7] Assuming null hypothesis significance test at the 95 per cent level and with 80 per cent power.

estimates of the treatment effect in marks from each simulated trial are distributed around the central effect size. The sample size is derived such that were an effect size of 0.25 to be true, in 80 per cent of trials with a sample size of 400 subjects, we would reject the null hypothesis. That is, such a statistical test has power of 80 per cent and its inverse, the Type II error rate, is 20 per cent.

Given that we also know how the underlying mark distribution relates to the grade awarded for English and mathematics GCSE in 2019, we can convert the distribution of effects measures in marks consistent with the effect size = 0.25 into a grade distribution across all 5,000 simulations for each design. We can then calculate the Type II error rate for this distribution of grades equivalent to the distribution of effects in marks we have generated. We can then see if the Type II error rate differs for the distribution of grades compared to that for marks, where we know that by definition for the former the benchmark Type II error rate is 20 per cent. If the resulting Type II error rate for the grade distribution is higher than that for marks, statistical power is lost, if it is lower, statistical power is gained.

In practical terms, what this shows is that a trial powered to detect an effect size of 0.25 but where the researchers have access to examination results in grades only, may not have the power the researcher supposes it to have. Given that an intervention must have an effect on the marks awarded for there to be a consequence in terms of the grade awarded, it is statistical power associated with the mark distribution that counts when determining sample size. If power for a given effect size is different in the mark and grade distributions then judging power in grades can lead to experiments with less than optimal sample sizes or put differently with actual Type II error rates that differ to the nominal rate used in the sample size calculation.

Data

The results presented in this paper are for both GCSE English and mathematics. These two examinations were chosen chiefly because results from them are widely used as outcomes and they are examinations taken by nearly all English school children at age 16. We obtained details of the distribution of marks received by all pupils sitting English language and mathematics GCSEs in summer of 2019. Data comes from each examination awarding organisation (AO) in England[8] and were provided by the Joint Council for Qualifications (JCQ). Results obtained were for students aged 15 or 16 years in summer 2019 only, thereby excluding early-entry, re-sitting and mature candidates. These data were supplemented with publicly available details published by each AO of their respective grade boundaries in relation to the underlying distribution of marks for each subject.

Each AO delivers a GCSE examination in English language and mathematics that will have a different grade/mark distribution, with different though similar grade boundaries. The analysis we report here was carried out separately for each AO and each qualification. Results for each AO and an overall weighted average are reported. Adjustments are also made due to the existence of higher and foundation tier GCSEs in mathematics, which vary in their difficulty and permit pupils to achieve a different subset of the 9-1 grades.

---

[8] In England a 'free market' approach to examinations exists, with multiple awarding organisations which offer different versions of each GCSE. Schools are free to choose which AO's GCSE they administer for each subject, and can mix and match between AOs.

Results

As mentioned previously, the distribution of simulated sample effects around the associated central estimate, recorded as a standardised mean difference, for each trial design can be converted into a distribution on the mark scale. This is achieved by multiplying the effect size by the standard deviation in marks for each AO, and within each AO, for each subject. Therefore for each of the 5,000 simulations, for each study design, we have the simulated result in marks. The data we have which combines the mark distribution with grade boundary information, permits us to locate each mark achieved within a given grade boundary (i.e. the calculation is not based on standard equations but done mechanically), meaning we can take our simulated distributions in marks and convert these to grades. On this basis we have a full distribution of simulated effects in grades, that would reflect the grade differences that would be observed if a hypothetical intervention had produced the effects in marks we have simulated for study designs 1-5.

We know by design that null hypothesis significance tests across simulated trials for a given design would result in the null hypothesis being rejected in 80 per cent of instances, and the observed effect declared as statistically significant. The question is, what would be the rejection rate for repeated tests on the distribution with exactly the same design but where we did not have access to marks but instead had to rely on measuring attainment in grades? If the rejection rate is greater than 80 per cent all else being equal, power is gained, if it is less, power is lost. The results of our analysis are presented in Tables 2 and 3.

Table 2 presents the proportion of tests of statistical significance at the 95 per cent level that would result in rejection of the null hypothesis, across 5,000 simulated trials conducted for each AO, and each design, after the distribution of effects is

converted from marks to grades using the grade/mark distributions we have obtained and published grade boundaries. To repeat, we know that the benchmark rejection rate is 80 per cent. For a design powered to detect an effect size (standardised difference in means) of 0.05 for AO 1, the actual reject rate is 0.75 rather than 0.80, representing an increase in the Type II error rate from 20 to 25 per cent. Thus there is an increase of five percentage points in the probability that we would find in favour of the null hypothesis, even though the alternative hypothesis ES=0.05 is true by definition, in repeated trials of the same design and sample size. Further, this means that a sample size chosen on this basis would be sub-optimal.

[table 2 near here]

Overall, the analysis in Table 2 shows that for GCSE English language, trials powered to detect an effect size where the outcome is measured in grades are likely to be underpowered relative to the situation where marks are the outcome. All rejection rates are lower than 80 per cent, though admittedly, in the case of English language, for a minority of award body/design combinations, the problem is either non-existent or trivial.

[table 3 near here]

Table 3 contains the results for mathematics. Here the position is slightly starker, with in general, a greater loss of power. For most awarding body/design combinations, the loss of power is at least five percentage points when the simulated effects are converted from marks to grades, and in some cases the loss in power is quite a bit more.

Discussion

What these analyses show is that not all of the change in marks achieved by pupils in an intervention group as a result of their exposure to a hypothetical intervention will be

reflected in a change in the grade they are awarded. Some gains in marks do not change the grade a pupil achieves. When the change in grade that has been mechanically converted from marks is further converted to an effect size, we can, holding the sample design constant in all other respects, calculate the power of the design on the grade scale. In most cases, statistical power is lower and it is appreciably so for mathematics.

In practical terms what are the consequences of these findings? Most studies in England for which performance at GCSE is declared the primary outcome will have access to the grade attained by sample members, rather than marks. What this analysis shows is that a trial powered to detect an effect size where grades are the fundamental unit in which the outcome is measured can be under-powered relative to the sample size required to detect an equivalent effect in marks. When the required effect in marks is understood it is likely that the equivalent grade change is quite a bit lower than anticipated. This is so because the relationship between marks and grades for any given effect size to which a trial might be powered is not linear or smooth (see Figure 2). In fact, it is a quite complex relationship that even Figure 2 does not fully capture. Further, the complexity of this relationship will vary across subjects depending on the structure of the award.

The complexity in the relationship is further illustrated through the different results we have obtained for English language and mathematics. The loss is power is more pronounced for mathematics. This result is obtained because the standard deviation of marks in mathematics equates to a smaller fraction of a grade than the standard deviation of the grade distribution. However, for English, the standard deviation in marks equates to only a little less than the standard deviation of the grade distribution, indicating that there is a reduced loss in sensitivity entailed when working in grades for this subject.

A factor in why this is the case is that, in England, mathematics is a tiered subject, meaning that students can sit two different tiers of examination paper that vary in their demand. The higher tier paper is tougher. Results from both tiers are however combined and mapped on to the same 9-1 grade scale as other GCSEs. What this means is that within each tier there are fewer grades available relative to the range in possible marks. Put differently, the tiering system works to produce a much wider range of possible marks per grade. This means that the coarsening effect of collapsing marks into grades is significantly greater for mathematics than it is for English. This finding can be generalised to any other qualification – the key consideration is how coarse primary metrics are relative to summary ones.

What these analyses also show is that researchers relying on grade outcomes or summary measures more generally, and planning their trials accordingly, need to think very carefully about whether the effect sizes they are powering their trial to are plausible, not just in terms of summary outcome (in our case grades), but in the primary metric (in our case marks). This should be a key consideration in feasibility/pilot studies assessing whether an intervention should proceed to a full trial.

Two key factors are worthy of further consideration. First, and more generally, even if we care about performance in national examinations, are such tests necessarily sufficiently proximal to the intervention under consideration? The content of interventions and their theories of change, or general theoretical basis, need to be very carefully considered in selecting primary and secondary outcomes. Even if interventions are designed with distal measures in mind, such as performance in national examinations, it may be of greater plausibility to capture intervention effects in more proximal standardised measurement instruments. What is key here is that a convincing and logical account can be given in theoretical terms, as to the location of

any chosen standardised measure, on a causal pathway between the intervention and the more distal outcome - performance in national examinations. Thus if a measurable effect can be detected on a standardised measure it is reasonable to suppose that a contribution will be made to improved examination performance, albeit one not directly measurable in the particular setting and under the conditions of the evaluation.

Second, if researchers conclude that performance in national examination is the appropriate outcome measure, in England they should seek to obtain marks as these are what we have termed the primary metric. Trials should be designed such that they are powered to detect a plausible or meaningful effect size in the primary metric (marks) rather than a summary measure of that metric (the grade). In other international settings, our advice would be for researchers to strive to obtain the primary metric. If summary grade-like measures are all that is available, or all that can be obtained for a reasonable cost then adjustments will need to be made to power and sample size calculations. It is likely that sample size will need to be larger than would otherwise be the case. Fundamentally these adjustments will depend on the nature of the award and the relationship between the primary metric and the summary measure. Here we have shown that the width of the summary measure expressed in units of the primary metric is key in this regard.

Finally, the analysis presented here is based on simulating the very simplest trial design involving the randomisation of individual pupils. Many, if not the majority of trial designs in education are more complex, and frequently take the form of cluster or group randomised designs, in which whole schools or classes are allocated to intervention and control conditions. For the sake of brevity and that attention to be drawn to the crucial issues we have here deliberately focused on the simplest case, that of a trial design in which individual students are randomised to intervention and control

group. To summarise, the magnitude of the loss in power in group and cluster trials is related to the so called design effect. The design effect enables researchers to adjust sample size calculations made on the basis of individual random allocation to take account of the effect of clustering when conducting a group randomised trial (Campbell & Walters, 2014). Space prohibits a more expansive discussion though interested researchers may consult Smith, et al. (2021) a companion piece to this paper which demonstrates our findings hold in the case where more complex designs are used.

Conclusion

Where researchers select national examinations results as a source from which to obtain outcome measures in randomised trials, they are often presented with summary or coarsened measures that are constructed from underlying or primary performance metric. In England, GCSEs examinations sat by pupils at age 16 have as their summary measure examination grade, which is coarsening of a primary metric - examination mark. We have shown that effects in marks translate to effects in grades that are smaller when expressed as a standardised mean difference and that this disparity can complicate sample size calculations. We recommend that researchers think carefully about whether outcome measures obtained from national examinations are sufficiently proximal to treatments and strive to obtain primary metric measures for outcome definition and derivation where at all possible.

Acknowledgements

The authors would like to thank the Joint Council for Qualifications (JCQ) for kindly supplying the mark distribution data necessary to enable this research to take place

Disclosure statement

The authors have derived no personal financial interest nor benefit from the research described in this paper.

Funding details

This work was supported by the Education Endowment Foundation under the grant titled ‘The effect of using grades as an outcome variable in evaluation’.

References

Allen-Platt, C., Gerstner, C.-C., Boruch, R., & Ruby, A. (2021). Toward a Science of Failure Analysis: A Narrative Review. *Review of Research in Education*, *45*(1), 223–252. https://doi.org/10.3102/0091732X20985074

Biesta, G. J. J. (2007). Why 'what works' won't work. Evidence-based practice and the democratic deficit. *Educational Theory*, *57*(1), 1–22.

Boutron, I., DG, A., Moher, D., KF, S., Ravaud, P., & Group, for the C. N. P. T. (2017). Consort statement for randomized trials of nonpharmacologic treatments: A 2017 update and a consort extension for nonpharmacologic trial abstracts. *Annals of Internal Medicine*, *167*(1), 40–47. http://dx.doi.org/10.7326/M17-0046

Campbell, M. J., & Walters, S. J. (2014). *How to design, analyse and report cluster randomised trials in medicine and health related research*. John Wiley & Sons.

Cheung, A. C. K., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, *45*(5), 283–292.

Davies, P. (1999). What is Evidence-based Education? *British Journal of Educational Studies*, *47*(2), 108–121. https://doi.org/10.1111/1467-8527.00106

Dawson, A., Yeomans, E., & Brown, E. R. (2018). Methodological challenges in education RCTs: reflections from England's Education Endowment Foundation. *Educational Research*, *60*(3), 292–310.

Dong, N., & Maynard, R. A. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and

quasi-experimental design studies. *Journal of Research on Educational Effectiveness*, *6*(1), 24–67. https://doi.org/10.1080/19345747.2012.673143

Edovald, T., & Nevill, C. (2020). Working Out What Works: The Case of the Education Endowment Foundation in England. *ECNU Review of Education*, 2096531120913039.

Hedges, L. V, & Schauer, J. (2018). Randomised trials in education in the USA. *Educational Research*, *60*(3), 265–275. https://doi.org/10.1080/00131881.2018.1493350

Herrington, C. D., & Maynard, R. (2019). Editors' Introduction: Randomized Controlled Trials Meet the Real World: The Nature and Consequences of Null Findings. *Educational Researcher*, *48*(9), 577–579. https://doi.org/10.3102/0013189X19891441

Jacob, R. T., Doolittle, F., Kemple, J., & Somers, M.-A. (2019). A Framework for Learning From Null Results. *Educational Researcher*, *48*(9), 580–589. https://doi.org/10.3102/0013189X19891955

Lortie-Forgues, H., & Inglis, M. (2019). Rigorous Large-Scale Educational RCTs Are Often Uninformative: Should We Be Concerned? *Educational Researcher*, *48*(3), 158–166.

Lortie-Forgues, H., & Inglis, M. (2020). On the Practicality of Extremely Large Educational RCTs. *Educational Researcher*, *49*(4), 291–292.

Morrison, K. (2021). *Taming randomized controlled trials in education: Exploring key claims, issues and debates*. Routledge.

Norwich, B., & Koutsouris, G. (2019). Putting RCTs in their place: implications from an RCT of the integrated group reading approach. *International Journal of Research & Method in Education*, 1–14.

Oakley, A. (2006). Resistance to new technologies of evaluation: Education research in the UK as a case study. *Evidence and Policy*, *2*(1), 63–87.

Pontoppidan, M., Keilow, M., Dietrichson, J., Solheim, O. J., Opheim, V., Gustafson, S., & Andersen, S. C. (2018). Randomised controlled trials in Scandinavian educational research. *Educational Research*, *60*(3), 311–335. https://doi.org/10.1080/00131881.2018.1493351

Slavin, R., & Madden, N. A. (2011). Measures inherent to treatments in program effectiveness reviews. *Journal of Research on Educational Effectiveness*, *4*(4), 370–380.

Smith, B., Morris, S. P., & Armitage, H. (2021). *The effects of using examination grade as a primary outcome in education trials to evaluate school-based interventions*. https://educationendowmentfoundation.org.uk/public/files/Publications/The_effects_of_using_examination_grade_as_a_primary_outcome_in_education_trials.pdf

Spybrook, J., & Raudenbush, S. W. (2009). An examination of the precision and technical accuracy of the first wave of group-randomized trials funded by the Institute of Education Sciences. *Educational Evaluation and Policy Analysis*, *31*(3), 298–318.

Thomas, G. (2020). Experiment's persistent failure in education inquiry, and why it keeps failing. *British Educational Research Journal*.

Torgerson, C. J., & Torgerson, D. J. (2001). The need for randomised controlled trials in educational research. *British Journal of Educational Studies*, *49*(3), 316–328.

Torgerson, C. J., & Torgerson, D. J. (2007). The need for Pragmatic Experimentation in Educational Research. *Economics of Innovation and New Technology*, *16*(5), 323–330. https://doi.org/10.1080/10438590600982327

Torgerson, C. J., Torgerson, D. J., Birks, Y. F., & Porthouse, J. (2005). A comparison of randomised controlled trials in health and education. *British Educational Research Journal*, *31*(6), 761–785.

Table 1. Randomised designs, effect sizes and sample size

Design	effect size	sample size
1	0.25	400
2	0.20	620
3	0.15	1,107
4	0.10	2,476
5	0.05	9,894

Source: authors calculations (using PowerUp program available in the statistical software R)

Table 2. Proportion of simulations H_0 rejected in (outcome variable = grades) – English language

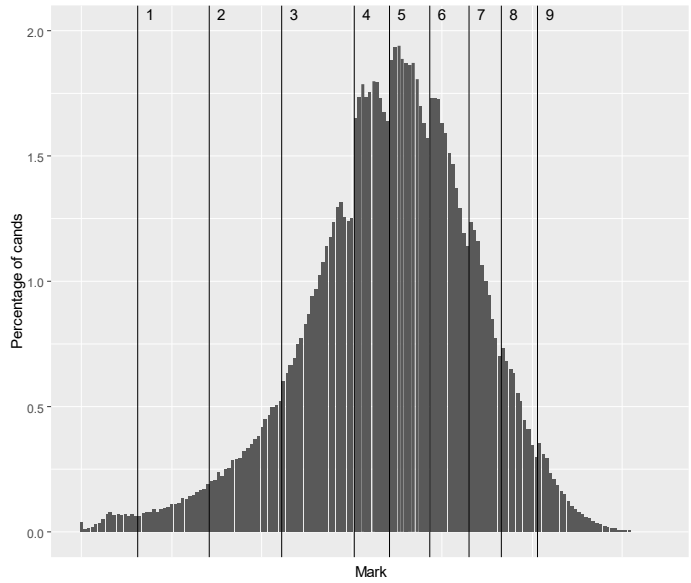
Design	MDES	AO 1	AO 2	AO 3	AO 4	All AOs
1	**0.05**	0.749	0.702	0.730	0.687	0.739
2	**0.1**	0.753	0.702	0.733	0.721	0.746
3	**0.15**	0.777	0.725	0.771	0.738	0.770
4	**0.2**	0.791	0.734	0.776	0.763	0.785
5	**0.25**	0.799	0.752	0.788	0.765	0.793

Source: authors calculations

Table 3. Proportion of simulations H_0 rejected in (outcome variable = grades) – Mathematics

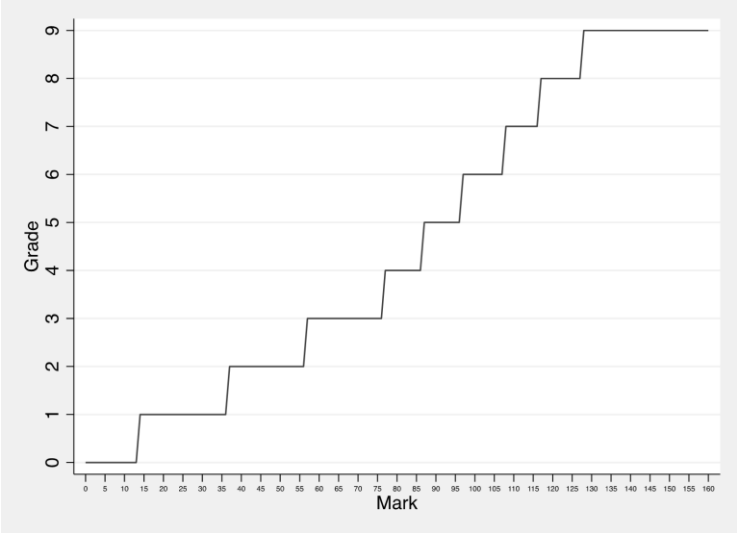
	Foundation tier				Higher tier				All AOs, both tiers
MDES	**AO 1**	**AO 2**	**AO 3**	**All AOs**	**AO 1**	**AO 2**	**AO 3**	**All AOs**	
0.05	0.668	0.661	0.658	0.661	0.675	0.639	0.641	0.650	0.655
0.1	0.696	0.711	0.719	0.713	0.742	0.708	0.705	0.716	0.714
0.15	0.728	0.742	0.736	0.735	0.762	0.735	0.730	0.739	0.737
0.2	0.728	0.748	0.756	0.748	0.758	0.750	0.748	0.751	0.750
0.25	0.741	0.746	0.747	0.745	0.767	0.751	0.755	0.758	0.752

Source: authors calculations



Source: data received from JCQ concerning a particular awarding organisation

Figure 1.



Source: data received from JCQ concerning a particular awarding organisation

Figure 2.

Figure 2. GCSE English Language mark distribution (with grade boundaries 1-9)

Figure 2. Relationship between marks and grades – English language GCSE 2018/19 – chosen awarding organisation