


Please cite the Published Version

Smith, Ben, Morris, Stephen  and Armitage, Harry (2021) The effects of using examination grade as a primary outcome in education trials to evaluate school-based interventions. Research Report. Education Endowment Foundation.

Publisher: Education Endowment Foundation

Version: Published Version

Downloaded from: <https://e-space.mmu.ac.uk/628151/>

Usage rights:  Open Government Licence 3.0

Enquiries:

If you have questions about this document, contact openresearch@mmu.ac.uk. Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from <https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines>)



Education
Endowment
Foundation

The effects of using examination grade as a primary outcome in education trials to evaluate school-based interventions

EEF Research Paper
July 2021

Authors:

Ben Smith, Stephen Morris and Harry Armitage



Summary

This paper aims to assess the impact of using GCSE grades as a primary outcome in educational evaluations and trials, compared to using marks.

The choice of grades or marks is relevant as many evaluations use GCSE performance as an outcome measure. For such evaluations, the National Pupil Database (NPD) is used as a source of data by the vast majority, and it contains grade information only; no information on the underlying mark distribution is available. The use of the NPD has largely been a pragmatic choice, as historically it has been comparatively easy to access grades via the NPD than gather marks (or indeed grades) from schools or awarding organisations. However, there has been little consideration of whether only having access to grade information, as opposed to the marks, has consequences for evaluations. This paper seeks to fill this gap in the literature.

Both grades and marks measure students' performance on national assessments, but grades are a coarser measure of this performance. Crucially, grades are not an independent measure, they are simply a summary of marks. As such, we hypothesised that there would be a loss in sensitivity and statistical power entailed in using grades as an outcome variable rather than marks. In other words, due to grades' coarseness, the intervention may need to have a larger impact (i.e. students in the treatment group would need to improve their performance by a greater number of marks) for an evaluation to be capable of detecting a given effect size. If a fall in sensitivity proves to be the case, then using grades as an outcome variable would necessitate a larger sample size in order to maintain the desired level of statistical power, with commensurate consequences for the cost of trials.

Data was provided by the Joint Council for Qualifications detailing mark distributions for GCSE English Language and Mathematics in the summer 2019 series. A simulation approach was used to model thousands of trials in each subject, for a range of different target MDES values when power is held constant at 0.8 (a value selected on the basis of typical design choices for EEF-funded trials). In each simulated trial, the effect was measured both in marks and in grades (through converting marks to grades), and statistical tests used to determine whether a result reached standard thresholds for statistical significance when using each metric. Across each set of simulations we then compare the proportion of trials in which we reject the null hypothesis where the outcome is measured first in marks, and then after converting the outcome, in grades. By design, in each of our simulations, the rejection rate for outcomes measured in marks will be 80 per cent across all simulated trials (that is we will find-against the null in 80 of 100 simulations and thus reject it). After converting marks to grades, however, the rejection rate may be different. We hypothesise that it will be lower. If it falls below 80 per cent then this is consistent with a loss of power; that is, a loss of sensitivity.

We found that as hypothesized, there was a drop in sensitivity associated with using grades rather than marks as the outcome variable, but that this drop was notably larger for Mathematics:

- In English Language, the decline in statistical power associated with using grades rather than marks was at least 0.9 per cent (over the nominal rate of 80 per cent).
- In Mathematics, the average loss in power is larger, at at least 4.7 per cent.

The reductions in sensitivity which emerged would, in practice, mean that any trial using grades as an outcome variable needs to recruit more students and schools in order to maintain the same level of statistical power. We quantified the increase in sample size associated with the above simulated figures for each subject:

- For English Language, in some cases no increase in sample size was necessary, and at worst a sample size 1.18x larger was needed to maintain 0.8 statistical power when using grades as the outcome variable.
- For Mathematics, a sample size between 2.53x and 3.19x larger was needed to maintain 0.8 statistical power when using grades.

From this, it is apparent that the choice of marks vs grades as the outcome variable has a fairly negligible impact on sensitivity and thus sample size for English Language, but a considerable one for Mathematics.

Ultimately, the reason why there is a greater effect of using grades rather than marks as the outcome variable for Mathematics is attributable to how many grades an standard deviation of marks equates to. This is because

how easy an effect is to detect is heavily impacted by how spread student scores are; if scores are distributed tightly around the mean, then it is much easier to detect an effect amongst the noise of student-level variations in performance.

There are some important implications of these findings; chiefly, for evaluations where GCSE Mathematics is the outcome variable. Where Mathematics at GCSE is declared the study's primary outcome, it seems very beneficial (in terms of sample size required) to use marks rather than grades as an outcome variable. This in turn means that such evaluations should avoid the NPD as a source of outcome data, and instead gather mark information from other sources. However, these findings can reasonably be generalised to any other assessment where number of grades an SD of marks equates to is higher than the SD of the grade distribution.

More widely, we believe this research highlights a potentially overlooked step in trial design; there must be a coherent logic in place as to how the intervention will improve students' GCSE marks (not grades, as marks always underpin grades even if not the outcome variable) – and what the magnitude of this improvement could plausibly be. Our feeling is that in many cases, interventions are likely to only result in a handful of marks gained (on average) and thus a small effect size. The key judgement is then whether the plausible mark gains the intervention could result in are larger than the 'minimally important mark difference' (in terms of effecting a meaningful change in grade), and thus whether the intervention is likely to prove economically viable when costs of implementation are considered.

Introduction

This paper examines the consequences for the planning and conduct of randomised trials in education that stem from using results in national examinations as outcome measures. Many trials in England rely on pupil attainment in GCSE maths or English language as an outcome; obtained typically from extracts of the National Pupil Database (NPD) linked to trial records. Thus average grades achieved by pupils in an intervention or treatment group are compared to those achieved by control group pupils. The resulting difference is interpreted as the causal effect of the treatment. As is well known, trials should be designed with reference to a minimally important difference. That is, a sample design should be chosen that is capable of detecting a difference between treatment and control groups consistent with some minimally important effect that would warrant further interest. In other words, there is no point designing a trial to detect a trivial effect that would not be worthy of further consideration. The focus of this paper, is the implications that arise from choosing grade as an outcome for judgements around minimally important differences in the planning of trials. Essentially the issue is this: treatments do not directly affect the grade obtained; they affect the mark obtained. The grade is a summary measure of the mark. The relationship between the mark and the grade a pupil achieves is surprisingly complex and varies across subjects. If researchers plan their trials on the basis of a minimally important difference defined in grades, given choices around desirable levels of statistical significance and sample power, then it could quite easily be that the planned effect size is simply unrealistic when viewed in terms of marks. That is, a difference defined in grades implies an implausibly large effect in marks. This paper explains why this is and what might be done about it.

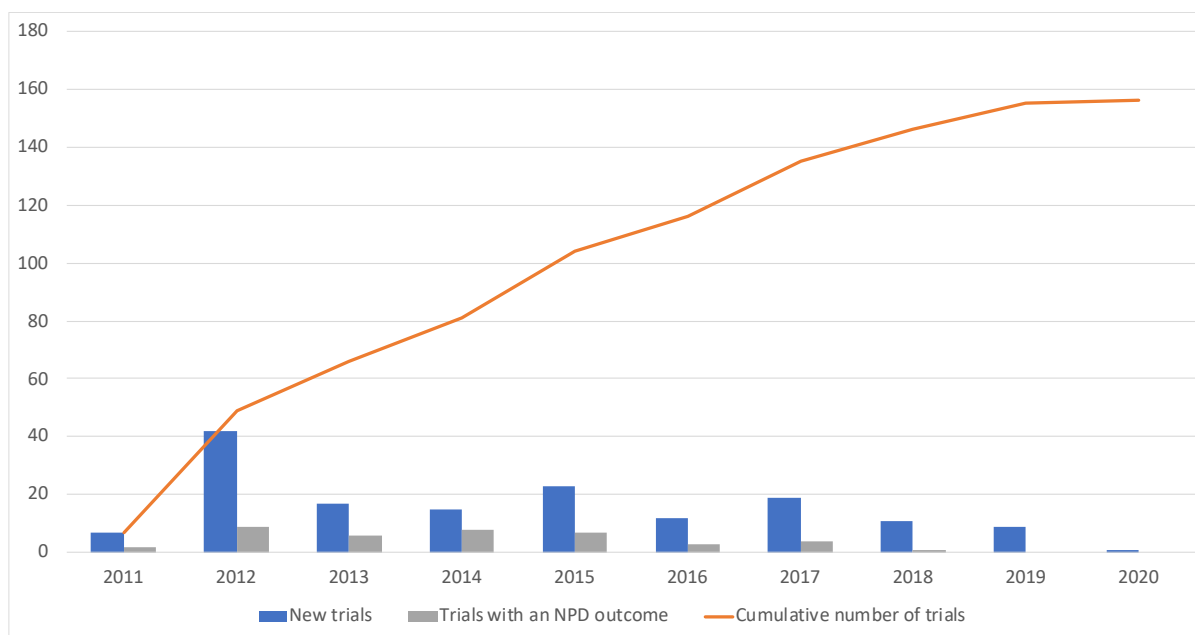
In just under a decade educational research in England has been transformed. Until the advent of the Education Endowment Foundation (EEF), very few randomised trials testing the effects of treatments in education had been conducted, a state of affairs long bemoaned by some authors (Davies, 1999; Oakley, 2006; Torgerson & Torgerson, 2001, 2007). This is despite the fact that education, as a field, had a strong claim to be one of the first in which randomised trials were seen as possible and even desirable (Oakley, 1998).

The EEF was established in 2011 with a £125 million endowment from the Department for Education, but independent of government, with an envisaged lifespan of 15 years (Edoald & Nevill, 2020). Its remit was to address relative under-achievement among children from less advantaged backgrounds through helping close the 'attainment gap' (Dawson et al., 2018; Edoald & Nevill, 2020). It would seek to do this through bringing together and synthesising the best available evidence on the effectiveness of school-based interventions that had as their aim raising attainment, and presenting this evidence in a format accessible to both policy makers and educational practitioners. Furthermore, the EEF would aim to fill gaps in the evidence base by funding new primary studies conducted by independent evaluators (Edoald & Nevill, 2020; Xiao et al., 2016). It is within this context that role of the EEF and its North American counterparts is seen. Specifically to provide the resources

and evidence to education practitioners that will enable them to improve educational standards as measured, ultimately, by examination results (Murnane & Nelson, 2007). It is in terms of examination results that the attainment gap is typically understood and defined.

Taking inspiration from evidence-based medicine, from its inception the EEF has had a strong commitment to the use of randomised controlled trials (RCTs) (Dawson et al., 2018; Edovald & Nevill, 2020; Norwich & Koutsouris, 2019). At the time of writing, we counted 108 studies¹ on the EEF website with published findings that involved the randomisation of pupils, classes or whole schools to intervention or control groups, and a further 46 with published protocols² yet to reach the reporting stage. The cumulative number of EEF-funded trials over time can be seen below, in Figure 1. It also shows the number of these trials with outcome measures derived from National Pupil Database (NPD) data, which amount to just over a quarter of EEF-funded trials in total.

Figure 1: New and cumulative number of RCTs commissioned by the Education Endowment Foundation, including trials with an NPD primary or co-primary outcome



Source: Authors' calculations

Edovald & Nevill (2020) claim that globally, EEF-funded randomised controlled trials account for nearly one in five of all education trials. The situation in England mirrors that in the United States, where the activities of the Institute of Education Sciences and its subsidiary the National Centre for Education Evaluation and Regional Assistance (NCEE) (established somewhat earlier than the EEF in 2002), has also seen a dramatic increase in the number of randomised studies in education (Makel & Plucker, 2014; Murnane & Nelson, 2007; R. E. Slavin, 2017).

There is therefore now a discernible history of EEF activity in England stretching back over nearly ten years. Across this period, a significant number of EEF-funded trials have specified primary outcomes derived from the National Pupil Database, notably outcomes derived from national examination results and standardised assessment tests (SATs). This is in contrast with studies that have primary outcomes focusing on student achievement in standardised (often commercial) tests (such as GL Assessment, the Centre for Evaluation and Monitoring and Hodder Education, etc.). These tests almost always require extensive fieldwork in schools leading to spiralling costs and problems with missing data. Reliance on national examinations and standardised

¹ These were trials the EEF designated either efficacy or effectiveness studies as at 29th May, 2020

² This is as of Friday 29th May, 2020. A full list of these trials 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/>

assessments was to some extent adopted as a strategy to avoid these and other problems associated with collecting standardised assessment results through primary data collection.

EEF's interest in the use of examination results as outcome measures was initially as a source for longitudinal outcomes for its trial samples, where studies tended to focus on primary and early secondary education (Edovald & Nevill, 2020). In other words, even if trial outcomes initially took the form of a standardised test, in the longer term pupil examination results might be linked to trial samples (Education Endowment Foundation, 2012). In the criteria EEF published providing guidance to evaluators around the most suitable standardised assessments, the ability of results from a standardised assessment to be predictive of national examination results is mentioned explicitly (Education Endowment Foundation, 2012). Most frequently these will be results from English Language, Mathematics and Science examinations at GCSE (at the end of secondary school in Year 11 of the English schooling system) and KS2 SAT scores in Mathematics, English grammar, punctuation and spelling, and English reading, from assessments at the end of primary school in Year 6 of the English schooling system. Rapidly after the EEF's establishment in 2011, use of pupils' attainment in national examination as primary outcomes increased markedly (though this trend has diminished somewhat in recent years, in line with an overall reduction in the number of trials commissioned). As a result the findings of this paper have a direct bearing on current practice.

The work presented follows on from a pilot study conducted to assess the extent to which an intervention known as 'Deeper Thinking' might be evaluated experimentally drawing on national examination results in GCSE Science as a primary outcome (Smith et al., 2020). Smith et al's key finding was that the sensitivity of the proposed trial of 'Deeper Thinking' would depend on which variable was used as the primary outcome: GCSE Science grades, or marks. As noted, grades are a summary measure in which the full variation in the underlying mark distribution is not always maintained, and are therefore a 'coarser' measure of performance than marks. For example, an intervention could see a student gain several marks on their GCSE, but this may not always translate into achievement of a higher grade. To re-emphasise the point, the effects of an intervention are primarily and always captured in the mark but grades are much more commonly used as the primary outcome in EEF trials³; in large part due to their being (comparatively) easily accessible via the National Pupil Database.

In line with the above, Smith et al found that using grades as an outcome variable would result in a loss of sensitivity, relative to using marks. The key implication for evaluators was that a larger sample size would be required to maintain desired levels of statistical power when using grades as an outcome variable as opposed to an outcome measured in marks, with implications for the cost and complexity of a trial. To the best of our knowledge, this work is the only published research to evaluate the impact of using marks compared with grades as primary outcomes in randomised trials. However, Smith et al's (2020) work had several limitations. First, it focused solely on one awarding organisation's GCSE Science qualification – attainment in Science is not widely used as a primary outcome in trials, with far greater emphasis placed on attainment in Mathematics and English Language. Because GCSE Science has a different structure to Mathematics and English Language, typically being delivered as either a double award or three separate GCSEs (Biology, Chemistry and Physics), it is therefore not straightforward to generalise findings from Science to the other two subjects. Second, the distribution of the underlying marks had to be simulated for this paper.

The research reported here considers both mathematics GCSE as well as English language, and uses the actual mark distributions and grade boundaries for all four of the GCSE awarding bodies in England by way of addressing these limitations. These improvements aim to, respectively: render the results of greater relevance to the design of randomised trials using GCSE attainment as the primary outcome, and render the results more valid and robust. However, although the focus of this paper is the situation in England and specifically the use of GCSE examination results, we believe much of the evidence presented is applicable more widely; to any graded assessment being used as an outcome measure.

³ From our scrutiny of trials above, we identified no reported trials for which GCSE marks were available or used as outcome variables, and only four trials not yet at the reporting stage which intended to use GCSE marks in this capacity.

Background & context

In this section we discuss the context and broader issues that set the scene for this work. In particular, the following are considered:

1. The challenges of outcome measurement in RCTs and the use of attainment in national examinations as a primary measure
2. The National Pupil Database as a data source
3. Minimum detectable effect sizes (a key concept in this analysis)

The challenges of outcome measurement in RCTs and the use of attainment in national examinations as a primary measure

The importance of outcome definition and measurement within the context of randomised trials is widely appreciated; the canonical contribution in this area, at least within the social sciences, being Shadish, Cook, & Campbell (2002), particularly their discussion of construct validity. More recently, with education RCTs specifically in mind, Connolly et al., (2017, page 43) describe an outcome as a ‘measurable characteristic of human development that can change over-time’, making the crucial distinction between an output and outcomes in the context of logic model development for RCTs. Logic models are often used in the identification of relevant outcomes in education RCT evaluations; particularly in assessing the extent to which it is plausible that a programme might change an outcome sufficiently for such change to be measurable. In terms discussed by Connolly et al., (2017), our attention here focuses on outcomes derived from the NPD that are understood as cognitive outcomes. As Connolly et al., (2017, page 46) point out, cognitive outcomes are ‘often of most interest to potential funders like government. Furthermore, cognitive measures often have the most secure operational definition and as a result are more predictive of future outcome change’.

With respect to outcome measurement in the context of trials, concerns are frequently raised in relation to the issue of ‘blinding’ (Boutron et al., 2017). That is, those administering assessments are often aware of the intervention/control group status of pupils and this can influence both the administration and scoring of assessments and tests (Torgerson et al., 2005). Other literature makes distinctions between different forms of outcome measurement in the context of experimental research. These distinctions include those between “researcher developed” and “standardised assessments”, and between “treatment inherent” versus “treatment independent” measures (Cheung & Slavin, 2016; R. E. Slavin, 2008; R. Slavin & Madden, 2011). These distinctions reflect the fact that the choice of outcome measure is motivated by both scientific concerns as well as practical considerations.

Turning first to scientific considerations: a substantial effort is expended by the English and Welsh awarding organisations (AOs) in ensuring that equivalent standards are maintained in the design of examination questions year by year, and that the inevitable slight variations in performance are accounted for during the setting of grade boundaries. As Boyle & Mellor (2020, page 4) make clear:

“GCSE examinations are well designed – with skilled teams devoting considerable expertise to writing questions, setting standards, etc.. Also, we can be reasonably sure that exams were conducted without collusion between candidates, and that candidates were trying their hardest”

This means that although examination results cannot be aligned precisely with a particular independent or objective measure of actual attainment or mastery, they measure relative performance accurately and consistently year to year (Ofqual, 2011).

In relation to the notion of ‘treatment inherency’ in outcome measurement, it is much to the EEF’s credit that they recognised the problem with both researcher developed and treatment inherent measures from the outset (Edoald & Nevill, 2020). This concern manifested itself in many EEF trials through an emphasis on primary outcomes from standardised instruments measuring various dimensions of, for example, literacy and numeracy. These avoided the problems of treatment inherency and researcher developed measures. But they came with

other, more practical as well as some scientific disadvantages. Firstly, the assessments required administration in school settings either in the form of online or paper and pencil tests. This presented schools with a significant burden creating an incentive for schools to leave studies when they realised the potential disruption testing might cause. Second, administration of tests is a costly exercise requiring the presence of the research team or their representatives 'on the ground' in schools to collect the data, or, to contain costs, requiring teachers and/or teaching assistants to administer tests. As such, those administering tests and schools themselves are sometimes not blind with respect to the intervention/control group status of the setting, even if the scoring of tests was often blinded, semi or fully automated. Thirdly, experience suggests that it was often quite difficult to obtain details on the reliability and validity of commercial assessment instruments and scales (Allen et al., 2018).

These disadvantages acted together and individually to incentivise the adoption of attainment in national examinations as outcome measures for trials. It is quite obvious that national examinations are marked by those with no knowledge of a given trial; thus marking is 'blind' – though teachers will of course know their school is participating in a trial and therefore might be more inclined to 'teach to the test' for national examinations; though we doubt that participation in a trial adds much to the existing incentives for this. The issue of treatment inherency is also not a concern with national examinations, in the sense that examinations are general and cannot be tailored to the content of specific interventions.

In practical terms outcome measures obtained from national examinations do not involve additional costs of data collection. There may be costs associated with accessing examination results data, their extraction and manipulation into a form that can be analysed, though these costs are likely to be modest relative to those associated with administering standardised/commercial assessments to experimental samples. Perhaps the greatest cost associated with the use of NPD data, especially post-GPDR's introduction, is caused by delays in accessing data (whereas a standardised assessment's results would be available swiftly upon its completion). Whilst such costs are difficult to quantify, if research staff are required to be kept in post long enough to conduct analysis of NPD data, then the net cost could certainly equal that entailed in buying and using commercial tests.

The collection of outcome measures in the form of standardised/commercial assessments are affected quite appreciably by initial non-response and subsequent attrition. Some early EEF-funded trials in England suffered from high loss to follow-up in administering standardised/commercial tests, with some trials seeing a quarter of participants being lost following randomisation (Dawson et al., 2018). A chief concern is where missingness varies by intervention and control groups, potentially leading to biased estimates of treatment effects (Gerber, Alan & Green, Donald, 2012). National examinations, particularly those in English Language and Mathematics, are taken by virtually all students and therefore, in theory at least, trials deriving outcome measures from such examination results should encounter next to no missing data and patterns of missingness that do not differ by intervention and control groups.

Placing reliance on national examination data for outcome measurement is not, however, a silver bullet. Missingness can still occur. Results from examinations recorded in the National Pupil Database or obtained through some other means can be incomplete. Small numbers of pupils do not sit examinations. Furthermore, records from examination results recording systems need to be linked to those from the trial. Data fields used in such linking of records include Pupil Matching Reference (PMR) number, Unique Pupil Number (UPN), full name and date birth. Recording errors in one or more of these fields can lead to a failure to link records resulting in missing data. There can also be considerable challenges associated with getting schools to transfer this information correctly.

There is also a potential disadvantage in terms of treatment inherency; national assessments by design measure a broad construct – reading, literacy, or numeracy and so forth. If an intervention is very targeted on a specific aspect of one of these domains then the intervention would likely require a very sizeable impact to be able to detect an effect when measuring overall scores on the assessment as a whole. In such cases there may be merit to using narrower assessment instruments as outcome measures. Whilst many commercial standardised instruments by design mirror that of national assessments and thus are subject to this same limitation, there are instruments which focus specifically on one aspect of a wider domain and could serve this purpose.

Finally, deriving outcome measures from examination results should reduce the burden on schools from having to administer or facilitate standardised tests, and as a result of this reduced burden, lead to fewer schools

dropping out of studies. School drop-out has been a major source of loss to follow-up in early EEF trials, particularly among control schools (Dawson et al., 2018). Unfortunately, however, schools can opt to leave experiments for reasons other than the burden of data collection.

The National Pupil Database

The National Pupil Database (NPD) is an administrative dataset collated and managed by the English government's Department for Education (DfE). The NPD holds information from a wide range of sources on children and young adults (aged between two and twenty-one) in England as they progress through the education system (Jay et al., 2019). Key data sources contained within the NPD include the school census, through which a range of demographic information about children is gathered, and information on children's performance in general qualifications sat by entire national cohorts, including KS1 and KS2 National Curriculum Tests (NCTs), GCSEs (and equivalents) and A-levels (and equivalents). A number of statistics available in the NPD are not directly gathered but are computed from other variables, such as EBacc achievement: whether a child has achieved certain grades in various combinations of GCSE subjects⁴, for example Attainment and Progress 8 measures.

Whilst NPD extracts have long been utilised for research, facilitating research is not the database's primary purpose. It is chiefly intended to produce key accountability measures for English schools, in the form of DfE's "school performance tables", with research a secondary or tertiary aim. Indeed, this is the case across many jurisdictions, with examination results increasingly seeing use to assess school/teacher performance as well as act as a target variable for a range of policy interventions (Murnane & Nelson, 2007).

Nonetheless, the DfE have remained committed to opening up the NPD to researchers, who are able to request access to NPD data to facilitate their work, subject to a number of rules and conditions⁵. Prior to 2018, these conditions were not particularly onerous and researchers were able to remotely receive data extracts with little direct oversight. In 2019 a new model was adopted, whereby access to NPD data is permitted only through the Office for National Statistics (ONS)'s Secure Research Service (SRS). The main implications of this are that all researchers who wish to access the NPD must have an accreditation under the ONS approved researcher scheme, and can only access the NPD in this secure environment (either from ONS premises, or through those of an organisation approved for remote SRS access).

Despite the increased complexity of accessing NPD data, performance data from general qualifications has substantial potential in terms of providing an outcome measure for EEF trials. Indeed, we calculate that of the 154 trials commissioned by EEF since 2011 (at the time of writing), 40 rely on the NPD for their primary outcome and 28 specify attainment at GCSE for this purpose. Note that whilst many GCSEs are not taken by the entire cohort, key subjects such as GCSE English Language and GCSE Mathematics are, and these are indeed the GCSEs most commonly used as outcome measures in EEF-commissioned studies (21 of the 28 studies that declared GCSE attainment to be the primary outcome had either English Language and/or Mathematics as its focus; the remainder largely focused on overall attainment figures such as Attainment 8).

The attainment information held for each type of general qualification varies somewhat, with GCSE attainment being recorded as the grade each student achieves. Though grading scales have changed over time, currently GCSEs grades run from 9-1, with a grade U to capture students who do not perform at a level that warrants the award of a grade.

GCSE grades loosely reflect a particular level of attainment; whilst the purpose of Ofqual's comparable outcomes awarding approach (Ofqual, 2011) is to maintain the standard of performance needed to achieve each grade over time. There is no explicit "meaning" to achieving a particular grade, in that achieving a grade does not indicate that a student is competent at particular set of defined tasks. This has been conceptualised in the literature as "weak criterion referencing" (Baird, Cresswell, & Newton, 2000).

KS2 (and KS1) NCTs, however, are norm-referenced (Bond, 1996), meaning their scaled scores (STA, 2020) are standardised scores (Lawley, 1950), and thus indicate not an absolute level of performance, but students'

⁴ <https://www.gov.uk/government/publications/english-baccalaureate-ebacc/english-baccalaureate-ebacc>

⁵ <https://www.gov.uk/guidance/how-to-access-department-for-education-dfe-data-extracts>

relative position within the distribution of performance achieved by an entire national cohort. KS2 scaled scores range from 80-120, whilst KS1 scaled scores range from 85-115. This means that NCT scaled scores are a more fine-grained measure of performance than GCSE grades. Whilst each of these outcome measures is available in the NPD, the above differences necessitate a different approach when working with NCT scaled scores compared to GCSE grades.

However, the raw marks underlying GCSE grades and NCT scaled scores are not available in the NPD. This is an important factor to consider in electing to use the NPD as a data source, since grades and scaled scores are a more coarse metric than marks; not all changes in marks result in a GCSE grade changing, for instance. The increased coarseness of grades is the key issue this paper seeks to investigate: whether there is a disadvantage of using grades (rather than marks) as an outcome measure in that their increased coarseness impacts on our ability to detect an effect. If there is a material advantage to using marks, then an evaluator looking to gather raw mark data would need to approach AOs with schools' permission, which introduces additional considerations and costs in evaluation design. Historically, accessing data via the NPD has been more straightforward and cheaper than gathering marks, with a significant majority of evaluations with GCSE as an outcome measure relying on grade information from the NPD. However, with the advent of GDPR there are arguably now challenges inherent in accessing results from the NPD which render this process more comparable to that of gathering marks from AOs.

The information held in the NPD also provides a useful means of assessing and tracking over time the 'attainment gap', a key interest of funders such as EEF. The attainment gap is the difference in average levels of attainment (in national examinations and NCTs) between pupils deemed 'disadvantaged' and all other pupils. Arguably the most well-known metric used as a proxy indicator of socio-economic disadvantage in the UK is students' eligibility for free school meals (FSM). Whilst in many cases students are eligible for FSM due to coming from low-income households, there are also a number of other factors which make a student eligible (DfE, 2018); hence FSM is only a proxy for deprivation. FSM eligibility for a student can change over time, as parental income and circumstances changes. FSM eligibility information is reported by schools to the DfE in the termly school censuses, and for the NPD the DfE automatically computes a number of variables from this termly FSM information to provide more robust indicators of deprivation than whether the student was eligible for FSM in the most recent school census. These include whether students were eligible for FSM at any point over the last few (three or six) years, and whether students have ever been eligible for FSM at any point in their school career.

The other deprivation variable the NPD holds is the Income Deprivation Affecting Children Index (IDACI) (Ministry of Housing Communities & Local Government, 2019), which denotes the proportion of children in an area living in income deprived families. It is not reported on the school census like FSM eligibility, but is instead merged into the NPD from other government datasets based on the postcode each child lives at (home postcode is reported in the school census). The IDACI is presented in the NPD as both the raw index value and as a rank derived from the index value.

Taken together therefore, containing records of attainment in both national examinations and NCTs, as well as proxy indicators of student deprivation, the NPD is a vital source of information for both tracking attainment over time, assessing the extent of the attainment gap as well as a source of outcome measurement for evaluation.

The minimum detectable effect size

The purpose of this paper is to examine the implications of deriving outcome measures from the NPD in the form of grades achieved in GCSE English Language and Mathematics. To do this, we define a series of trial designs and a single metric to distinguish the design sensitivity of these designs, namely the minimum detectable effect size (MDES). The minimum detectable effect size was first introduced by Bloom (1995) as a way of summarising the sensitivity of a given trial design. At the planning stage of a trial, the MDES is the size the true effect of an intervention must reach for rejection of the null hypothesis in 95 per cent of identical trials with 80 per cent power. The MDES is expressed in units of standard deviations. Thus, for example, in a trial with an MDES of 0.25, the true population effect must reach a quarter of one standard deviation in magnitude for statistical tests of the null hypothesis to result in rejection of the null in 95 per cent of identical trials over the long run. A trial with an MDES smaller relative to some other trial design is capable of distinguishing a smaller difference between intervention and control groups for a given level of statistical significance and sample power, and is therefore more sensitive.

Typically, researchers use the MDES in planning a trial in two different ways. First, they may consider a range of possible sample sizes and the MDES associated with each of these respective designs. Researchers select the study design in terms of sample size consistent with the MDES equivalent to some minimally important difference established in light of theory and/or other existing evidence. Alternatively, if fixed and variable costs of data collection and programme delivery are known, an MDES can be calculated consistent with sample sizes relating to different budgets. Judgements can then be made as to whether the budget and implied sample size are sufficient that the MDES is again equivalent to a minimally important difference.

For the analysis we conduct here, a range of MDESs have been chosen that represent designs with different levels of sensitivity. A range of trial designs are simulated consistent with the chosen MDES values.

Whereas EEF trials tend to be designed as cluster randomised trials, involving the randomisation of pupils in groups to treatment and control, this paper focuses on individual-level randomisation. This is necessary given the structure of the data available, (discussed in the next section), where there is no indication of what school or class grouping an individual pupil belongs to⁶. As a result the simulations presented here assume that randomisation takes place at the individual student-level, the level at which outcomes are also measured. Clustering of observations at any higher level units (for example classes or schools) is ignored. Thus, we proceed on the basis that the estimator from which sample estimates of the average treatment effect are obtained is the bivariate regression model of the following form:

$$Y_i = \beta_0 + \beta_1 T_i + \varepsilon_i \dots \dots \dots [1]$$

Here Y_i represents either GCSE mark or grade for student i and T_i is an indicator variable which takes the value '1' for pupils allocated to the intervention, '0' otherwise. The model includes the disturbance term ε_i which is assumed to take the usual independent and identically normally distributed form. An effect size is derived from the sample estimate of β_1 via ordinary least squares and is standardised by dividing through by the total pooled variance of Y (Hedges, 2011). Given this set up Dong & Maynard (2013) provide an equation for the minimum detectable effect size:

$$MDES = M_{n-2} \sqrt{\frac{1}{nP(1-P)}} \dots \dots \dots [2]$$

Here M is 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. The term n represents the total number of pupils in the sample whilst ' P ' the proportion of pupils randomised to the intervention. The equation for an MDES can be re-arranged as follows in terms of the sample size ' n ':

$$n = \left(\frac{M_{n-k-2}}{MDES} \right)^2 \left(\frac{1}{P(1-P)} \right) \dots \dots \dots [3]$$

As a result, for any selected MDES that is used as a starting point in our calculations we can examine the consequences in terms of the number of pupils a study might require.

A note on terminology

The methodology that is used in this paper centres on simulating a large number of randomised trials based on different study designs, by drawing repeated samples from the population of pupils that sat English Language and Mathematics GCSEs in the summer of 2019. This is done so that we can explore the consequences for study design of measuring outcomes in examination grade instead of examination marks. As discussed, we use the MDES as the metric by which we distinguish between different trial designs. The MDES is expressed in units of standard deviations. As will be seen below, the data we have enable us to calculate the standard deviations for the population mark and grade distributions, so that we can express the MDES as minimum detectable effect

⁶ Note that the findings from our analysis are not substantively affected by the decision to conduct our analysis assuming individual randomisation rather than cluster randomisation; that is a design that involves randomising groups of students, typically classes or whole schools, rather than individual pupils. An empirical example illustrating these points is provided at Annex 1.

(MDE) in units of the grade and mark. This means we can provide a summary metric for the different trial designs examined in this paper in units of the standard deviation, the mark and the grade separately.

Once results from a simulation are available, the difference in the sample means between pupil marks or grades is calculated on the basis of the estimator at equation [1] for a simulated trial sample. We refer to these differences as estimates of the sample average treatment effect (SATE). For example, if we conduct 5,000 simulations of a particular study design we will obtain 5,000 SATEs estimates. The centre of the distribution of SATEs for a given design is referred to as the population average treatment effect (PATE). If for a particular design, a) the MDES is converted to MDE in marks, b) the relevant sample size chosen, c) 5,000 samples of the sample size chosen are drawn, d) 5,000 trials simulated and e) 5,000 SATEs obtained, the resulting PATE will by design be equal to the MDE in marks and the study's MDES.

Method

The purpose of our analysis is to examine the consequences for trial design of selecting GCSE attainment measured in units of the grade, in either English Language or Mathematics, as a primary outcome. We aim to address the following research questions, which focus on vital aspects of trial design:

- a) Does using grades (as opposed to marks) as the outcome variable lead to decreased sensitivity/statistical power, for each of:
 - i. GCSE English Language
 - ii. GCSE Mathematics
- b) Does using grades as the outcome variable necessitate a larger sample size to maintain comparable statistical power to using marks, for:
 - i. GCSE English Language
 - ii. GCSE Mathematics

Note that research question b) follows from question a); if there is a reduction in power associated with using grades as the outcome variable, then a larger sample size would always be required. As such the key factor is how much larger a sample size would be required if a reduction in power is observed.

As a summary of our method, the above research questions will be addressed in the following ways:

- a) Sensitivity will be investigated by simulating many trials and assessing if, across these simulated trials, the proportion of rejections of the null hypothesis alters during simulations using grades as the outcome variables, as this is indicative of a change in power.
- b) Sample size will be investigated by using standard formulae to deduce the sample size associated with the statistical power observed in practice in the above simulated trials. The sample size obtained in this manner can then be compared to that associated with the initial MDES, showing the extent to which a loss or gain in power manifests as an impact on the sample size required to detect an effect.

The detail of the analysis can be characterised as a series of steps, which we now discuss in more depth:

1. We select a range of minimum detectable effect sizes that represent a range of trial designs. By trial designs, we mean a design of a given sample size in terms of pupil numbers consistent with the chosen MDES, statistical significance of 95 per cent and sample power of 80 per cent. Each simulated trial replicates a study design in which individual pupils are randomised to treatment and control groups on a 1:1 basis. All tests for statistical significance are assumed to be uni-directional.
2. For each of these designs, represented by the chosen MDES, we calculate the required sample sizes for each set of simulated trials using equation [3] above. Table 1, below, sets out the chosen MDESs and associated sample sizes.

Table 1: Sample sizes commensurate with each MDES investigated

MDES	n (total sample size)*
------	------------------------

0.25	398
0.2	620
0.15	1,102
0.1	2,476
0.05	9,894
*calculations are performed using Power-Up in R statistical software (R Core Team, 2017; Bulus, Dong, Kelcey, & Spybrook, 2019). A one sided test of statistical significance is assumed, with random assignment of individual pupils to treatment and control on a 1:1 basis, statistical power of 80 per cent and statistical significance of 95 per cent.	

3. We obtained data on the entire distribution of marks received in English Language and in both tiers of Mathematics GCSE, in England, covering four Awarding Organisations (AOs) via the Joint Council for Qualifications, for examinations sat in the summer of 2019. The mark distributions were mapped onto published grade boundaries for each AO. These data are described more fully below. From these data a set of 5,000 trials were simulated⁷, with each set of simulated trials consistent with one of the chosen MDESs revealed in Table 2 and one of the AOs. For instance, for a design associated with an MDES of 0.25, 5,000 trials were simulated for AO 1. Each of these 5,000 simulated trials carried out the following four steps:
 - a. First, a sample of pupils was selected; for the trial associated with an MDES of 0.25, for example, for a given AO, the sample size for each simulation was $n=398$ pupils.
 - b. Second, this sample was allocated at random (using simple random assignment) to treatment and control groups.
 - c. The relevant MDES, in this case equal to 0.25, was then converted into a minimum detectable effect (MDE) in marks by multiplying the MDES by the standard deviation of the mark distribution for the relevant AO.
 - d. Finally, the resulting MDE was then added to the mark observed for each pupil assigned to the treatment group; marks for pupils allocated to the control group remained unchanged.
4. Having simulated distributions of marks under treatment and control conditions as described, the SATE *in marks* was then obtained for each simulated trial on the basis of equation [1] above. Through this process the distribution of the SATEs about the PATE was obtained, where the PATE in marks by design equals the MDES. Also by design, each of these sampling distributions represented power of 80 per cent, which meant that in 20 per of trials within each set of 5,000 simulations, a significance test would result in failure to reject the null, that is a Type II error would be made.
5. Based on our knowledge of the distribution of marks and how they relate to grades for each AO the mark distribution can be converted into grades for each trial sample. Thus we now have a distribution of grades achieved under treatment and control for each simulated trial derived directly from the underlying distribution of marks. For each simulated trial we re-estimate the SATE *in grades* and conduct a null hypothesis significance test⁸. Across each of the 5,000 simulated trials for a given set of trials, if the proportion of significance tests that results in rejection of the null hypothesis deviates from 80 per cent then power has either been gained or lost due to conversion of marks to grades. If the rejection rate falls below 80 per cent, which we might expect a priori, then power is lost as a result of the switch from marks to grade. Likewise, we might expect PATE in grades to no longer equal the MDES for a given set of simulations.
6. Another way of assessing the consequences of choosing grade as an outcome is to examine the impact on sample size. We have calculated the required sample sizes consistent with the chosen MDESs, which is by design consistent with the PATE *in marks* for each of the chosen study designs. We can, however, re-compute required sample sizes (based on equation 3 above) for an MDES equal

⁷ 1,000 trials was considered slightly too few to adequately minimise error resulting from the simulations; 5,000 was selected rather than an order of magnitude more (10,000) due to halving the considerable processing time required for the analysis.

⁸ Note that because we could expect the treatment group to improve their marks due to the intervention, we utilised a one-tailed test.

to the PATE *in grades*⁹, which as outlined above, may not be identical to the PATE in marks. This is again achieved through using the PowerUpR() package (Bulus, Dong, Kelcey, & Spybrook, 2019). Comparing the sample size obtained in this manner to that originally obtained based on the initial MDES shows the extent to which the loss or gain in power manifests as an impact on the sample size required to detect an effect.

Data

The key qualifications investigated in this paper are GCSE English Language and GCSE Mathematics; they were selected for a number of reasons. Firstly, these GCSEs are taken by almost every child in England, meaning a substantial population-level dataset is available. Secondly, unlike GCSE Science which is also taken by all children, English Language and Mathematics are more commonly observed as outcome measures in EEF trials, and are single award GCSEs (Science can be taken as a single or double award, or as separate sciences), simplifying the analysis considerably. Thirdly, whilst conducting similar analysis on KS2 SATs standardised scores would be interesting, GCSE grades are a considerably coarser metric than KS2 standardised scores, so a larger impact of using grades as an outcome measure than standardised scores can be presumed.

To investigate the impact of using grades or marks as an outcome variable in GCSEs, it is desirable to know the precise distribution of raw marks. Marks underly and determine the grade each pupil receive, and as a result, the effect of an intervention must influence the underlying mark distribution. The most recent GCSE series unaffected by the Covid-19 pandemic is summer 2019 (with GCSEs now linear in structure, the entire cohort sits their GCSEs in summer; other series are solely for resits). We requested and were provided each English AO's subject-level mark distribution for GCSE English Language and Mathematics in summer 2019 from the Joint Council for Qualifications (JCQ). JCQ represents the UK's eight largest awarding organisations (AOs), and thus handles such data requests on behalf of them.

Some important constraints were placed on the data requested: only data from candidates in England was requested, as the regulation and grading of GCSEs is not comparable across the various jurisdictions within the UK. Further, only candidates in the typical year for sitting GCSEs were included (Year 11, so age 15 or 16). This excludes resitters and early-entry candidates (in addition to any adult learners), since educational interventions are most likely to follow students on a typical progression through school examinations.

In addition to mark distribution data from JCQ, publicly available grade boundary position information was collated from each English AO's website. This permitted the subject-level mark distribution data to be converted to a grade distribution as described above. Whilst grades are strictly speaking, categorical, it is common to treat them as numeric data to facilitate quantitative analysis (indeed, the NPD's conversion of grades into Points Scores does this explicitly). With the new 9-1 grading scale in use for GCSEs this is fairly intuitive, with the exception of grade U (unclassified, given to candidates who fail to achieve the lowest possible grade), which is recoded as "0" in all subsequent analysis.

It is important to note that GCSE Mathematics is tiered, with a more challenging (Higher) and less challenging (Foundation) tier, each permitting candidates to achieve a different range of grades. The Higher tier allows grades 9-3, with candidates falling below the standard of a 3 receiving a U, whilst Foundation tier allows grades 5-1 (with U for anyone who does not achieve the standard of a 1). The cohort is split across the two tiers, with a slight majority of candidates sitting Higher.

Whilst the data supplied from JCQ comprises the entire (typical age) 2019 summer cohort and can thus be considered population-level data, each of England's AOs delivers a different GCSE exam per subject. Each AO's GCSE in each subject will see a different mark and grade distribution, different grade boundaries, and potentially different numbers of marks available. As such, the analysis was carried out separately on each AO's data for a given subject, and slightly different findings for each AO's GCSEs are to be expected.

⁹ A complexity again arises due to the nature of tiered assessments: for Mathematics specifications, there will be two PATEs in grades; one for Foundation and one for Higher tier. In order to derive a single PATE to compare to the MDE in grades, a weighted average of the two tiers' PATE in grades is computed (with the number of candidates taking each tier applied as weights). This average will then be used to derive the sample size needed to detect an effect of this size.

The AO-level analyses are, however, individually not nationally representative, and pragmatically many educational interventions are *unlikely* to recruit candidates sitting only one AO’s GCSE English Language or Mathematics examination. So, in order to produce more informative population-level results, the majority of the analyses include an “all AOs” weighted average figure (with weighting conducted by the number of candidates per AO).

Similarly, Mathematics having two tiers means that each AO’s Mathematics GCSE has two separate mark distributions, despite the fact that the resulting two grade distributions can be combined. As such, when dealing with mark-level analyses for Mathematics, a further level of weighted averaging is conducted to produce “all AOs” figures for each tier, and for Mathematics overall, combining both tiers (the weights used being the number of candidates sitting each tier in each AO).

Results

Change in sensitivity when grades are the outcome measure

Mark differences required to detect an effect

Tables 2 and 3 below show, for each trial design and its associated MDES and for each of the four AOs, the MDE in marks for GCSEs English Language (Table 3) and mathematics (Table 3). To convert the MDES to an MDE expressed in units of marks we multiply the MDES by the standard deviation of the mark distribution for the relevant AO. Because the standard deviation of the mark distribution varies by AO, the MDE in marks will also vary by AO.

For example, consider ‘AO 1’ in Table 2. For a design with an MDES of 0.05 (that is, five per cent of one standard deviation) the MDE in marks is 1.167 marks, or 0.05×23.338 . Incidentally, an MDE of this order of magnitude is unlikely to be meaningful in policy or operational terms. Small MDES such as this are included to help illustrate the consequence of using grades as an outcome even though such effect sizes are unlikely to be a target for inference in an actual trial. Likewise for a design with an MDES of 0.25 for AO 1, the MDE in marks is 5.834. What this means is, given a trial performed on pupils sitting this AO’s GCSE, where attainment in GCSE English Language (in marks) is the trial outcome and the trial sample size $n=398$ (see Table 2), the true effect of the intervention would have to be 5.8 marks for tests of the null hypothesis in repeated trials of the same size to reject the null in 80 per cent of those trials.

Table 2: Mark differences required to detect an effect – English Language

	MDES	AO 1	AO 2	AO 3	AO 4
Mean mark	n/a	85.508	88.978	94.750	87.121
SD	n/a	23.338	28.233	24.519	29.331
MDE (marks)	0.05	1.167	1.412	1.226	1.467
	0.1	2.334	2.823	2.452	2.933
	0.15	3.501	4.235	3.678	4.400
	0.2	4.668	5.647	4.904	5.866
	0.25	5.834	7.058	6.130	7.333

Table 3: Mark differences required to detect an effect – Mathematics

		Foundation tier			Higher tier		
	MDES	AO 1	AO 2	AO 3	AO 1	AO 2	AO 3
Mean mark	n/a	106.408	128.490	131.020	122.144	167.126	126.660
SD	n/a	43.964	54.316	45.426	49.605	58.888	43.441
MDE (marks)	0.05	2.198	2.716	2.271	2.480	2.944	2.172

0.1	4.396	5.432	4.543	4.960	5.889	4.344
0.15	6.595	8.147	6.814	7.441	8.833	6.516
0.2	8.793	10.863	9.085	9.921	11.778	8.688
0.25	10.991	13.579	11.357	12.401	14.722	10.860

Grade differences required to detect an effect

In addition to the above two tables showing MDEs in marks, we present the MDEs in grades necessary to detect true effects for the same five MDES values. This is done by following exactly the same process described above for marks using the grade distribution. Note that whilst the two tiers in Mathematics have completely distinct mark distributions, once converted to grades they comprise the same grade distribution.

Table 4: Grade differences required to detect an effect – English Language

	MDES	AO 1	AO 2	AO 3	AO 4
Mean grade	n/a	4.702	5.210	4.862	4.712
SD	n/a	1.862	2.112	1.916	1.978
MDE (grades)	0.05	0.093	0.106	0.096	0.099
	0.1	0.186	0.211	0.192	0.198
	0.15	0.279	0.317	0.287	0.297
	0.2	0.372	0.422	0.383	0.396
	0.25	0.466	0.528	0.479	0.495

Table 5: Grade differences required to detect an effect – Mathematics

	MDES	AO 1	AO 2	AO 3
Mean grade	n/a	4.674	4.285	4.676
SD	n/a	2.102	2.116	2.075
MDE (grades)	0.05	0.105	0.106	0.104
	0.1	0.210	0.212	0.208
	0.15	0.315	0.317	0.311
	0.2	0.420	0.423	0.415
	0.25	0.526	0.529	0.519

The above tables can be used to practically highlight a key element of our rationale for this analysis. Taking AO1's English Language specification as our example, [Table 2](#) shows that for an MDES of 0.25 the intervention must result in a true effect 5.8 marks in size to be able to reject the null hypothesis in 80 per cent of trials. [Table 4](#) shows that for the same MDES, if measuring outcomes in grades rather than marks, a true effect 0.47 grades in size must manifest to achieve the same level of power. It is not possible to linearly convert grades to marks, but the average grade width for this AO's English Language specification was 13.9 marks. Multiplying this by 0.47 gives 6.5 marks, which is higher than the 5.8 marks evident in [Table 2](#). This implies that using grades as the outcome measure means a larger PATE is necessary to achieve the same power (but does not prove it as the conversion from marks to grades has been simplified).

Is statistical power lost or gained when re-estimating SATEs using grades?

Having shown how we can convert MDES to MDEs defined in terms of marks and grades we return to the question at hand, namely the consequences of measuring outcomes in grades instead of marks (in a more empirical manner than considering average grade widths). To reiterate, for each subject and AO, we simulate 5,000 trials, one set of simulations for each design consistent with a specified MDES. In these simulations we

use marks as the dependent or outcome variable. For each set of simulated trials we calculate 5,000 SATEs in marks, one for each separate simulation, and from the resulting distribution the PATE in marks.

For each simulation we then convert the simulated outcomes under treatment and control conditions into grades by examining the mark distribution in relation to its grade boundaries and allocating each mark a grade on this basis. Having done this, we then re-estimate the SATEs in grades. We can then observe how many of these SATEs fall within the rejection region. If the proportion of trials in which the null hypothesis is rejected differs from 80 per cent, statistical power has either been gained if this figure is greater than 80 per cent, or lost if it is less than 80 per cent. [Table 6](#) provides the results of this analysis for English and [Table 7](#) for mathematics.

Table 6: Proportion of simulations H_0 rejected in (outcome variable = grades) – English language

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

Table 7: Proportion of simulations H_0 rejected in (outcome variable = grades) – Mathematics

MDES	Foundation tier				Higher tier				All AOs, both tiers
	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

From the above two tables, it is immediately clear that all values are somewhat below 0.8, indicating that when grades are the outcome variable, there is a tendency for the statistical power to fall relative to power where outcomes are measured in marks.

In order to better compare the results in marks and grades (the latter of which is shown in the tables above), the difference between the proportion of simulations the null hypothesis was rejected when working in marks and grades is displayed in the tables below. These figures are a further quantification of the change in power when using grades rather than marks as the outcome variable.

Table 8: Difference between the proportion of simulations H_0 rejected in (grades - marks) – English language

MDES	AO 1	AO 2	AO 3	AO 4	All AOs
0.05	-0.055	-0.093	-0.067	-0.102	-0.062
0.1	-0.046	-0.097	-0.056	-0.084	-0.053
0.15	-0.037	-0.077	-0.035	-0.067	-0.041
0.2	-0.010	-0.068	-0.022	-0.038	-0.015
0.25	-0.006	-0.046	-0.004	-0.033	-0.009
0.5	0.002	-0.041	0.004	-0.015	-0.001

Table 9: Difference between the proportion of simulations H_0 rejected in (grades - marks) – Mathematics

MDES	Foundation tier				Higher tier				All AOs, both tiers
	AO 1	AO 2	AO 3	All AOs	AO 1	AO 2	AO 3	All AOs	
0.05	-0.142	-0.141	-0.146	-0.145	-0.121	-0.164	-0.157	-0.147	-0.146
0.1	-0.094	-0.089	-0.086	-0.088	-0.054	-0.087	-0.091	-0.080	-0.084
0.15	-0.072	-0.066	-0.062	-0.065	-0.046	-0.056	-0.068	-0.061	-0.063
0.2	-0.055	-0.057	-0.058	-0.057	-0.035	-0.060	-0.056	-0.050	-0.054
0.25	-0.060	-0.060	-0.052	-0.054	-0.031	-0.047	-0.044	-0.041	-0.047
<i>0.5</i>	<i>-0.062</i>	<i>-0.068</i>	<i>-0.049</i>	<i>-0.054</i>	<i>-0.028</i>	<i>-0.034</i>	<i>-0.038</i>	<i>-0.035</i>	<i>-0.044</i>

From [Table 8](#), we can see that in English Language, across all AOs, the average loss in power associated with using grades rather than marks is between 0.9 and 6.2 per cent. For Mathematics in [Table 9](#), the average loss in power is larger at comparable MDES values, ranging from a loss of 4.7 to 14.6 per cent. For both subjects and across AOs, there is a consistent trend for a loss in power to occur, and that this loss tends to be greater with designs with smaller MDES.

This trend across MDES values is to be expected. With a larger MDES we model larger MDES for the treatment group. Because grade changes are governed by candidates' marks passing over one (or even several) grade boundaries, as the MDES get closer to grade boundaries' widths, more of the treatment group will gain a grade and there will be less loss in power associated with using grades as the outcome variable.

If anything, it is perhaps slightly surprising that the values observed in [Table 8](#) for an MDES of 0.25 (and AOs 1 and 3) are so close to zero, as the PATE in grades modelled for these AOs is still around half their average grade boundary width. This appears to be attributable to the PATE in marks also being close to around half the average grade boundary width, leading to there being little difference between using marks and grades as the outcome variable in terms of sensitivity/power.

To check whether the observed trend would continue at substantially higher MDES values (i.e. grades would eventually be found to be more sensitive than marks), an additional MDES scenario was run for an MDES of 0.5. For brevity, the results are only included in the bottom italicised row of [Table 8](#) and [Table 9](#). These show very similar results to those for an MDES of 0.25 for both subjects, namely near zero difference in power for English Language, and 4.5 per cent fewer rejections for Mathematics.

This suggests that there is a cap to the loss of power; once MDES reaches around 0.25. Indeed, the loss in power with higher MDES appears to follow a negative exponential relationship; there is great change in power at smaller MDES values, but as higher MDES values are reached the change in power plateaus. It is however interesting that the plateau for Mathematics is not at zero difference; in our simulations there is always around a 4.5 per cent power loss at minimum when grades are used as the outcome variable.

Change in sample size required when grades are the outcome variable

In the subsequent tables, for each set of simulations we convert the PATE obtained in marks (which by design is nearly identical to the MDES modelled) into grades, through consulting the distribution of marks and using grade boundaries to impose grades upon each simulated student's mark. Subsequent to this, we can use the standard deviation of the grade distribution for each AO to standardise the PATE; the resulting standardised PATE can be directly compared to the initial MDES. Note that as outlined above, the subject grade distribution spans both tiers, so for Mathematics no tier breakdown of information is needed, and both tiers' data are summarised in a single column.

Table 10: Standardised PATEs (grades) – English Language

	MDES	AO1	AO2	AO3	AO4	All AOs
SD	n/a	1.862	2.112	1.916	1.978	n/a

Standardised PATE (grades)	0.05	0.046	0.044	0.045	0.043	0.046
	0.1	0.094	0.088	0.092	0.088	0.093
	0.15	0.147	0.135	0.144	0.138	0.145
	0.2	0.200	0.184	0.196	0.189	0.198
	0.25	0.254	0.234	0.250	0.239	0.251

Table 11: Standardised PATEs (grades) – Mathematics

	MDES	AO1	AO2	AO3	All AOs
SD	n/a	2.102	2.116	2.075	n/a
Standardised PATE (grades)	0.05	0.030	0.027	0.028	0.028
	0.1	0.063	0.059	0.059	0.060
	0.15	0.097	0.092	0.092	0.093
	0.2	0.128	0.123	0.124	0.125
	0.25	0.162	0.152	0.155	0.157

In accordance with our findings above discussing the loss of power that results from the switch from marks to grades, the standardised PATE in grades is typically smaller than the MDES associated with the specific trial design under consideration. There is one exception to this: AO1's English Language for an MDES of 0.25, where the observed standardised PATE is 0.254. The standardised PATEs for Mathematics are also markedly smaller than those for English Language, again in accordance with the finding of a greater loss of power where attainment in Mathematics is considered, when using grades as the outcome variable.

Having converted the PATEs each set of trials into grades and standardising these, we can now ask the question as to what change in the sample size would be necessary if the resulting standardised PATEs were used to inform MDES for future trials. As discussed above, each initial MDES represents a different trial design, which in effect means a different trial sample size. Now, we can use the standardised PATEs which emerged from the analysis (Table 10 and Table 11) as if they were MDES informing a new set of trials, and ask: how much larger/smaller would these new trials' samples have to be?

We can then calculate a ratio of these new sample size estimates over the initial sample sizes associated with the starting MDESs in Table 2. If this ratio is greater than '1', then relying on grades as an outcome measure (rather than marks) requires a larger sample size for a trial of the same MDES. Likewise if the ratio is less than '1' a smaller sample size is required. Table 12 and Table 13 below show these sample size ratios for English Language and Mathematics respectively.

Table 12: Ratio of sample sizes needed to detect an effect (marks:grades) – English Language

MDES	AO 1	AO 2	AO 3	AO 4	All AOs
0.05	1.159	1.307	1.219	1.358	1.187
0.1	1.123	1.295	1.151	1.285	1.146
0.15	1.065	1.233	1.074	1.188	1.082
0.2	1.013	1.191	1.032	1.121	1.029
0.25	0.980	1.149	0.993	1.091	0.995

Table 13: Ratio of sample sizes needed to detect an effect (marks:grades) – Mathematics

MDES	AO 1	AO 2	AO 3	All AOs
0.05	2.868	3.366	3.308	3.189
0.1	2.521	2.842	2.847	2.756

0.15	2.420	2.700	2.673	2.607
0.2	2.389	2.689	2.610	2.556
0.25	2.391	2.671	2.566	2.531

From the above tables, we can see that for English Language, switching from outcomes measured in grades rather than marks necessitates a slightly larger sample size (with the exception of AO1 and AO3's 0.25 scenarios). For example, an English Language trial powered to MDES 0.15 would need to be approximately 89 pupils larger (1,172 as opposed to 1,083) if grades are used as the outcome variable. Consistent with discussions around the loss of power, at lower MDESs the additional sample required is larger than for MDESs larger in magnitude. Turning to Mathematics, the position is much more stark. Particularly for lower MDES scenarios, samples nearly three times those where outcomes are measured in marks are required when outcomes are measured in grades.

As should be apparent from these figures, for English Language the increase in sample size needed if using grades is relatively modest, or even non-existent for higher MDES values. However for Mathematics, the sample size needed to detect the same effect if measured in grades as opposed to marks is always at minimum 2.5x that which marks would require¹⁰. A sample size this much larger would have substantial cost implications in a RCT.

Discussion

Key findings

Is there a loss in sensitivity associated with using grades as an outcome variable?

The key finding from the above is that there is a loss in sensitivity/power associated with using grades rather than marks as the outcome variable for a given true effect size. From [Table 8](#), in English Language across all AOs, the average loss in sensitivity associated with using grades rather than marks as the outcome ranges from a decline in statistical power of 0.9 per cent, to a fall in statistical power of 6.2 per cent (over the nominal rate of 80 per cent). For Mathematics ([Table 9](#)), the average loss in power is larger at comparable MDES values, ranging from a loss of power of 4.7 per cent to a loss of 14.6 per cent. For both subjects and across AOs, there is a consistent trend for greater loss in power to occur with smaller MDES values (with AO1's Foundation tier Mathematics the sole exception, as there is a greater loss in power for an MDES of 0.25 than 0.2).

What are the implications of this finding for necessary sample size?

The implications these findings have for sample size, whilst not our primary metric of power, are perhaps more notable than the loss of sensitivity outlined above. From [Table 12](#), we can see that for English Language, the sample size needed to detect a specified MDES, if grades are used as the outcome variable instead of marks, can be up to 1.18x larger than required if marks were the outcome variable. For Mathematics ([Table 13](#)), the increase in sample size associated with using grades is much larger, at between 2.53x and 3.19x the sample size needed if marks are used. In both cases, larger increases in the necessary sample size occur for lower MDES values, which mirrors the above findings for loss of statistical power.

How does the observed effect vary for trials powered to detect different effect sizes?

To clarify the relationship between the loss of power that results from moving from marks to grades for different trial designs we plot the target MDESs against the statistical power achieved in our simulations when grades are the outcome variable (from the All AOs columns in [Table 6](#) and [Table 7](#)) in the following two charts. The

¹⁰ As noted above, the same bears out for trials using cluster randomised designs (as opposed to individual level randomisation as this paper simulates). For further detail on why this is the case, see Annex 1.

relationship between the loss of statistical power at different levels of MDES appears to follow an approximate inverse exponential trend.

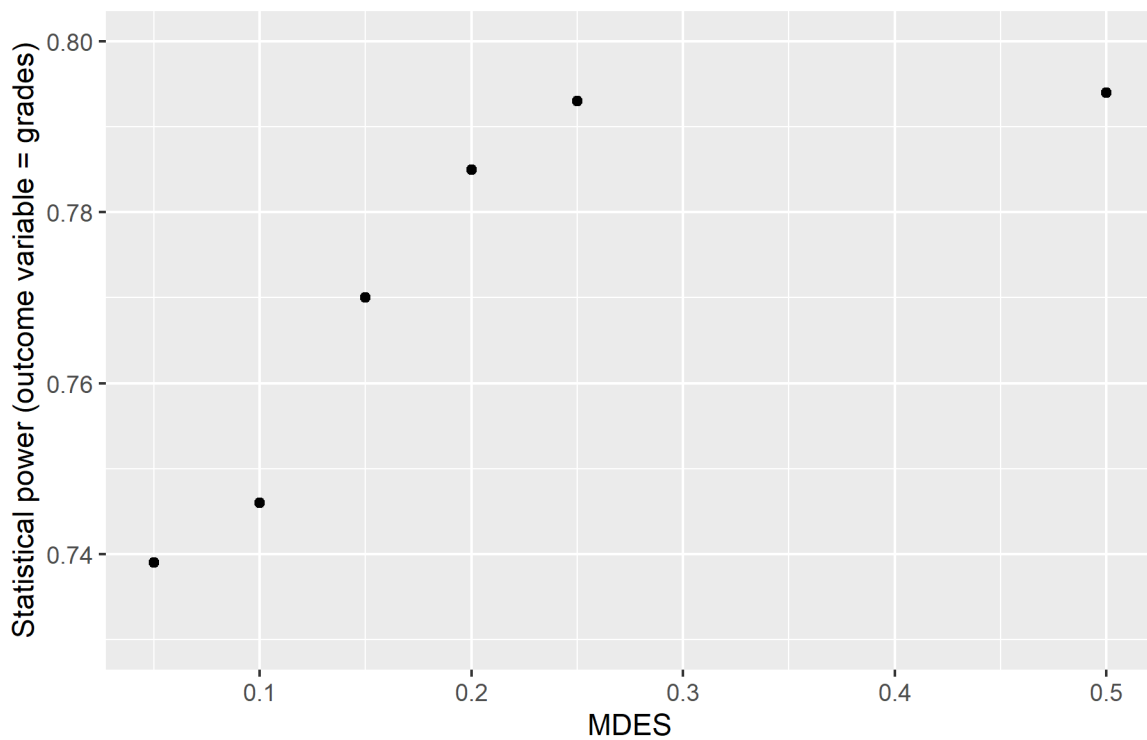


Figure 2: Loss in statistical power when grades are the outcome – English Language

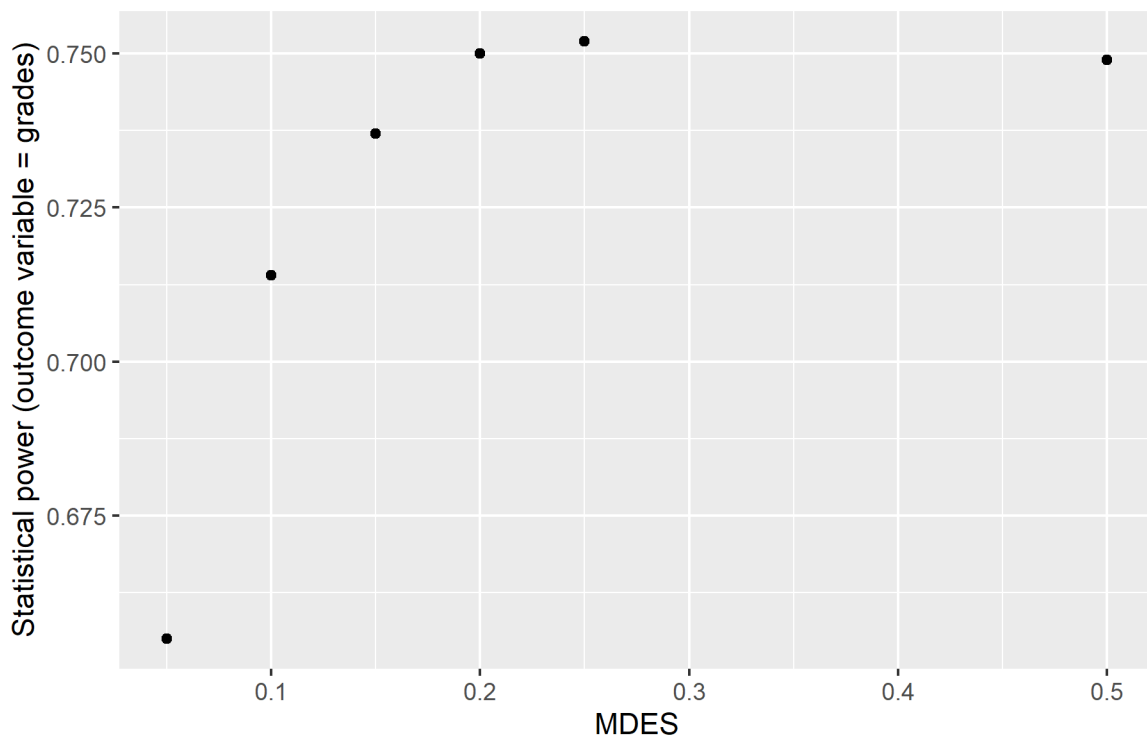


Figure 3: Loss in statistical power when grades are the outcome – Mathematics

Figure 3 shows that for Mathematics, the loss in power plateaus at around 0.755. This means that unlike English Language, where Figure 2 shows that for higher MDES values the plateau appears to be at around 0.793

(indicating little difference between using marks or grades as the outcome), for Mathematics there is always at least a 4.5 per cent reduction in statistical power when using grades.

Why is there a difference in the magnitude of the observed loss in sensitivity across English Language and Mathematics?

This is likely attributable to the differing characteristics of the Mathematics mark distributions, relative to those for English. English Language qualifications are consistently scored out of a total of 160, or 200 for one AO. Mathematics, meanwhile, is scored out of 240, or 300 for one AO. Furthermore, English Language is un-tiered, and thus includes nine grade boundaries; Mathematics qualifications only include five or seven depending on tier (for Foundation and Higher respectively). The knock-on effect of this is that in English Language, grade boundaries are substantially closer together than in Mathematics; on average a mere 14.4 marks apart whilst Mathematics sees an average of 30.2 or 37.1 marks between boundaries (for Foundation and Higher tiers respectively). This means that the grade awarded in English is more sensitive to changes in the underlying mark distribution. As has been noted, grades are a coarser measure of performance than marks; but it is important to note that in English Language, marks are substantially more coarse than in Mathematics, due to Mathematics' substantially higher mark tariff. In effect, each grade in Mathematics is split into around 30-40 increments, whilst in English Language each may only be split into 15 increments.

We can use the SD values in [Table 2](#) through [Table 5](#) as a starting point to explain how this impacts differently upon the results for the two subjects. When considering grades ([Table 4](#) and [Table 5](#)), we can see that across all AOs, the SD of each grade distribution is around 2 grades in width for both subjects. When considering marks ([Table 2](#) and [Table 3](#)), the SD of the mark distribution is around 23-30 marks wide for English Language, and 43-55 or 59 (depending on tier) marks wide for Mathematics (reflecting the higher number of marks available for Mathematics qualifications). Knowing this, we can assess the extent to which these two standardised spread metrics "match up".

To use an example above to explain this, we can rephrase this step as a question: are the grade boundaries on English Language generally spread such that 23-30 marks equates to around two whole grades (i.e. the SD of the grade distribution)? If the SDs in each metric do line up at two grades fairly well, then there is not likely to be much impact of using either metric as an outcome measure. However, if (returning to the above example) the SD in marks for English Language is closer to a single grade boundary in size, then marks is a more sensitive metric: if a trial can detect a difference between groups one SD in magnitude, one grade difference would achieve 0.8 power if marks were the outcome measure, versus a difference of two grades if grades were (per the SD of grades for this subject being ~2).

Does this hypothetical difference in SDs bear out in the data?

We can double-check whether this may be the case in the GCSEs we analyse. Considering the average grade widths outlined above, the SD of marks for English Language equates to 1.6-2.1 grades on average. For Mathematics, the SD of marks is equivalent to only 1.5-2.0 grades for Higher tier, so fairly similar to English, but for Foundation tier, the SD of marks equates to a lower 1.2-1.5 grades, on average. This means that when considering both tiers together, the SD of marks equates to fewer grades for Mathematics than for English Language. Because SD is instrumental in determining DDEs, the knock-on effect of this is that an effect X SDs (of marks) in size for Mathematics results in a smaller change in the commensurate grade distribution than an effect the same X SDs (of marks) in size for English Language would.

It is worth noting that the above discussion is a slight simplification; grade boundaries are not uniformly wide within a qualification, nor are candidates uniformly spread amongst the grades. Often, the majority of candidates will achieve a limited range of grades (the central grades, typically), meaning that the width of these 'more-used' grades is more impactful to the consideration of how many grades the SD in marks equates to. An example of one AO's GCSE English mark distribution is presented below to help visualise this.

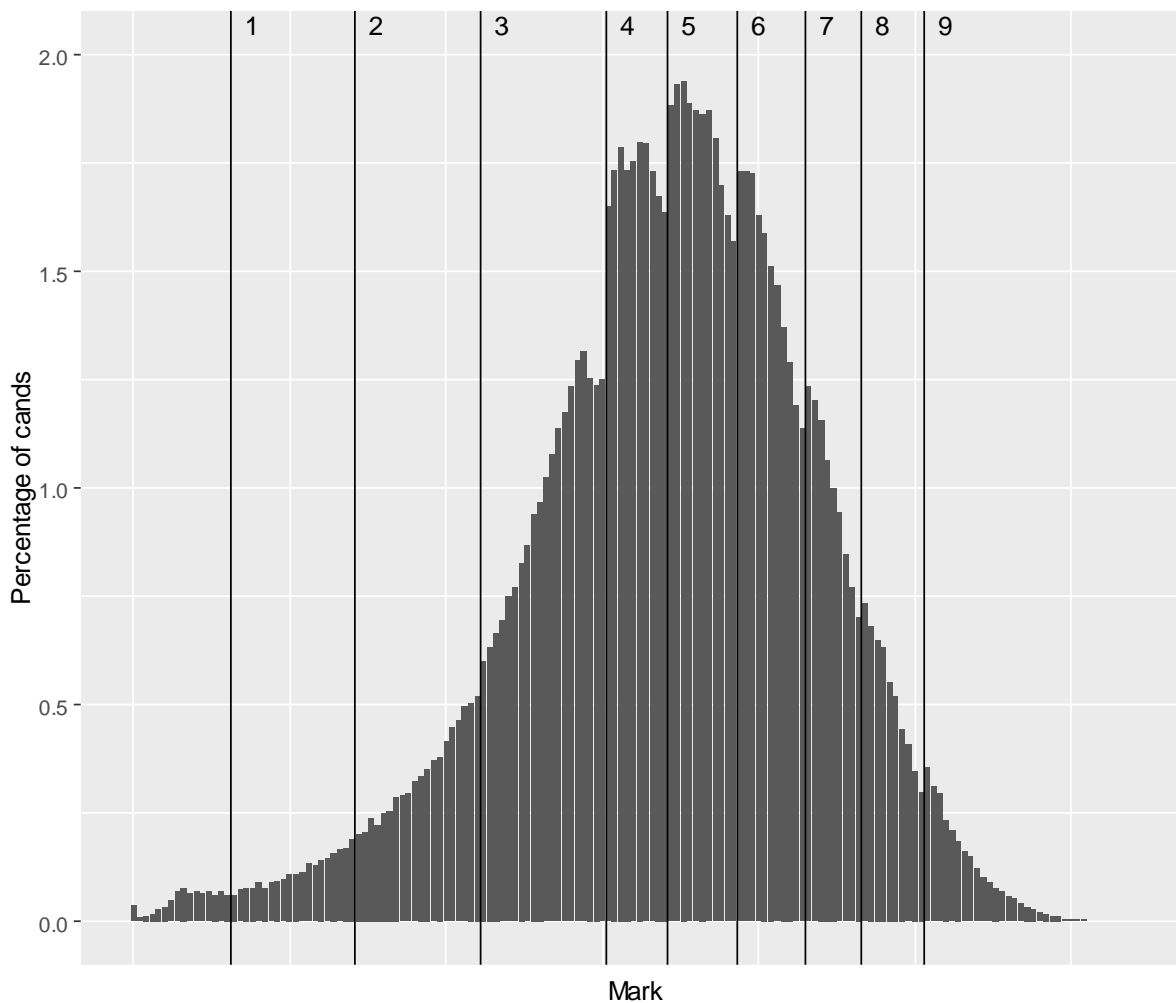


Figure 4: Example English Language mark distribution (with overlaid grade boundaries)

In the data visualised here, the lower less-used grades tend to be up to twice as wide as the higher grades, which will have the effect of reducing the apparent difference between it and Mathematics; for example, if we only considered grades 4-9 (which 71 per cent of candidates achieve), the average grade width for English Language is 11.7, and the commensurate grades the SD of marks equates to is markedly higher, at 2.0-2.6. Regardless of this, our conclusions about the root cause of the difference between the findings for Mathematics and English Language hold, and if anything are only strengthened.

This discussion highlights the crux of the difference observed in our results for the two subjects; ultimately, the root cause of the difference is that in English Language, one SD of marks equates to more grades than one SD of marks does in Mathematics. This in turn is attributable to the shape of the English mark distribution and how widely spread its grade boundaries are when compared with Mathematics. Indeed, Mathematics' more widely spread distribution with fewer, further apart boundaries will intuitively see much higher power when marks rather than grades are the outcome variable. The increased coarseness of marks is also a factor (and is likely to explain, at least to some degree, the slightly increased irregularity of our findings for English Language when graphed in [Figure 2](#)), but importantly it is not the root cause.

Limitations

[Figure 4](#) highlights one interesting feature of the data used for this analysis. Instead of a smooth mark distribution, there are distinct 'troughs' just below each grade boundary. This occurs because, following the initial marking of examination papers and delivery of results to candidates, an appeals process takes place. Since an unsuccessful appeal carries a financial cost for the school, the vast majority of appeals are entered for candidates just below a grade boundary, in the hope that the appeals process might result in the candidate gaining a handful of marks

and slipping over the boundary. Whilst not all appeals succeed in this and candidates' marks can in theory go down as well as up, around 70 per cent of appeals that change candidates' marks result in an increase in marks (Ofqual, 2019).

The net result of these 'troughs' is that in the data we use for our simulations, there are fewer candidates just under a boundary who (if selected into the treatment group) could increase their grade by gaining marks equal to the DDE – relative to the number of candidates just under the boundary were the distribution were more smooth/normal (as it is pre-appeals). This will have a greater impact on lower MDES scenarios, because almost all of the candidates who *could* gain a grade if selected to the treatment group and their mark is increased (by the DDEs associated with these MDES values) will fall in the 'trough' in the distribution where there are fewer candidates than we might expect (because many have already improved their grade via appeals).

Whilst this is a limitation of our results, it is important to consider that for the larger MDES scenarios we modelled, the DDE mark gain implemented for the treatment group approaches (or in some cases surpasses) the average grade width for the qualification. For example, in the 0.5 MDES scenario modelled to provide an additional data point, the average DDE for English language is 13.1, whilst the average grade width is 14.4. This means that almost every single candidate in the treatment group will gain a grade when the DDE is added to their actual achieved mark, and thus the impact of appeals is broadly negated.

To summarise, we can therefore conclude that our results for higher MDES values provide an estimate of the lower limit of the loss in power that is likely accurate, and is unaffected by the use of the post-appeals dataset. However, our results for lower MDES values are potentially inflated by the use of the post-appeals mark distribution, and are thus somewhat less reliable. As such, we focus in the remaining discussion primarily on the higher MDES 'lower limit' loss in power.

The other potentially key limitation is that the data available did not permit us to conduct clustered simulations more representative of a majority of EEF trials' designs. However, as detailed in Annex 1, we do not consider that this is a significant concern. To summarise the Annex here, this is because the two additional variables which impact on the determination of sample sizes and MDESeS in cluster randomisation designs are the average cluster size and intra class correlation coefficient. In short, there is no reason why a shift from marks to grades as the outcome variable would systematically impact either of these two variables. As such, whilst the absolute value of some of our findings would undoubtedly be different under a cluster randomisation design, our conclusions would still stand as the same trend should emerge. The equations presented in Annex 1 permit the reader to check this on any of our key results; a worked example is also provided to verify that this is indeed the case.

One other limitation is that the data available makes no distinction between FSM-eligible and not-FSM eligible students. As mentioned in the Background section, one of EEF's key objectives is closing the attainment gap, and as such evaluations commonly conduct additional analyses investigating the extent to which interventions help to close this gap. Because FSM data was not available, it was not possible to simulate these analyses and thus to quantify the impact of using marks vs grades on the ability to detect a closing of the attainment gap.

To briefly summarise other minor limitations; simulations are never entirely accurate, but we believe that 5,000 iterations of each scenario is sufficient to adequately minimise error in our results. One AO's Mathematics qualification did not have sufficient volume of data to support our sampling of candidates in the simulations; therefore our results for Mathematics are not completely representative of the population, but because the cohort was so small, our results for this subject are still broadly accurate.

Implications

Above we have detailed the key findings and their limitations. With this done, it is necessary to consider the implications of the findings for EEF trials and evaluators. Because of the difference in results for the two GCSEs, we discuss each in turn.

For English Language, it appears to be the case that there is in general only a negligible loss in statistical power when grades rather than marks are used as the outcome variable. This is because the SD of the mark distribution converted into grades is very similar to the SD of the grade distribution, meaning a standardised effect in marks

equates to more or less the same change in grades in most situations. Even in the worst case, for the lowest MDES considered (0.05), the loss in statistical power is around 6.2 per cent; i.e. a power of around 0.74 rather than 0.8. In terms of sample size, this equates to an increase in sample sizes of around 18 per cent in order to maintain power at 80 per cent.

For Mathematics the choice of marks or grades is of greater consequence. In the best case, power falls to 0.76 from 0.8. However, even here, the sample size must rise two and half times in order to maintain power of 80 per cent. This is because when the SD of the mark distribution is converted into grades, it is somewhat lower than the SD of the grade distribution, meaning a given standardised effect in marks equates to a smaller change in grades, and thus that marks are a more sensitive metric. In the worst case, sample power can fall by around 15 percentage points (from 80 per cent to around 65 per cent) and sample sizes three times the size are required in order to maintain power at 80 per cent.

Even considering only the 'lower limit' of results for GCSE Mathematics, the increased sample size needed to detect the same 'effect' in particular is problematic for evaluations. Recruiting 2.5x as many candidates or schools carries substantial cost (and time, and complexity) implications, and is likely to make many trials prohibitively expensive.

If offering advice to evaluators whose prospective trial designs use GCSE Mathematics as an outcome variable, bluntly the question to be considered is "is the cost of collecting mark information from schools or AOs (as opposed to accessing their grade information from the NPD) likely to increase costs more than recruiting a sample 2.5x as large would?" This is the key trade-off to be considered. We would suggest that, given the increased challenges associated with accessing NPD data given DfE's tightened data security policy in recent years, going to the effort of collecting mark information may be a more attractive option than it has been historically. Logistically speaking, there are two options here; either ask schools to provide this information based on their records (as total marks should be reported along with grades on certificates) or make a direct request to the AOs with school consent. We would suggest that whilst the latter is likely to result in higher quality data with less missingness, the challenges associated with getting AO sign-off on both data provision and data sharing approach are significant, so asking schools to provide mark information may be more practical. Another option (albeit one which requires systemic change) would be for the DfE to begin collecting mark information from AOs in addition to grades, and to store this in the NPD or at least make it available for research purposes in the ONS SRS.

Whilst our analysis focuses solely on GCSE Mathematics and GCSE English in a single year, we believe that our conclusions can be generalised to other graded qualifications, and to other years¹¹. The crux of our findings about loss of power if grades are the outcome variable is that this hinges on how many grades an SD of marks equates to; if this value is close to the SD of the grade distribution, then there will be no loss of power. If this value is higher than the SD of the grade distribution, then the same standardised effect size is 'worth more grades' when measured in marks for this qualification, and there will be a loss in power when grades are the outcome variable (conversely, if it is lower than the grade distribution's SD, there will be a gain in power when grades are the outcome variable).

Considering other English General Qualifications, whilst SDs of grade distributions can be derived from the freely available JCQ results statistics for GCSEs and A-levels, the same is not true of the SD of mark distributions. However, given the freely available grade boundaries and maximum mark information AOs make available, it is possible to approximate the SD of marks and draw inferences about whether a qualification is likely to see a loss in power if marks are the outcome variable, using the methodology we apply to derive the necessary sample sizes associated with various simulated scenarios. This approach should allow any evaluator to make an informed guess as to whether they should seriously consider using marks (as opposed to grades) as their outcome variable.

¹¹ Whilst our findings represent a single snapshot of the effect of using marks vs grades for one year's GCSE papers, the use of data from all AOs serves as an example of the variation in results which results from slightly different mark and grade distributions. In none of the tables of results above are there particularly striking differences between the AOs, so it is reasonable to assume that our results at the whole cohort level would only be slightly different if another year's data had been used, and that our overall conclusions would still stand.

Finally, it is important to consider these findings in a broader context. As discussed in the background section, one of EEF's chief aims is to reduce the attainment gap, which is most frequently measured in terms of GCSE attainment. Interventions which go to trial are therefore vehicles by which this aim may be achieved. We believe the analysis conducted in this paper and in Smith et al (2020) highlights a potentially overlooked step in determining whether an intervention should go to trial and whether GCSEs are an appropriate outcome measure; there must be a coherent logic in place as to how the intervention will improve students' GCSE marks (not grades, as marks always underpin grades even if not the outcome variable) – and what the magnitude of this improvement could plausibly be. If an intervention lacks this logic or the plausible magnitude of improvement is minimal, then the obvious alternative would be to utilise a different outcome measure better tailored to detect the treatment enacted in the intervention.

For this step, it may be useful to consider the concept of the 'minimally important mark difference', which we can define as 'the size of effect an intervention would need to have to be worth the investment of trialling', and the 'plausible mark gains' a trial could reasonably result in for the treatment group. This helps to close the loop in trial design; evaluators currently pick a plausible MDES value, but without explicit knowledge of how great a shift in marks would be necessary to generate an effect that size.

In Mathematics, for instance, our analysis of mark distributions shows that an MDES of 0.25 requires a difference between group means of 10-15 marks (approximately 5 per cent of all the marks available for the qualification). Knowing this, an evaluator can assess whether this is a plausible average 'gain' for the treatment group to make due to the intervention. It is possible to support this decision by using reported trials (with GCSE marks as an outcome variable) as a barometer, to indicate the level of mark differences between the treatment and control groups that real interventions have manifested. Our feeling is that in many cases, interventions are likely to only result in a handful of marks gained (on average) and thus a small effect size. The key judgement is then whether the plausible mark gains the intervention could result in are larger than the 'minimally important mark difference' (in terms of effecting a meaningful change in grade), and thus whether the intervention is likely to 'break even' on the investment required to trial it.

Annex 1: Generalisability to clustered trials

Impact of the design effect

This paper simulates trials using individual-level randomisation. A key question readers are likely to have is whether our findings generalise to trials designed using cluster randomisation; that is a design that involves randomising groups of students, typically classes or whole schools, rather than individual pupils.

We believe that our findings would remain materially similar for such clustered designs. To explain why, we must consider the factors that impact on the design effect (DE; Campbell & Walters, 2014):

$$DE = 1 + (m - 1)\rho \dots \dots \dots [4]$$

Here ρ is the intraclass correlation coefficient and m is the average size of a cluster. For example were whole schools to be the unit of randomisation, m would be the average size of the relevant cohort within schools, assuming equal cluster sizes. The interclass correlation coefficient (ρ) would then be the proportion of the total outcome variance that is between schools.

Typically DE is used to adjust calculations performed on the basis of individual randomisation in order to adjust such calculations for the presence of clustering. So for sample size calculations typically performed in order to obtain an estimate of the required number of pupils n we see:

$$n_{cluster} = n_{individual} \times DE \dots \dots \dots [5]$$

This calculation is approximately correct in that it does not take account of the different degrees of freedom in cluster as opposed to individual level study designs, nonetheless it is a generally accepted solution. Likewise for the minimum detectable effect size:

$$MDES_{cluster} = M_{n-2} \sqrt{\frac{1}{nP(1-P)}} * \sqrt{DE} \dots \dots \dots [6]$$

DE can be therefore used to account for the impact of switching to a cluster design from an individual design using the above equations. Importantly, Equation 4 shows that DE is underpinned solely by m and ρ , and in for both of these there is no reason why switching from marks to grades as the outcome variable (or vice versa) would materially change their values. Cluster sizes (m) are clearly completely unrelated to outcome variable choice, and because grades are a summary of marks, there is no reason ρ should be affected either.

What this means is that for the analysis presented here, each trial design would have a fixed value for the design effect based on assumptions we would have to make about m and ρ . In other words, once we have made these assumptions any results we present for a given simulated sample design would only be affected by this single scaling factor (DE). There would be no non-linearities nor interaction effects between the elements of the DE and the disparities between marks and grades that are the focus of our attention.

As a result the substance of our findings, in terms of loss of power and the proportional increases in sample size associated with using grades as an outcome variable, are unaffected by the choice to conduct our simulations based on individual as opposed to cluster randomisation. Whilst the magnitude of some findings (such as the sample sizes involved) are affected due to the scaling factor DE applies, this does not materially change our conclusions and the implications thereof, because the same scaling is applied in both mark and grade based scenarios.

Worked example

To confirm the above, we present a worked example demonstrating that our key sample size finding persists. We use the example of Mathematics, where our key conclusion was that, across all AOs, using grades as the outcome variable entailed needing a sample size 2.5 to 3.2 times the size of the sample needed if using marks, depending on the MDES selected (Table 13).

Looking at the school census for 2018/19, the average size for KS4 GCSE cohorts across state secondary schools was about 110 pupils. At the protocol stage, most EEF trials assume, conservatively that $\rho = 0.20$. This

means that $DE = 22.8$ and its square root 4.77 . We can feed these values into Equation 5 to adjust the sample sizes underpinning Table 13 for a clustered rather than individual-level design.

The figure below shows, based on the individual-level sample sizes informing Table 13, the relationship between the sample sizes necessitated with marks and grades as the outcome variable.

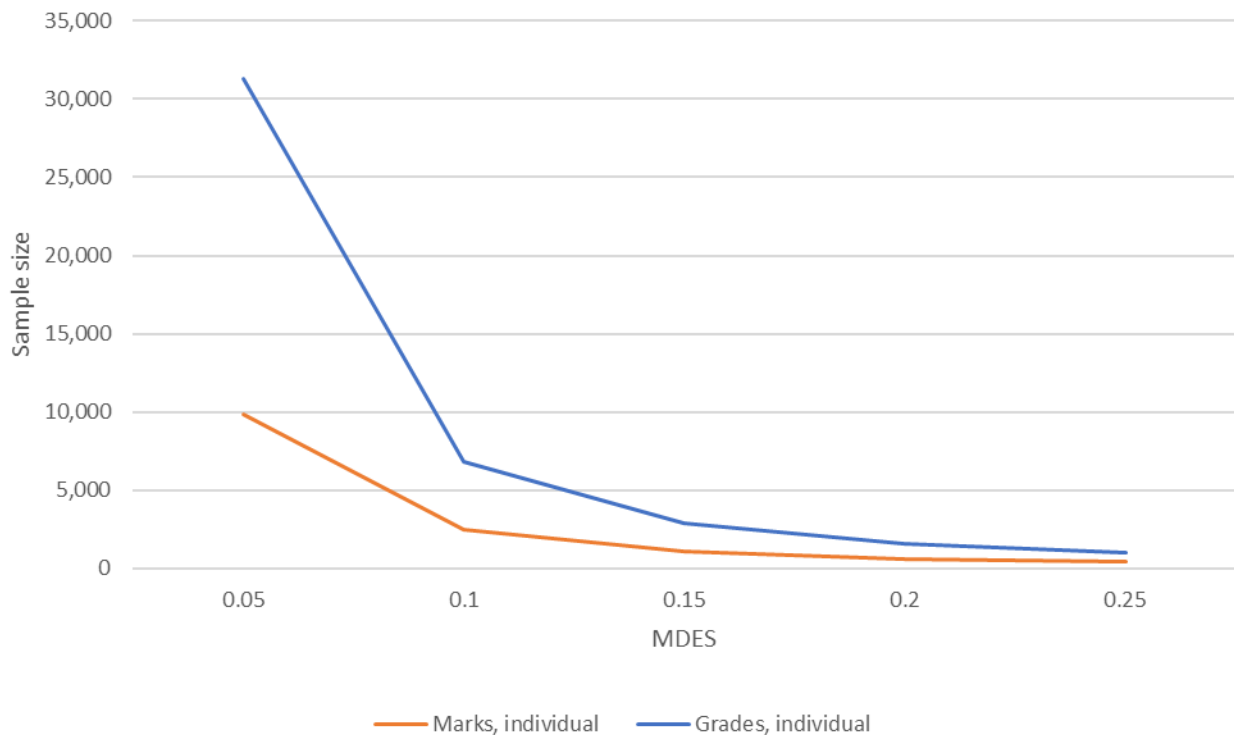


Figure 5: Necessary sample sizes – individual level design

The figure below shows the same relationship for the clustered design, based on having adjusted the sample sizes in Figure 5 according to Equation 5. Per the above, this essentially means multiplying them by 22.8 .

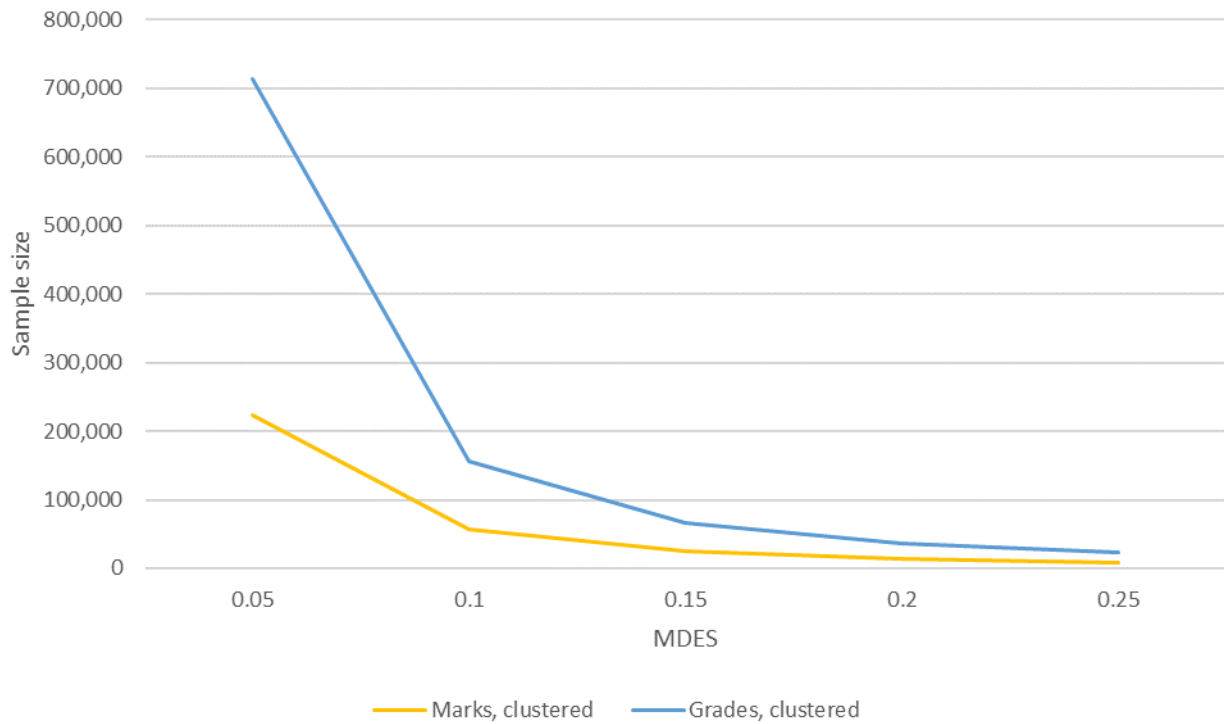


Figure 6: Necessary sample sizes – clustered design

The mark-grade trend shown in both figures is clearly extremely similar, if not identical. To verify this, the below table shows comparable ratios to [Table 13](#) for clustered designs.

Table 14: Ratio of sample sizes needed to detect an effect (marks:grades) – Mathematics, clustered design

MDES	AO 1	AO 2	AO 3	All AOs
0.05	2.912	3.368	3.352	3.228
0.1	2.495	2.841	2.884	2.773
0.15	2.412	2.738	2.685	2.616
0.2	2.381	2.689	2.653	2.585
0.25	2.368	2.678	2.560	2.512

The figures in the above table are extremely similar to those in [Table 13](#). They are not identical as the figures informing [Table 13](#) are integers (because one cannot sample a partial individual), so any rounding effects are compounded 22.8 times when the DE's impact is accounted for. This confirms the above suggestion that whilst the scale of sample sizes clustered trials require is always very different to trials randomising at the individual level, the substance of our conclusions is unaffected.

References

- Allen, R., Jerrim, J., Parameshwaran, M., & Thomson, D. (2018). *Properties of commercial tests in the EEF database*.
https://educationendowmentfoundation.org.uk/public/files/Publications/EEF_Research_Papers/Research_Paper_1_-_Properties_of_commercial_tests.pdf
- Baird, J.-A., Cresswell, M., & Newton, P. (2000, 1). Would the real gold standard please step forward? *Research Papers in Education*, 15(2), 213-229. doi:10.1080/026715200402506
- Bloom, H. S. (1995). Minimum detectable effects: A Simple Way to Report the Statistical Power of Experimental Designs. *Evaluation Review*, 19(5), 547–556.
- Bond, L. A. (1996). Norm- and Criterion-Referenced Testing. *Practical Assessment, Research, and Evaluation*, 5(2). doi:10.7275/dy7r-2x18
- 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>
- Boyle, A., & Mellor, D. (2020). *Implications of GCSE grade estimation on EEF projects*.
- Brennan, R. (2006). *Educational measurement*. Westport: Praeger Publishers Inc.
- Bulus, M., Dong, N., Kelcey, B., & Spybrook, J. (2019). PowerUpR: Power Analysis Tools for Multilevel Randomized. Retrieved from <https://CRAN.R-project.org/package=PowerUpR>
- 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. <https://doi.org/10.3102/0013189X16656615>
- Combined_science_grading. (n.d.).
- Connolly, P., Biggart, A., Miller, S., O'Hare, L., & Thurston, A. (2017). *Using randomised controlled trials in education*. British Educational Research Association/Sage Publications.
- Crawford, C., & Benton, T. (2017). *Volatility happens: Understanding variation in schools' GCSE results*. Retrieved from <http://www.cambridgeassessment.org.uk/>
- 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.
- Department for Education. (2018). *Free school meals Guidance for local authorities, maintained schools, academies and free schools*.
- 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>
- Edoald, T., & Nevill, C. (2020). Working Out What Works: The Case of the Education Endowment Foundation in England. *ECNU Review of Education*, 2096531120913039.
- Education Endowment Foundation. (2012). *EEF guidance on choosing and delivering attainment tests*.

https://educationendowmentfoundation.org.uk/public/files/Evaluation/Setting_up_an_Evaluation/EEF_testing_criteria_and_guidance_on_blinding.pdf

- Education Endowment Foundation. (2013). *Pre-testing in EEF evaluations*. https://educationendowmentfoundation.org.uk/public/files/Evaluation/Writing_a_Protocol_or_SAP/Pre-testing_paper.pdf
- GCSE_factsheet_for_employers_FE_and_HE_providers__final_. (n.d.).
- Gerber, Alan, S., & Green, Donald, P. (2012). *Field experiments: Design, analysis, and interpretation*. W. W. Norton & Company.
- Gill, T. (2014). *An investigation of the effect of early entry on overall GCSE performance, using a Propensity Score Matching method*. Retrieved from <http://register.ofqual.gov.uk/Qualification>
- Hedges, L. V. (2011). Effect sizes in three-level cluster-randomized experiments. *Journal of Educational and Behavioral Statistics*, 36(3), 346–380.
- Higgins, S., Katsipataki, M., Villanueva-Aguilera, A. B., Coleman, R., Henderson, P., Major, L. E., Coe, R., & Mason, D. (2016). *The Sutton Trust-Education Endowment Foundation Teaching and Learning Toolkit*.
- Lawley, D. (1950, 6). A METHOD OF STANDARDIZING GROUP-TESTS. *British Journal of Statistical Psychology*, 3(2), 86-89. doi:10.1111/j.2044-8317.1950.tb00286.x
- Jay, M. A., Mc Grath-Lone, L., & Gilbert, R. (2019). Data Resource: the National Pupil Database (NPD). *International Journal of Population Data Science*, 4(1).
- Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, 43(6), 304–316.
- Meadows, M., & Billington, L. (2005). *A REVIEW OF THE LITERATURE ON MARKING RELIABILITY*.
- Ministry of Housing Communities & Local Government. (2019). *The English Indices of Deprivation 2019*. Retrieved from <https://www.gov.uk/government/publications/english-indices-of-deprivation-2019-technical-report>
- Murnane, R. J., & Nelson, R. R. (2007). Improving the performance of the education sector: The valuable, challenging, and limited role of random assignment evaluations. *Economics of Innovation and New Technology*, 16(5), 307–322.
- 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. (1998). Experimentation and social interventions: a forgotten but important history. *BMJ: British Medical Journal*, 317(7167), 1239.
- 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.
- Ofqual. (2011). *Maintaining standards in GCSEs and A levels in summer 2011*.
- Ofqual. (2019). *Reviews of marking and moderation for GCSE and GCE: summer 2019 series*. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/851640/Reviews_of_marking_and_moderation_for_GCSE_and_GCE_summer_2019_series.pdf
- R Core Team. (2017). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from <https://www.R-project.org/>

- Reynolds, C., Livingston, R., Willson, V., & Willson, V. (2010). *Measurement and assessment in education*. Upper Saddle River: Pearson Education International.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin and Company.
- Slavin, R. E. (2008). Response to Comments: Evidence-Based Reform in Education: Which Evidence Counts? *Educational Researcher*, 37(1), 47–50. <https://doi.org/10.3102/0013189X08315082>
- Slavin, R. E. (2017). Evidence-based reform in education. *Journal of Education for Students Placed at Risk (JESPAR)*, 22(3), 178–184.
- 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., Boyle, A., & Morris, S. P. (2020). *GCSE science as an outcome measure: the capacity of the Deeper Thinking intervention to improve GCSE science grades*. https://educationendowmentfoundation.org.uk/public/files/GCSE_Science_paper.pdf
- Smithers, A. (n.d.). *GCSE 2021*.
- Standards and Testing Agency. (2020). *Understanding scaled scores at key stage 2*. Retrieved from <https://www.gov.uk/guidance/understanding-scaled-scores-at-key-stage-2>
- Taylor, R. (2013). *Centre for Education Research and Policy Early entry to GCSE*. Retrieved from www.cerp.org.uk
- Thorndike, R., & Thorndike-Christ, T. (2010). *Measurement and evaluation in psychology and education*. Upper Saddle River: Pearson.
- 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.
- Wheadon, C., & Stockford, I. (2008). *CLASSIFICATION ACCURACY AND CONSISTENCY IN GCSE AND A LEVEL EXAMINATIONS OFFERED BY THE ASSESSMENT AND QUALIFICATIONS ALLIANCE (AQA) Classification accuracy and consistency in GCSE and A level examinations*.
- Wilson, F., & Dhawan, V. (n.d.). *Capping of achievement at GCSE through tiering*. Retrieved from www.cambridgeassessment.org.uk
- Xiao, Z., Kasim, A., & Higgins, S. (2016). Same difference? Understanding variation in the estimation of effect sizes from educational trials. *International Journal of Educational Research*, 77, 1–14.

This work was produced using statistical data from the ONS. The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. This work uses research datasets which may not exactly reproduce National Statistics aggregates.

You may re-use this document/publication (not including logos) free of charge in any format or medium, under the terms of the Open Government Licence v3.0.

To view this licence, visit <https://nationalarchives.gov.uk/doc/open-government-licence/version/3> or email: psi@nationalarchives.gsi.gov.uk

Where we have identified any third-party copyright information you will need to obtain permission from the copyright holders concerned. The views expressed in this report are the authors' and do not necessarily reflect those of the Department for Education.



Education
Endowment
Foundation

The Education Endowment Foundation
5th Floor, Millbank Tower
21–24 Millbank
London
SW1P 4QP

<https://educationendowmentfoundation.org.uk>

 [@EducEndowFoundn](https://twitter.com/EducEndowFoundn)

 Facebook.com/EducEndowFoundn