BRILL

IMPLEMENTATION AND REPLICATION STUDIES IN MATHEMATICS EDUCATION 4 (2024) 83–124

IRME

brill.com/irme

Educational Performance over Time: Changes in Mathematical Attainment between 1976 and 2009 in England

Jeremy Hodgen | ORCID: 0000-0002-9196-4088
UCL Institute of Education, University College London,
20 Bedford Way, London WC1H 0AL, UK
Corresponding author, e-mail: jeremy.hodgen@ucl.ac.uk

Robert Coe
Evidence Based Education, 1 Grange Crescent, Sunderland SR2 7BN, UK
robert.Coe@cem.dur.ac.uk

Margaret Brown
School of Education, Communication & Society, King's College London,
Waterloo Bridge Wing, Franklin-Wilkins Building, Waterloo Road,
London SE1 9NH, UK
margaret.brown@kcl.ac.uk

Dietmar Küchemann
School of Education, Communication & Society, King's College London,
Waterloo Bridge Wing, Franklin-Wilkins Building, Waterloo Road,
London SE1 9NH, UK
dietmar.kuchemann@kcl.ac.uk

Received 9 December 2023 | Accepted 27 March 2024 |
Published online 3 May 2024

Abstract

Over the past fifty years, there have been substantial attempts to improve students' mathematical performance around the world. Many commentators have criticised the efficacy of these initiatives, arguing that performance in western developed countries has either stagnated or fallen. Yet, there is limited robust comparative evidence available. This paper reports a replication of a study of student performance from the 1970s.

Published with license by Koninklijke Brill BV | DOI:10.1163/26670127-BJA10018

In 2008 and 2009, in England, Grades 6–8 students ($N \approx 7000$), in a nationally representative sample based on a stratified random sample of schools, were tested on their understandings of algebra, decimals, ratio and fractions. The survey used tests administered in 1976 and 1977 to an equivalent nationally representative sample of students. The findings indicate that, at Grade 8, overall understandings have generally fallen, although there are different patterns of change across the topics. The challenges of replicating studies where the full statistical findings are not available are considered.

The impact sheet to this article can be accessed at 10.6084/m9.figshare.25507276.

Keywords

educational reform – educational standards – mathematical attainment – replication

1 Introduction

Educational performance over time is a perennial concern of politicians, policy-makers and other commentators across the world. Some have claimed that performance in the US, the UK and elsewhere in the developed world has stagnated or even fallen since the mid-1970s (e.g., Goldin & Katz, 2008; Hanushek et al., 2010) and have pointed to the comparative success of the Pacific Rim and others in international surveys (e.g., Oates, 2011). These claims are hotly contested (e.g., Kilpatrick, 2011; Openshaw & Walshaw, 2010). So, for example, increases in school leaving or graduation qualifications attained are cited to support both the case that schooling is becoming ever more successful and also the opposing case that schooling is becoming less successful because the 'standards' of such qualifications are 'falling'. Yet, the evidence on which these claims about changes in educational performance are based is weak, because only limited data comparing educational performance extends back before the mid-1990s. Comparisons based on the international surveys, TIMSS and PISA, are valid for only a relatively short period back to the 'anchor' years of 1995 and 2003 respectively (PISA mathematics scores are comparable to 2003, although reading and science scores are comparable to 2000 and 2006, respectively). In the absence of reliable data, Goldin and Katz (2008), for example, base their analysis largely on years of schooling, whilst others analyse IQ test data over time (Nuthall, 1985) or compare attainment on different tests (Afrassa & Keeves, 1999; Rashid & Brooks, 2010). The Long-Term Trend National Assessment of Educational Progress (NAEP-LTT) in the US, which dates back to the 1970s, is a notable exception (Kloosterman, 2010), although even these data are problematic, because the test focuses on a rather narrow range of topics (consisting in mathematics of largely computationally based items).

This debate is important, because it bears on the success or otherwise of educational reform historically, and thus can inform future policy. We contend that England provides a critical case that is of interest internationally. Since the 1970s, England has, in common with many other countries around the world, implemented a range of educational reforms aimed at raising educational (including mathematical) attainment for all students (Cockcroft, 1982; Hodgen et al., 2022). Indeed, many systems have looked to England as an inspiration or as a model for educational change. One measure of whether this extensive reform and investment has been successful is the extent to which it has improved educational attainment. The focus of this paper is on changes over time in attainment in mathematics in the context of major investment in reform in this subject.

In this paper, we report on a replication of several surveys of student attainment first carried out in England in the 1970s. Specifically, we report the findings of a study of lower secondary students' mathematical attainment in England in which the results of a national sample of students tested in 2008 and 2009 are compared to the results of a similar national sample of students tested in 1976 and 1977 using the same research-based tests of conceptual understanding in mathematics (Hart et al., 1981, 1985).

As such, this study is both a *constructive* replication, investigating the phenomena at a different time, and a *conceptual* replication, utilising different statistical methods (Melhuish, 2018). As such, this replication enables us to examine whether English students' mathematical understandings have changed over the period and, thus, to consider whether the various reform of mathematics education in England have been successful in raising attainment. The replication also enables us to investigate the original findings on students' misunderstandings and misconceptions, although these findings are reported elsewhere (e.g., Færch & Hodgen, 2023; Hodgen et al., 2012). In conducting this replication we re-validated the tests, using modern statistical methods that utilise the computing power now available. Specifically, we used Rasch modelling, now widely used in assessment (Bond & Fox, 2007; Cascella et al., 2023; Gilberti & Maffia, 2022), but which was not well-developed at the time of the original study. We also have the benefit of a research team consisting of both researchers who were not involved in the original study (the first two authors: Hodgen and Coe), as well as two who were involved in the original study (the third and fourth authors, Brown and Küchemann), thus enabling us to combine the strengths of internal and external replications (Aguilar, 2020).

Our focus is on two areas that are fundamental to both mathematics itself and the use of mathematics in other disciplines: algebraic and multiplicative reasoning. We report on how students' performance on items relating to the topics of algebra, decimals, ratio and fractions has changed, and also on the variations in performance across the attainment range. Our findings relate to students in lower secondary education, who in England are aged between 11 and 14, with a particular focus on comparing the performance of the oldest group of students (Grade 8, aged 13–14).

2 Educational Reforms in England

In England over the period between 1976 and 2009, there were various large-scale national initiatives directed at improving teaching and raising attainment and focused on mathematics either exclusively or as a core subject (Ball, 2013; Brown, 1996, 1999, 2011; Hodgen et al., 2022; Majewska et al., 2022). These include a National Curriculum in 1989 (with five major revisions since then), the introduction of national testing at ages 7 (1991), 11 (1995) and 14 (1998), the Primary and Secondary National Strategies (1998–2010) and school league tables (1996), the National Centre for Excellence in Teaching Mathematics (2006) and significant investment aimed at improving pedagogy, qualifications and school infrastructure, with spending on education more than doubled in real terms between 1977 and 2008 (Chowdry & Sibieta, 2011).[1] In addition, more generic reforms aimed at improving attainment generally have included the introduction of a 'universal' school-leaving examination at age 16, the General Certificate of Secondary Education (GCSE) (1988), regular high-stakes school inspection by the Office for Standards in Education (Ofsted) (1992), and Making Good Progress (2007), which was focused partly on encouraging greater use of formative assessment. Many of these initiatives were extremely wide-ranging and ambitious; Fullan (2000), for example, describes the National Strategies as internationally "the most ambitious large-scale strategy of reform witnessed since the 1960s" and as having "without question the most explicit and comprehensive *implementation-based* strategy" (p. 19, emphasis in original).

These initiatives had a variety of different, and sometimes conflicting, emphases. Indeed, some commentators refer to the policy context in England in terms of 'overload' (Ball et al., 2012). Nevertheless, given the explicit aims to

[1] Reform in mathematics education in England has continued at a pace since 2009, including the introduction of the 'Mastery' initiative (Boylan et al., 2018), although this is beyond the period of the current study.

raise attainment together with the level of resources committed and "pressure and support" mechanisms (Fullan, 2000, p. 15), one might expect that the combined effect of these reforms would be to produce at least a modest increase in educational performance in mathematics.

3 Mathematical Performance over Time

During this 30-year period of educational reform, results in national examinations of mathematics in England have shown steady and substantial rises. For example the proportion of 11-year-olds achieving the targeted level in national tests rose from 54% in 1996 to 79% in 2009. Similarly, the proportion of 16-year-olds achieving grade C or above at GCE O-level/GCSE rose from about 23% in the mid-1980s to 45% in 1992 to 55% in 2007.[2] Some commentators (e.g., Barber & Mourshed, 2007) have claimed this represents considerable success for educational reform in England.

It is notoriously difficult to compare performance over time (e.g., Goldstein & Heath, 2000). One of the particular problems with using national tests and GCSE examinations to compare performance is that all papers are released, so new tests have to be used each year. This makes it difficult to maintain standards over time, and there is some evidence that the standard represented by the award of the same grade or level in these examinations for successive years has decreased (Coe & Tymms, 2008; Jones et al., 2016; Tymms, 2004). For example, Coe (2008) shows that between 1996 and 2007 performance in mathematics GCSE for students with equivalent ability scores rose by 0.9 of a grade, which was a larger increase than for any other mainstream subject. Prior to this period, use of a grade criteria system by the Graded Assessment in Mathematics (GAIM) project suggested that there was an average rise of approximately one grade in 1988 (the first year of the new GCSE) in comparison with the previous year's results on the earlier system, so that students who would previously have been awarded a grade D would now receive a grade C (Brown, 1989, 1996). Taken together, these results suggest a shift of up to 2 grades in the standard required at the GCSE grade C boundary over 30 years.

The international surveys, the International Association for the Evaluation of Educational Achievement's TIMSS (Trends in International Mathematics

[2] The GCSE (General Certificate of Secondary Education) examination replaced a dual qualification system in 1988. GCSE is normally taken at age 16. Grade C at GCSE was intended to be equivalent to the same grade of the previous, more academic qualification, GCE O-level (General Certificate of Education).

and Science Survey) and the OECD's PISA (Programme for International Student Assessment), do have comparable data going back to 1995 for TIMSS and 2003 for PISA mathematics. In TIMSS, England's performance at Grade 8 rose from a mean of 498 in 1995 to 513 in 2007 (Mullis et al., 2012),[3] whereas in PISA, England's performance at age 15 showed a small decline between 2003 and 2012 (Wheater et al., 2013).[4] However, independent analyses of the surveys suggest that, after correcting for sample bias and changes to the date of test administration, the performance on PISA has probably been broadly stable over time (Brown et al., 2007; Jerrim, 2013). Before 1995, some notion of comparison can only be gained in relation to rank order against other countries. However, there is no clear evidence of significant improvement from the earlier tests run by the IEA in 1964 (FIMS) and 1982 (SIMS), in each of which England scored close to the international mean. Indeed, the OECD's 2012 survey of adult skills indicates that mathematical skills are lower for younger cohorts, suggesting a gradual decline over time (OECD, 2013).

In a review of the research evidence on educational attainment over time, Rashid and Brooks (2010) find limited comparative evidence of mathematical attainment over time, and the evidence that they report is over a relatively short timescale. For example, they report that the Assessment of Performance Unit (APU) National Monitoring Survey conducted between 1978 and 1987 indicated a small overall improvement in mathematics attainment between 1978 and 1982, but no overall change thereafter, although there were decreases for the topics of algebra and number.

Internationally, aside from TIMSS and PISA, one of the few longstanding rigorous national systems for monitoring mathematical performance is the NAEP-LTT. Although based in the US, this provides comparable data on the performance of students aged 9, 13 and 17 over the period 1978–2004, and shows statistically significant gains of 22, 17 and 7 points in the mean score, respectively (Kloosterman, 2010). We note, however, that the NAEP-LTT tested

[3] In 2011, England's performance at Grade 8 in TIMSS had fallen back to a mean of 507 (Mullis et al., 2012). We note that comparisons of educational performance over time, including the PISA and TIMSS surveys, have limitations. For example, in addition to contextual and demographic changes, methodological and sampling procedures change between administrations. Nevertheless, the survey teams go to great lengths to ensure that claims about comparability over time are sufficiently robust.

[4] In fact, England, as part of the United Kingdom, was excluded from the comparison tables in PISA 2003, because of a failure to meet the minimum sampling requirements. Hence, the OECD do not consider the comparison of scores over time for England to be sufficiently robust over this period to report.

a narrow range of procedural skills in contrast to the Main NAEP or the CSMS tests considered in this paper.[5]

In summary, the evidence on educational performance over time is limited, both in England and internationally, particularly prior to 1995. Since then, the evidence in England suggests a slight rise in performance at secondary, although these rises may be at least partly due to factors other than increases in genuine mathematical attainment or competence.

4 The Design and Content of the Tests

Increasing Competence and Confidence in Algebra and Multiplicative Structures (ICCAMS) was a 4½ year project in England and included a survey of 11–14-year-olds' understandings of algebra and multiplicative reasoning, and their attitudes to mathematics. This survey involved both cross-sectional and longitudinal samples. The ICCAMS study used three tests (Algebra, Decimals and Ratio) which were first developed and administered to nationally representative samples of English students in the 1970s as part of the Social-Science-Research-Council-funded Concepts in Secondary Mathematics and Science (CSMS) study.

4.1 *The Design of the CSMS Tests*

The development of the tests was in several stages (Hart & Johnson, 1983). First, an initial set of items was developed on the basis of a review of the research literature and an analysis of the then curriculum. Since there was not then a national curriculum in England, this analysis was conducted using the most commonly used textbooks and contemporary national attainment surveys; e.g., Tests of Attainment of Mathematics in Schools (TAMS) (Sumner, 1975). Second, items were trialed in 20–30 interviews with students across ages 11–14 and the mainstream ability range, and finally pilot versions of the tests were administered in one or two schools. During this whole process, some items were dropped, some added and others amended. The majority of items on the tests are short open-response tasks intended to capture students' approaches without disadvantaging students who have reading difficulties.

[5] The NAEP-LTT is designed to measure students' "knowledge of mathematical facts, ability to carry out computations using paper and pencil, knowledge of basic formulas such as those applied in geometric settings, and ability to apply mathematics to daily-living skills such as those involving time and money" (http://nces.ed.gov/nationsreportcard/ltt/moreabout.aspx#measure).

4.2 *The Operationalisation of "Understanding" in the CSMS Tests*

The focus of the CSMS tests is on what is often described as "conceptual" as opposed to "procedural" understanding (Hiebert, 1986); CSMS aimed to use "problems which were recognisably connected to the mathematics curriculum but which would require the child to use methods which were not obviously 'rules'" (Hart & Johnson, 1983, p. 2). "Excessive computation" was generally avoided in favour of non-routine "word problems where the numbers were simple" (p. 22). The CSMS tests deliberately did not attempt to cover all aspects of the secondary mathematics curriculum, but rather intended to "reveal the different strategies that children used (rather than just to determine whether an item was answered correctly, or was easy or difficult)" (Küchemann, 1981, p. 30).

4.2.1 Algebra

The Algebra test focuses on generalised arithmetic rather than algebraic structure (Küchemann, 1981). The test extended Collis's (1975) analysis of the different ways in which pronumerals can be interpreted, and items were devised to bring out the following six categories (Küchemann, 1981): Letter evaluated, Letter not used, Letter as object, Letter as specific unknown, Letter as generalised number, and Letter as variable. As an example, consider item 5c: "If $e + f = 8$, $e + f + g = \ldots$." Here the letter g has to be treated as at least a specific unknown which is operated upon: the item was designed to test whether students would readily 'accept the lack of closure' (Collis, 1975) of the expression $8 + g$, rather than give the closed response $8g$, or a numerical response such as 9, 12 or 15.

4.2.2 Decimals

This test[6] focuses on decimals as an aspect of rational number, including place value, and its full title is "Place value and decimals". Items were designed to assess meaning and structure "in the sense of understanding how [the place value system] works and how to apply it in appropriate situations both mathematical and drawn from real-life" (Brown, 1981, p. 214). Computational aspects were included, but the focus was on meaning rather than on methods of calculation. Meaning was defined as "how children conceptualise number operations, visually, verbally and symbolically, and how applications of them are recognised" (Brown, 1981, p. 75). These foci are illustrated with the

[6] The full title of this test is Numbers 2 (Place Value and Decimals), although, for convenience, we refer to it as the Decimals test. It is important to note that the Decimals test was not a test of all aspects of rational number. It was developed alongside other tests of rational number that focused more specifically on fractions, computation with fractions as well as the Ratio test. In addition, a further test focused on children's understanding of number operations in the context of whole numbers.

item: 19d, “The cost of 6.22 litres of petrol was £4.86. What would the price of one litre be?”, requires students to identify the correct calculation rather than to calculate the answer. Students need to reason that the context requires division (4.86 ÷ 6.22), moreover, a division where the divisor is larger than the dividend, and that this makes sense. The Decimals test was developed at a time when research on rational number was at a relatively early stage of development and the foci of meaning and structure reflected the state of research at that time. Research in the United States had concurrently identified seven inter-related sub-constructs to rational number (Kieren, 1976), a framework that was subsequently refined into a smaller number, four or five, inter-related sub-constructs: quotient, ratio, operator and measure (and/or part-whole relations) (Behr et al., 1992; Kieren, 1988; see also, Lamon, 2007). Although developed independently, test items were matched at the time to cover all the sub-constructs in Kieren’s (1976) early version of this later framework (Brown, 1981, p. 66).

4.2.3 Ratio

The Ratio Test focuses on the application of ratio-based thinking in increasingly complex situations, with a particular focus on whether students used multiplicative rather than additive approaches (Hart, 1980). The test, titled “Test R”, deliberately does not mention ratio, so that students are not cued to use taught techniques, and instead have to work out the relationships inherent in the situation. Complexity was operationalised in terms of the multipliers involved, the steps required and the context (Hart & Johnson, 1983, p. 23). The test development was influenced by Piaget’s ideas and two items, involving eels and similar rectangles respectively, were drawn from questions used in his research (Piaget et al., 1960). Another item was a version of the well known ‘Mr Short and Mr Tall’ task developed by Karplus and colleagues (Karplus & Petersen, 1970) that requires a conversion between two non-standard units of measurement. The correct response demands multiplicative reasoning (or at least rated addition, Carraher, 1986)[7] rather than a direct additive approach. In order to avoid “excessive computation”, emphasis was placed on relatively simple ratios like 3:2 and 5:2. These were judged to provide substantive evidence of proportional thinking whereas, importantly, simple integer ratios did not (Hart & Johnson, 1983, p. 19; Karplus et al., 1972). A few items involved more numerically ‘complex’ ratios such as 5:3 and the technical report suggests that the team considered such a ratio to be sufficiently complex to indicate

[7] This additive strategy for correctly solving ratio problems has been variously termed Rated addition (Carraher, 1986), Scalar Decomposition (Vergnaud, 1983) and a Build Up method (Küchemann, 1981).

near-complete understanding of ratio as a multiplicative relationship (Hart & Johnson, 1983, p. 31).

4.2.4 Fractions

Some additional items focusing on fractions were included at the end of the ICCAMS tests (14 in the Ratio test and one in the Algebra test). This enabled a broader assessment of multiplicative reasoning, and made the Ratio test a similar length to the other tests. These items were drawn from the two CSMS Fractions tests, which took a similar approach to that described above for place value and decimals, with a focus on the 'measurement' and the 'multiplicative' areas of quotient and operator.

4.2.5 The CSMS Hierarchies of Levels

In the original CSMS data analysis in the 1970s, items were selected empirically from each test to form a series of hierarchical levels of difficulty. For each level, groups of items were identified based on the strength of phi correlations between the items within the context of a Guttman scaling model. (The method is described in detail by Hart and Johnson, 1983.) These hierarchies used 30 out of the 51 items in the Algebra test, 39 out of the 72 items in the Decimals test, and 20 out of the 27 items in the Ratio test. There were different numbers of levels associated with each test, since the levels were derived empirically; Algebra and Ratio both had 4 levels and Decimals had 6. Students were judged to have been successful at a specific level if they had successfully answered two-thirds or more of the items at that level. Students who had not achieved two-thirds of Level 1 items were said to be 'at Level 0'. It was possible to broadly describe the type of mathematical understanding required for the items in each level in each topic, although these were not always neat descriptions, since the items and levels were assigned mostly on empirical rather than theoretical grounds. These hierarchies provide a basis for comparing the understandings of students across the attainment range in 1976–1977 and 2008–2009. The additional 15 items on fractions were drawn from items that appeared in the hierarchy of the CSMS Fractions test, which also had 4 levels. However, in the case of Fractions, only single item comparisons with the 1976 data are possible, since the full set of Fractions items was not used.

4.3 *The Relationship to the English School Mathematics Curriculum of the 1970s*

The original CSMS study was conducted prior to the introduction of a statutory National Curriculum in England. However, as described above, considerable efforts were made to ensure that the test items related to the curriculum at the time and that the items would engage students in the 1970s. The team's subsequent analysis (Hart & Johnson, 1983, pp. 30–33) shows that, within the overall focus of each test on particular aspects of understanding, the test content could be shown to be covered in the curriculum, if not always explicitly, and that a relatively large proportion of the curriculum was covered by the whole set of tests.

4.4 *The Relevance of the CSMS Tests to the 2008–2009 Mathematics Curriculum in England*

In the 30 years between the CSMS study and this replication, there were considerable changes to the school mathematics curriculum in England. The first National Curriculum introduced in 1988 was influenced by the results of the CSMS study, particularly in relation to progression within and between topics and the recognition of the frequent use of informal methods. As part of its introduction, several aspects of topics, and most notably formal symbolic algebra, were introduced in Grade 6, which was earlier than before. (The introduction and revision of the National Curriculum is described in Brown, 1996.) Around a decade later, alongside the second revision of the National Curriculum, two National Strategies were introduced (the primary National Numeracy Strategy in 1999, and the secondary Key Stage 3 Strategy in 2001), which, amongst other things, introduced a very much more detailed year-by-year teaching framework as a programme for teaching mathematics (DfEE, 2001). A further amended national curriculum was introduced for secondary schools in 2007, but with little change in relation to content. Our analysis of the two National Strategies documents indicates that the CSMS tests are still relevant to the curriculum. Indeed, compared to 1976–1977, the CSMS tests appear in some ways to be slightly closer to the curriculum in place during 2008–2009, since the framework references some specific items or contexts utilised in the CSMS tests. A more significant change is that in 2008–2009 greater emphasis is placed on measurement and computation with decimals, rather than with fractions, due to the increasing prevalence of calculators, computers and the metric system.

4.5 *Survey Methods*

The ICCAMS and the CSMS projects each administered the tests to nationally representative samples of English lower secondary students, across 2008 and 2009, and in 1976 and 1977, respectively.[8] At both times, the Algebra and Ratio tests were administered to students in Grades 7 and 8, and the Decimals test to

[8] The 2008–2009 dataset is publicly available via the UK Data Service: https://doi.org/10.5255/UKDA-SN-851382. Further details and statistics for the 1976–1977 dataset are available in Hart & Johnson (1983), although the full dataset is no longer available.

students in Grades 6, 7 and 8. The focus will therefore be on those age groups, and, more specifically, on Grade 8.[9] The sampling for each project is described below. It is clear that in order to achieve a reliable comparison over time, the two samples would have to be selected in similar ways.

4.6 *The 1976 and 1977 CSMS Samples*

Fifty-four schools from across England were involved in the whole testing programme. The schools were selected from among those where a teacher had volunteered the participation of their school, responding to requests from the research team, either during professional development events or following an article in the professional press. The sample was not formally stratified, although there was a deliberate balance between rural and urban schools, schools with different ranges of social class and ability in their intakes, and state-funded and independent sector (private) schools. For each test in each grade, a sub-sample of at least 6 schools was selected. Care was taken to ensure that the score distribution across the sample on a (then) widely used test, the Calvert Non-Verbal IQ Test (Calvert, 1958), matched the national norms. The Calvert test was administered only to students in Grade 7, the age group for which it was designed; at the time, there was no equivalent test available for other age groups, and it was checked in each school that there was no obvious reason why the IQ score distribution of other grades would differ significantly from that of the Grade 7 students.

In 1976, the Algebra and Ratio tests were administered to entire year groups at Grades 7 and 8 within a school. In 1977, the arrangements were similar, except that the Decimals and Fractions tests were administered to a randomly selected proportion of the students in each class, since most schools were organised into classes for mathematics by attainment level. Again, the achieved samples were checked for their matches to national IQ norms. No full record exists of the details of the matching, except in the case of the Decimals test in 1977, where the means for each of four age groups were all in the 99.0–101.5 range, with the standard deviations all in the 14.0–15.0 range (Calvert norms were 100 and 15, respectively); the shapes of the four distributions were found not to be significantly different ($p > 0.2$) from the standard normal curve, using the Kolmogorov-Smirnov goodness-of-fit test (Brown, 1981, pp. 237–230).

[9] In England, Grade 6, 7 and 8 are known as Years 7, 8 and 9 (of ages 11–12, 12–13 and 13–14 respectively). These Year Groups, or Grades, are collectively referred to as Key Stage 3 and constitute the first three years of secondary school for most students. In this paper, given the international audience, we refer to Grades rather than Year Groups. Grade cohorts in England correspond well to age cohorts, because very few English students repeat Grades.

TABLE 1 The samples in 1976–1977 and in 2008–2009

	Students (and schools)		
	Grade 6	Grade 7	Grade 8
1976–1977			
Algebra (1976)	–	1128 (8)	961 (7)
Decimals (1977)	170 (5)	294 (8)	247 (7)
Test R (Ratio) (1976)	–	800 (8)	767 (6)
Fractions (1977)	246 (6)	309 (9)	308 (6)
2008–2009			
Algebra	1681 (19)	1810 (19)	1647 (18)
Decimals	1703 (19)	1784 (19)	1661 (18)
Test R [Ratio]	1677 (19)	1738 (19)	1595 (18)

It is believed that these values were typical of those for all the tests, especially since the Decimals test had the smallest sample sizes. Because of the different administrations, the sample sizes of students for the 1977 tests were smaller than for the 1976 tests, although the number of schools was similar in each case. The samples are shown in Table 1.

4.7 *The 2008–2009 ICCAMS Sample*

The aim was to draw the sample in a similar way, but since it was no longer possible to use the Calvert test as a control for the representativeness of the sample, it was decided instead to employ the MidYIS (Middle Years Information System), a value-added reporting system, which is widely used by schools across England (Tymms & Coe, 2003). The control test used to measure representativeness is a measure of developed ability, and consists of verbal (receptive vocabulary), numerical (everyday mathematics) and spatial (3D visualisation) problems.

The MidYIS system held a database of schools, so that it was possible to draw a random sample of schools. The intention had been to test a sample (20 schools) in the summer of 2008; however, owing to delays in the approval of funding for the project and higher than expected refusal from schools under time pressure, only 10 schools actually completed the tests (and one school did not test any Grade 8 students). A further round of testing was therefore conducted in the summer of 2009 to make up the sample.

4.7.1 The 2008 Sample

In order to obtain the right proportion of students from each sector, the sample was made up of two independent (private) schools and 18 government-funded schools.

The group of government-funded schools in the MidYIS database turned out to be a very close match to the group of all government-funded schools in England,[10] so a simple random sample of 18 was drawn. The characteristics of this sample were well matched to the population. For each sampled school, a reserve school was selected with matching characteristics in order to maintain the balance of the sample characteristics in the event of any school non-response. The use of a stratified sample was considered, using a range of variables to define strata, but the trial samples did not appear to produce a better fit to the population than a simple random sample.

The group of independent schools in the MidYIS database had slightly higher scores than the average for other independent schools (there were 264 independent schools with three years of MidYIS paper test data in the database). The sample was therefore limited to schools whose average MidYIS score over the three years was within one standard deviation of the mean for all schools, in order to ensure that the overall sample of 20 schools would be close to being nationally representative. Two schools were chosen at random from this subgroup. Given the much smaller numbers, reserve schools were simply chosen at random from the remaining eligible schools.

4.7.2 Selecting the 2009 Sample to Balance the 2008 Sample

A weighted random sample of schools was selected. Such a small sample has a lot of variability in sample mean. Therefore sampling was repeated until the mean MidYIS score was within 0.5 points of the desired value (the population standard deviation of MidYIS is 15 points). Reserve replacement schools were identified in the same way as for the 2008 sample.

4.7.3 The Achieved 2008–2009 Sample

Although 20 schools agreed to complete the surveys, only 19 actually managed to do so. Altogether, we approached 86 schools, which represents a school-level response rate of 22%, a value that is lower than might have been hoped for, but is within a typical range for studies of this kind (Coe & Hodgen, 2012, 2017c;

[10] The database was restricted to schools with three years of MidYIS test data in order to ensure that MidYIS data were available for all three Grades tested. There were 301 such schools out of a total of 1164 in the database, which is around a third of secondary schools in England.

Education Endowment Foundation, 2013). A range of student- and school-level characteristics were compared for the achieved sample of 19 responders, the 67 non-responders and the wider population.[11] Most differences were small and within chance variation for a sample of this size, though, overall, the achieved sample contained pupils from schools with slightly higher than average levels of Free School Meal[12] eligibility (18% vs 13% nationally), lower than average attainment (44% 5A*–C vs 50%), but above average value-added, both in mathematics and overall (Coe & Hodgen, 2012) (standardised effect sizes for the difference of 0.25 and 0.13, respectively). Nevertheless, there may be some bias due to the response rate of 22% in the 2008–2009 sample and, although we are able to show that there was no significant non-response bias in the school average MidYIS score for the students in the study, we did observe some differences between responders and non-responders in the attainment and progress of a previous cohort of students in those schools. A small difference in attainment (0.15 SD) suggests that the achieved sample might underestimate attainment in the national population, while a slightly larger difference in value-added (0.25 SD) points in the opposite direction.

In addition, we were able to match individual students' MidYIS scores with their ICCAMS scores. Correlations between MidYIS score and the total score on each ICCAMS test in each grade varied between 0.679 and 0.746, and all were statistically significant ($p < 0.001$). The strength of this relationship, combined with the availability of national norms for MidYIS scores, allowed us to increase the precision of sample-based estimates of population parameters by weighting the achieved responses to make their MidYIS scores fit the national distribution.

As the use of MidYIS scores was central to the approach used to obtain accurate estimates, we also investigated the sensitivity of the results to unmatched or missing MidYIS scores (Coe & Hodgen, 2017b). Overall, 9.5% of ICCAMS scores could not be matched and lower scores were more likely to be missing. Despite this, estimates using observed scores were within 0.02 of a standard

[11] The variables available for the population of all schools in England were: whether the school is single sex or mixed; whether the school is selective or not; whether the school is independent or maintained; total number of pupils in the school; school percentage FSM; school percentage achieving 5 + A* – C at GCSE (the equivalent of a Level 2 qualification, see Footnote 3 above); overall school value-added; mathematics value-added for the school. Full details of the comparisons can be found in Coe & Hodgen (2012).

[12] The proportion of students eligible for Free School Meals (FSM) is commonly used as a measure of deprivation in England. The data available to compare responding and non-responding schools relate to different cohorts of students from those who were involved in the study.

deviation of those derived from multiple imputation; using weighted MidYIS scores seemed to give small but appropriate corrections (Coe & Hodgen, 2017b).

4.8 *Test Administration*

The administration of the tests was the same for ICCAMS and CSMS and took place at the end of the school year in June and July. Each test was designed to be taken in one mathematics lesson and was administered by the students' regular mathematics teacher. In 2008–2009, to reduce test fatigue, each student completed just two of the tests on separate occasions. In 1976–1977, a sub-sample of students took two tests. Detailed instructions were provided, with only minor updating of language for the ICCAMS administration. In addition, whilst the tests were administered under examination conditions, teachers were encouraged to "ensure that all the students understand what the questions are asking of them … [but not to] give any information about how to tackle the questions" and to read the questions to students if required. In the 1970s, Algebra and Ratio, both shorter tests, were sometimes taken together in one lesson. Hence, in 2008–2009, students may have had a longer time to complete these tests than some students in the 1970s.

For the 2008–2009 sample, the performance of all three tests was analysed using both classical and item response theory (Rasch) models; full details are in Coe & Hodgen (2014). All tests performed well on dimensionality tests, and had high levels of internal consistency (e.g., Cronbach's alpha values: Algebra 0.95; Decimals 0.96; Ratio: 0.94). Almost all items provided an excellent fit to the Rasch model, with occasional misfit well below the level that would degrade measurement.

4.9 *Are the 2008–2009 and the 1976–1977 Samples Equivalent and Comparable?*

While, as far as possible, the 2008–2009 sample was constructed in the same way as that in 1976–1977, so as to enable valid comparisons of results, there were inevitably some differences. First, in order to improve precision, in 2008–2009, a larger sample of schools contributed to the results for each test in each grade, although the total number of schools involved in 1976–1977 was greater. In retrospect, a larger sample of schools for each test in each grade in 1976–1977 would have been preferable in order to reduce the extent to which the schools involved did not reflect schools nationally, but there was nothing that we could do to about that. Second, in 2008–2009 the schools were selected at random from the MidYIS database of schools in England, using the National Pupil Database (NPD) to establish the representativeness of the sample. In practice, however, the low response rate means that even a systematic sampling process does not guarantee that the 2008–2009 sample achieved is representative. In the 1970s, no equivalent database was available, so schools of different types and from different regions were asked to participate on a rather opportunistic basis. Third, the Calvert Non-Verbal Reasoning test originally used to establish the national representativeness of the sample for each test is no longer available and therefore, as already noted, an alternative, the MidYIS test, was substituted. It seems very unlikely that this change had anything but a very small effect.

Overall, therefore, it seems that in relation to the national distribution of IQ, the samples could be judged to be equivalent and comparable, and we know that there was a high correlation between these ability measures and scores on the mathematics tests. However, in relation to the effectiveness — or other characteristics — of the schools involved in the two samples, it is not possible to be completely confident whether either sample was nationally typical or whether they were strictly comparable. We know that schools in the 2008–2009 sample were slightly more effective than the national average in terms of their value-added progress in mathematics from grade 5 to 10, but also that their overall attainment in grade 10 was below the national average for England. In 1976–1977, many of the schools became part of the sample through a staff member volunteering at a professional development event, or through some other personal connection, so it is possible that these teachers were more confident, enthusiastic and perhaps more effective than typical. These are all limitations that we are unable to overcome and that should be borne in mind in interpreting any claims about national performance.

Nevertheless, we believe that some claims about national performance can still be made on the basis of these samples, for the following reasons. First, for all its limitations, both samples were the result of a systematic process to select a representative group and check its representativeness. Second, both samples used matching to a highly correlated, nationally standardised measure to limit the size of any variation from national norms. Third, no other longitudinal surveys exist, especially not on this scale, involving thousands of pupils at two time points. Ideally, our knowledge of changes in performance would be based on the interpretation of multiple and independent studies, each using different methodologies to give a balance of different strengths and weaknesses. Our study is far from the final word in such a process, but we hope it provides a start.

4.10 *The Estimation of Item Facilities and Confidence Intervals*

Bootstrapping was adopted as an approach to estimating the sampling error on 2008–2009 item facilities, after weighting to make the distribution of MidYIS scores in the sample for each test and grade nationally representative. Because we employed a two-stage sampling process (selecting schools, then pupils), estimation of standard errors must take account of possible clustering (the tendency for pupils in the same school to be more similar than pupils chosen independently). Although this can be done with standard statistical adjustments to the data from a single sample (e.g., using multilevel modelling or the Huber-White correction), the bootstrap approach is preferable. Our procedure for estimating item facilities was more complex: drawing a sample of schools, testing a sample of pupils in those schools, then applying weights to those test scores to achieve the same distribution in our sample of MidYIS scores as was known to be nationally representative. Part of the reason for using this weighting approach was to reduce the standard errors of our facility estimates: estimates from different samples chosen and weighted this way should be expected to vary less than they would if no weighting were applied. In the absence of an analytical way to calculate standard errors, the bootstrap approach allowed us to estimate the variation in facility parameters from repeated samples by simulating a process of repeated sampling and calculating that variation. For the Algebra test, 3000 bootstrap samples were generated for each grade in order to check the agreement across three different bootstrap methods: Standard 'Bootstrap-t' confidence interval; Simple percentile method; and BCa (bias corrected accelerated) percentile method (Efron & Tibshirani, 1993). In addition, the standard Bootstrap-t method was applied to item facilities without weighting based on MidYIS scores. Full details of the approach and results can be found in Coe & Hodgen (2017a).

The three methods that used MidYIS weightings were found to agree extremely well, with all inter-method correlations in excess of 0.99, and over 90% of pairwise comparisons within 10% of each other. Comparison with confidence intervals estimated without using weighting showed that weighting typically reduced the width of confidence intervals by 20%–30%, though for some items the reduction was much greater (Coe & Hodgen, 2017b). For the other tests, the simplest method (standard 'Bootstrap-t') was therefore used to estimate 95% confidence intervals.

The CSMS results were published in a lengthy technical report (Hart & Johnson, 1983), several doctoral theses (Brown, 1981; Hart, 1980; Küchemann, 1981) and a book (Hart et al., 1981). Although we had access to the detailed results from these sources, we did not have access to the full CSMS dataset. This meant that we could not reanalyse the data and, since standard errors and confidence intervals were not calculated for the original survey, we had to estimate these through a simulation process as described below. Hence, bootstrapping was also used to estimate confidence intervals for item facilities from the 1976–1977 round of testing, in the absence of any direct estimates from the original study. A bootstrap sample of six schools was taken to represent a typical sample, and the same method as that used in the 2008–2009 sample (Coe & Hodgen, 2017a) was applied. An estimate of the 95% confidence interval around the change in facility was calculated from the standard errors of measurement at each point.

5 Results: Changes to Students' Understanding over Time

In order to examine and compare overall performance between 1976–1977 and 2008–2009, we examine how item facilities as a whole have changed, together with the mean item facility in each topic. We begin by focusing on the comparison of the results over time for the oldest group of students tested in 2008–2009, those in Grade 8. We examine how the performance on the items in each topic has changed and then how performance has changed across the attainment range in the cohort by comparing the proportion of students at the different levels in the hierarchy of understanding. We then briefly consider how mathematical understanding has changed for younger students in order to consider the changes across lower secondary.

5.1 *Changes to Item Facilities at Grade 8*

In this section, we discuss the changes to the item facilities across all four topics — Algebra, Decimals, Ratio and Fractions — in order to examine overall how students' understanding has changed over time. In Table 2, we summarise these changes by using the mean facilities for each topic together with an overview of the numbers and percentages of items where the facilities have increased, decreased or not changed significantly. It can be clearly seen that the overall mean facilities have declined in all topics, with the decline smallest for Decimals and greatest for Fractions. The decline is statistically significant for all topics except Decimals, and the effect sizes range from $d = 0.18$ for Decimals to $d = -0.45$ for fractions. Over time, mean facilities on roughly half of the items have decreased significantly, and roughly half have not changed significantly. Only 5 (or 3%) of the total 163 items have mean facilities that have increased significantly, and all of these are from the Decimals test.

In Figures 1–4, the facilities of 2008–2009 are plotted against those of 1976–1977 and any significant changes are indicated. These scatterplots show that for Decimals, Ratio and Fractions, items that have declined significantly are spread across the range of item facilities.

TABLE 2 Summary of change to mean facility of items on each test at Grade 8

	Items (*n*)	1976–1977	2008–2009	Change					
		Mean proportion (SE)	Correct (SE)	Difference in proportion correct		Number (%) of items			Effect size (*d*)
				Mean (SE)	95% CI	Increasing	No change	Decreasing	
Algebra	51	48.8% (1.69%)	41.9% (0.98%)	−6.8% (1.96%)	[−11.5%, −2.3%]	0	24 (47%)	27 (53%)	−0.32
Decimals	73	60.0% (2.88%)	55.7% (1.33%)	−4.4% (3.17%)	[−10.1%, 1.4%]	5 (7%)	44 (60%)	24 (33%)	−0.18
Fractions	15	50.8% (2.13%)	39.3% (1.14%)	−11.5% (2.41%)	[−10.8%, −3.1%]	0	3 (20%)	12 (80%)	−0.45
Ratio	24	47.0% (1.92%)	40.0% (0.99%)	−6.9% (2.16%)	[−16.5%, −6.5%]	0	7 (29%)	17 (71%)	−0.29
Total	163					5 (3%)	78 (48%)	80 (49%)	

Effect size calculated using score change based on mean facility as a proportion of standard deviation in 2008–2009.

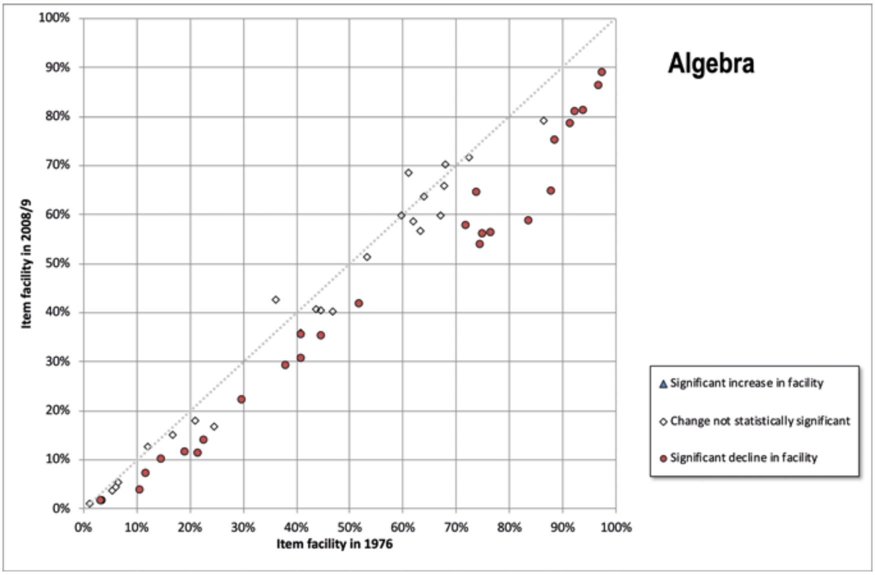


FIGURE 1 Scatterplot of 51 matched items facilities for Algebra test at Grade 8

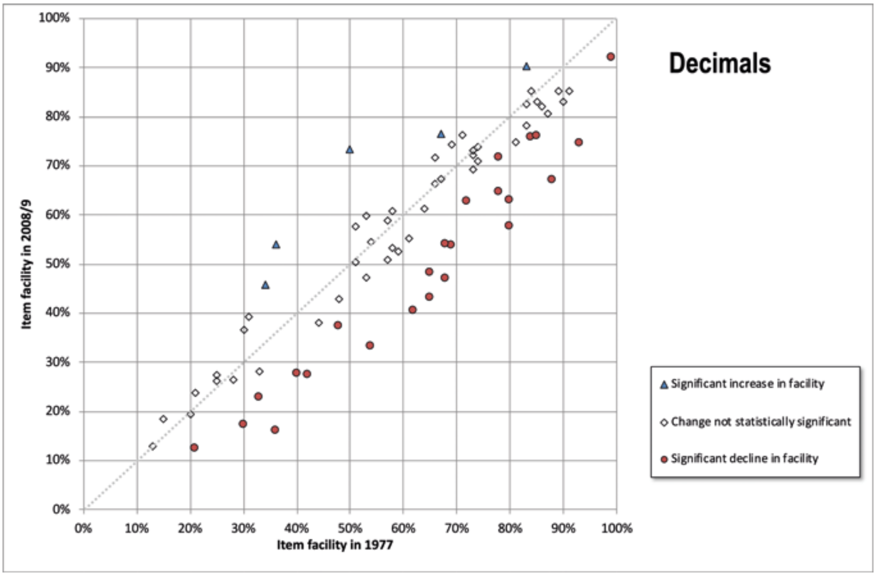


FIGURE 2 Scatterplot of 73 matched items facilities for Decimals test at Grade 8

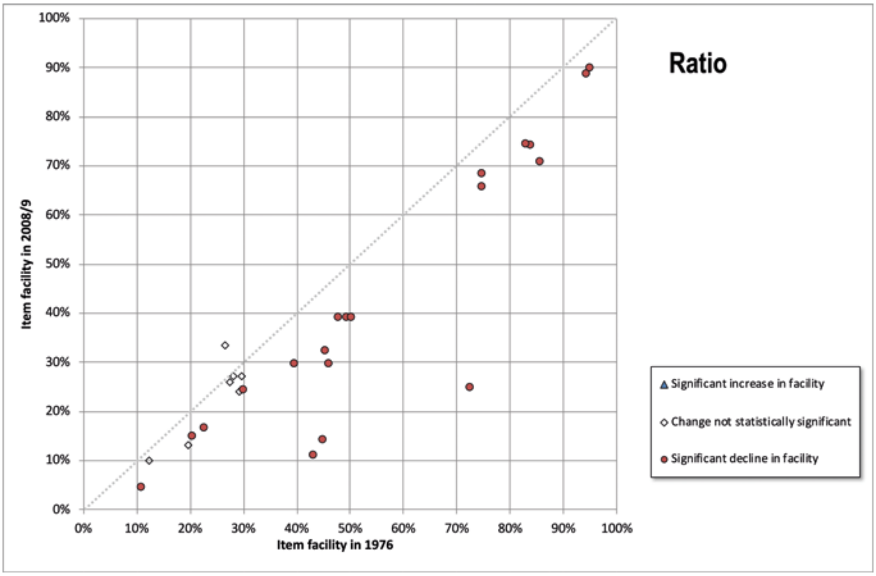


FIGURE 3 Scatterplot of 24 matched items facilities for Ratio test at Grade 8

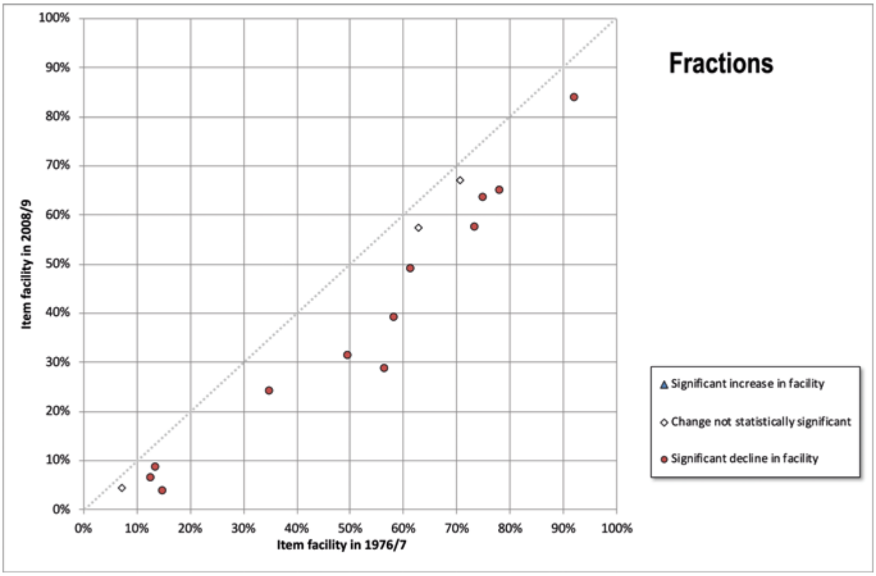
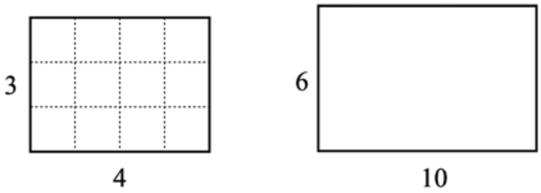


FIGURE 4 Scatterplot of 15 matched items facilities for Fractions items at Grade 8

The picture for Algebra shows a similar decline across the range of facilities, except that there are significant declines for a cluster of items with facilities greater than 75% in 1976. In 2008–2009 several items involving apparently simple arithmetic appear to be surprisingly difficult when presented in the context of the algebra test. For example, two items (see Figure 5) involving numerical calculations of area, a 3 by 4 rectangle (with a grid) and a 6 by 10 rectangle (without a grid), show declines from 91.4% to 78.4%, and 88.6% to 75.0%, respectively. Other items presented in geometric contexts also showed considerable declines in facilities. For example, two items (see Figure 6) involved enumerating diagonals in a polygon, given example of a five-sided polygon. The item facilities for 57 and *k* sided polygons had reduced from 74.6% to 53.8%, and 52,0% to 41.7%, respectively. This may be due to less emphasis being placed on geometry in general than in the 1970s, but is nevertheless salutary considering the extensive use of geometric contexts in the teaching of algebra at low secondary.

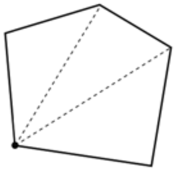
What are the areas of these shapes?



A = ... A = ...

FIGURE 5 Items 7a and 7b involving the numerical calculation of area from the Algebra test

In a drawing like this
you can work out the number of diagonals by
taking away 3 from the number of sides.



So, a shape with 5 sides has 2 diagonals;

a shape with 57 sides has .... diagonals;

a shape with *k* sides has .... diagonals.

FIGURE 6 Items 15a and 15b involving the enumeration of diagonals in a polygon from the Algebra test

TABLE 3 Facilities, standard errors and change over time for the four items involving percentages on Ratio (Test R) at Grade 8 (age 13–14) with abbreviated descriptions

Item and description	1976 Facility (SE)	2008–2009 Facility (SE)	Change Difference (SE)
8a: 4 out of 100 children, what percentage?	85.7% (3.93%)	70.6% (2.05%)	−15.1% (4.43%)
8b: 6% of 250 children?	45.5% (3.65%)	32.3% (1.88%)	−13.2% (4.10%)
8c: 24 out of 800 Avenger cars, what percentage?	39.6% (2.65%)	29.5% (1.40%)	−10.1% (3.00%)
8d: £20 coat reduced by 5%, cost now?	26.5% (3.44%)	33.5% (1.73%)	7.0% (3.85%)

The Ratio test included four items involving percentages (8a–d), which provide further striking examples of the decline. The facilities at Grade 8 for 1976 and 2008–2009 are presented in Table 3 together with the change in facility over time. It can be seen that the facilities for three of the items show considerable declines of between 10% and 15%. However, the final item (8d), which asks students to work out the cost of a £20 coat when reduced by 5%, shows a non-significant percentage point increase in the facility of 7% from 26.5% in 1976 to 33.5% in 2008–2009. This may be due to greater emphasis being placed on mental and other 'informal' methods for calculating percentages.[13]

One additional and potentially important change is that the proportion of blank or non-responses has increased over time. The frequencies of blank responses have risen from means of 12.8%, 6.7% and 7.6% in the 1970s to means of 21.0%, 17.9% and 17.9% in 2008–2009 for the Algebra, Decimals and Ratio items, respectively. This is a curious result, which we address in the discussion below. The blank responses for the Fractions items have also increased (from a mean of 14.6% in the 1970s to 31.6% in 2008–2009). This increase is perhaps less surprising, given that much less emphasis was placed on fractions in 2008–2009 than was the case in the 1970s.

[13] In 2008–2009, students in Year 7 (age 11–12) were required to "Know that 10% is equivalent to 1/10 = 0.1, and 5% is half of 10%". Valued Added Tax (VAT) is a sales tax levied across the UK.

5.2 *Changes across the Attainment Range at Grade 8*

We now turn to examine how students' understanding has changed across the attainment range. We examine how the proportions of students at each level in the CSMS topic hierarchies have changed. Again, we focus on the oldest students tested, those at Grade 8. Here, we report on Algebra, Decimals and Ratio, but not Fractions, because only a small subset of Fractions items was used in the 2008–2009 administration.

In Table 4 and Figure 7, we show the change in the proportion of Grade 8 students achieving each Level or above in the CSMS hierarchy for each test. As noted earlier, the levels were well-ordered in both administrations of the tests and there were very few students who achieved a higher level but not a lower level.

TABLE 4 Change over time of proportions of students achieving CSMS hierarchy levels in Algebra, Decimals and Ratio at Grade 8

	Proportion of cohort (SE)		**Change**	
	1976–1977	**2008–2009**	**Difference (SE)**	**Significant?**
Algebra				
At Level 0	6.0% (3.28%)	16.3% (1.68%)	10.3% (3.69%)	UP
Level 1 or above	94.0% (3.28%)	83.7% (1.68%)	–10.3% (3.69%)	DOWN
Level 2 or above	59.0% (4.50%)	58.8% (2.40%)	–0.2% (5.10%)	n.s.
Level 3 or above	35.0% (2.06%)	27.2% (1.02%)	–7.8% (2.30%)	DOWN
Level 4	6.0% (1.06%)	2.6% (0.56%)	–3.4% (1.20%)	DOWN
Decimals				
At Level 0	7.0% (3.40%)	11.5% (1.94%)	4.5% (3.92%)	n.s.
Level 1 or above	93.0% (3.40%)	88.5% (1.94%)	–4.5% (3.92%)	n.s.
Level 2 or above	84.0% (3.58%)	83.7% (2.05%)	–0.3% (4.13%)	n.s.
Level 3 or above	67.0% (4.17%)	68.7% (2.39%)	1.7% (4.81%)	n.s.
Level 4 or above	44.0% (4.17%)	51.9% (2.35%)	7.9% (4.78%)	n.s.
Level 5 or above	32.0% (3.65%)	26.6% (2.09%)	–5.4% (4.20%)	n.s.
Level 6	16.0% (2.43%)	12.3% (1.31%)	–3.7% (2.76%)	n.s.
Ratio				
At Level 0	7.0% (2.59%)	15.7% (1.34%)	8.7% (2.91%)	UP
Level 1 or above	93.0% (2.59%)	84.3% (1.34%)	–8.7% (2.91%)	DOWN
Level 2 or above	44.0% (2.77%)	34.8% (1.49%)	–9.2% (3.14%)	DOWN
Level 3 or above	21.0% (2.77%)	18.5% (1.45%)	–2.5% (3.13%)	n.s.
Level 4	9.0% (1.86%)	6.2% (1.01%)	–2.8% (2.12%)	n.s.

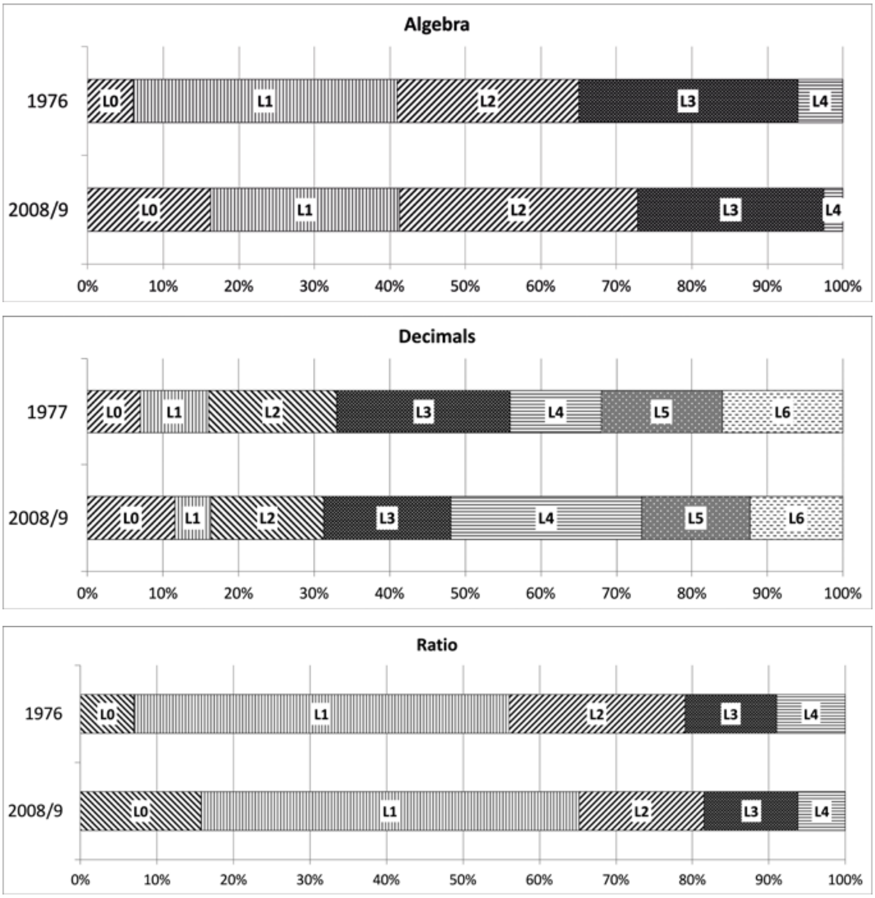


FIGURE 7 Proportional bar charts showing achievement of CSMS hierarchy levels in Algebra, Decimals and Ratio across the cohort at Grade 8. Key: L0, Level 0, etc.

In Algebra and Ratio, the proportion of the lowest achievers, i.e., those at "Level 0", has increased dramatically over time. In Algebra, the proportions have significantly declined for those achieving each level or above, except Level 2 or above. The proportion of students achieving at least Level 3 is of particular interest, since this is when students begin to understand the key algebraic concept of variable. For Ratio, the proportions achieving at least Level 1 and at least Level 2 have declined significantly. Level 2 is important, since this is when students begin to understand contexts involving non-integer ratios as being multiplicative. The proportion achieving at least Level 2 has declined to only around a third of the cohort. The results for Decimals indicate a slightly more positive picture, in that there has been some improvement for the middle range of attainment, with an increase in the proportion achieving Level 4 or above, although this is offset by what appears to be a corresponding decline at the highest levels. As with Algebra and Ratio, the proportion of the lowest achievers "at Level 0" has increased, although the change is smaller and is not statistically significant.

5.3 *Comparing Changes at Grade 6 and Grade 7 to Those at Grade 8*

In this section, we compare progression in Algebra, Decimals, Ratio and Fractions. Thus, in Table 5, we compare the mean facilities for 1976 or 1977 and 2008–2009 at Grade 8 with those at Grade 7 (aged 12–13) for all four areas and at Grade 6 (aged 11–12) for Decimals and Fractions. Changes across

TABLE 5 Change over time for Algebra, Decimals, Ratio and Fractions at all grades

	1976–1977	2008–2009	Change
Algebra (51 items)			
Grade 8	48.8%	41.9%	–6.8%
Grade 7	39.1%	37.9%	–1.2%
Progression Grade 7 to Grade 8	+9.7%	+4.0%	–
Decimals (73 items)			
Grade 8	60.0%	55.7%	–4.4%
Grade 7	53.1%	53.7%	0.5%
Grade 6	46.6%	49.6%	3.0%
Progression Grade 6 to 8	+13.4%	+6.0%	–
Progression Grade 7 to 8	+6.9%	+3.0%	–
Ratio (24 items)			
Grade 8	47.0%	40.0%	–6.9%
Grade 7	44.2%	35.1%	–9.2%
Progression Grade 7 to 8	+2.8%	+4.9%	–
Fractions (10 items)			
Grade 8	66.3%	53.5%	–12.9%
Grade 7	61.1%	50.7%	–10.3%
Grade 6	62.5%	47.5%	–15.0%
Grade 8 [All 15 common items]	*50.8%*	*39.3%*	*–11.5%*
Progression Grade 6 to 8	+3.8%	+6.0%	–
Progression Grade 7 to 8	+5.2%	+2.8%	–

For Fractions, there are more common items at Grade 8, and the mean facilities at each administration for these 15 common items are shown in italics.

TABLE 6 Effect sizes (Cohen's *d*) of gain in attainment from the end of Grade 6 to the end of Grade 8 across the attainment range for the Algebra, Decimals and Ratio tests in 2008–2009

	Gain in attainment at percentiles of cohort							Increase in attainment 'gap': 10th to 90th percentile
	5th	10th	25th	50th	75th	90th	95th	
Algebra	0.08	0.17	0.51	0.68	0.59	0.59	0.59	0.42
Decimals	0.03	0.06	0.23	0.34	0.23	0.29	0.23	0.23
Ratio	0.11	0.11	0.21	0.32	0.53	.063	0.48	0.53

the 30 years at Grades 6 and 7 were rather different across the four areas. In Ratio (Grade 7 only) and Fractions, the changes were similar to Grade 8; in Algebra (Grade 7 only), there was only a small decline. In decimals, there were small rises, greater for Grade 6 than Grade 7. In Table 6, we compare progression from Grade 6 to Grade 8 across the attainment range in 2008–2009 for Algebra, Decimals and Ratio. It can be seen that the gap in attainment from the 10th to the 90th percentile increases for the older students with greater increases for Algebra and Ratio. This is in large part due to much smaller gains for the lowest attainers.

An important change to the curriculum is that symbolic algebra is now generally introduced in Grade 6, which, as we noted previously, is earlier than in the 1970s (Brown, 1996). One might have expected that this earlier introduction would have boosted performance at Grade 7, so the fact that the 2008–2009 mean facility is below that for 1976 is surprising. Moreover, since the Grade 8 students in 2008–2009 had a longer exposure to symbolic algebra than their counterparts in 1976, three years rather than two, one might have expected a further boost in performance at Grade 8. Hence, it is very striking indeed that the change in mean facility, or progression, from Grade 7 to Grade 8 has more than halved to 4.0%. Moreover, and that over time the gap between the highest and lowest attainers actually widens from Grade 6 to Grade 9. This suggests that the earlier introduction of symbolic algebra may have had little or no lasting effect beyond a possible 'initiation' effect, particularly for the lowest attaining group of students.

For Decimals, the increase at Grade 6 may reflect the fact that many aspects of decimal number, particularly measurement aspects such as place value and the use of number lines, are now taught in primary (DfEE, 1999). Decimals are

TABLE 7 Proportion of students "at Level 0" in the CSMS hierarchy in Algebra, Decimals and Ratio at all grades

	Proportion of cohort (SE)		Change	
	1976–1977	2008–2009	Difference (SE)	Significant?
Algebra				
Grade 8	6.0% (3.28%)	16.3% (1.68%)	10.3% (3.69%)	UP
Grade 7	10.0% (3.82%)	18.7% (2.14%)	8.7% (4.38%)	UP
Decimals				
Grade 8	7.0% (3.40%)	11.5% (1.94%)	4.5% (3.92%)	n.s.
Grade 7	10.0% (3.24%)	10.0% (1.76%)	0.0% (3.69%)	n.s.
Grade 6	17.0% (3.60%)	13.4% (1.92%)	–3.6% (4.08%)	n.s.
Ratio				
Grade 8	7.0% (2.59%)	15.7% (1.34%)	8.7% (2.91%)	UP
Grade 7	7.0% (2.07%)	21.9% (1.20%)	14.9% (2.40%)	UP

taught much more extensively and very much earlier than in the 1970s and there is much greater emphasis on decimals outside school. It is therefore somewhat surprising that this has not resulted in better performance on the Grade 8 tests.

For Ratio, the decrease at Grade 7 is of a similar order to the decrease at Grade 8, suggesting a consistent decline in the understanding of ratio, but it is striking that the 2008–2009 mean facility for Grade 8 is below that for Grade 7 in the 1970s.

For Fractions, the mean facilities have declined substantially for all three age groups, with a slightly greater decrease for Grade 6. The current Grade 8 facility is well below that for Grade 6 in the 1970s. This is perhaps to be expected, because as we have already noted there is now much less teaching of fractions than in the 1970s. The change in mean facility, or progression, from Grade 6 to Grade 8 has increased from 3.8% in 1976–1977 to 6.0% in 2008–2009.

It can also be seen from Table 7 that the proportion of the lowest attaining group of students, those "at Level 0" in the CSMS hierarchy, has increased significantly for all ages of students in Algebra and Ratio. For Decimals, the proportion of low attaining students has fallen slightly at Grade 6 (but not significantly) and remained stable at Grade 7.

6 Discussion

6.1 *On Replication*

Before discussing the results of the replication further, we consider the challenges that we faced in replicating a study carried out in the 1970s and how we used methods now available to address these replication issues. The CSMS study was at the time one of the largest and most rigorous studies that had been carried out worldwide. Yet, viewed with a modern eye, the study has some limitations. The original analysis was limited by the methods and computing power then available. In addition, the expectations regarding statistical practice and reporting in the 1970s were lower. Hence, the original study reported point estimates but not standard errors or confidence intervals. Such measures of precision are critical to judging the significance of changes over time. These kinds of issues are likely to affect all, or most, of the significant studies carried out prior to the mid-1990s before academic journals in education began to establish clear guidelines for statistical reporting (Hill & Shih, 2009). We addressed this by making use of statistical techniques and computing power now available. First, we used a more robust modern method, Rasch modelling, to re-validate the tests. Second, we used statistical simulation to estimate standard errors, thus enabling the comparison. Ideally, our claims would be validated against other assessments and samples, but no such evidence is available. In the absence of any other data, as is likely to be the case in any other similar replication, we believe modern statistical methods, such as simulation, have an important role to play in replicating studies and comparing the results over time.

Some might argue that the study carried out here is simply a comparison of two national-scale studies and, whilst the comparison itself is of value, the study does not provide a specific contribution to the replication literature. Certainly, the examination of national-level reform has received little attention in the literature on replication (although this has been a significant concern of the implementation literature, see, e.g., Helenius, 2021). Following Hüffmeier et al., (2016), Melhuish (2018) provides a typology of replication types: exact, close, constructive and conceptual. Of most relevance to the research presented here are the constructive and conceptual variants, which Melhuish argues "may contain divergences from the original study to better test, refine, or expand a theory or theoretical propositions" (p. 11). Constructive replications do this by introducing a new element, whilst conceptual replications contain changes to the methodology. The study reported here involves both a different element to the original study, a sample from the population of English students at a different time point, and changes to the methodology, specifically the use of modern statistical methods. This combination of changes has enabled us to refine and expand the original findings about how students understand these key areas of mathematics and how these understandings develop over the lower secondary phase. This aspect of the replication is certainly very important, but is reported elsewhere (e.g., Hodgen et al., 2012). In contrast, the focus of this paper is on examining whether the levels of understanding, which were identified in the 1970s, still hold 30 years later and, thus, assessing whether reform in England has had a positive effect on student attainment. There are, of course, other tests and surveys that enable a comparative analysis of these long-term trends over time, such as national tests (such as GCSEs in England) and international surveys (such as PISA and TIMSS), although our replication demonstrates how comparisons can be made when other tests and surveys are either not available or are limited in scope. Importantly, this study focuses specifically on conceptual understanding, a key strand of mathematics that is often underemphasised in official tests and surveys. Thus, as a replication, this study provides an independent research-led assessment that is not influenced by national or international political concerns and makes an additional contribution by demonstrating how modern statistical methods can overcome the challenges of comparisons with studies where the original data is either not available or available only in a limited aggregated form.

6.2 *Comparing Performance over Time*

A major conclusion of our replication study is that, at Grade 8, there has been an overall decline in students' attainment since the mid-1970s in each of the areas tested. There is a more mixed picture for Decimals, where students' understanding appears to have slightly increased over time for the middle attaining students, although this is in the context of an overall decline. This general decline is a surprising result, since, as we have already noted, England has seen a concerted attempt to improve educational performance in mathematics over the past 30 years.

Ultimately, if our aim is to measure the change in attainment, the benefits of further refining our estimates of population parameters from the 2008–2009 sample are constrained by the much larger uncertainty around the estimates from 1976–1977.

The decline is equivalent to effect sizes of Cohen's $d = -0.32$, -0.18, -0.45 and -0.29 for Algebra, Decimals, Ratio and Fractions, respectively. Effect sizes of this order are often classed as low to moderate in the educational literature when judging the impact of educational interventions. However, these are arguably large systemic effects and are of a similar order to major changes in systems' performances on TIMSS and PISA. They are also large in relation to the typical growth we observed in students' scores on the tests between Grades 6 to 8. The decline in performance in Fractions is equivalent to the progress typically made in two years of schooling, while for the other three tests it is of the order of over a year. In other words, students at the end of Grade 8 in 1976–1977 were well over a year ahead of their counterparts in 2008–2009.

This overall decline is in marked contrast to the increase in examination results, which have risen dramatically over the period. There are several possible reasons for this anomaly. One possible hypothesis is that the nature and value of qualifications has changed. There is a great deal of recent research indicating that grade standards in English mathematics examinations may have 'slipped' over time (e.g., Coe & Tymms, 2008; Jones et al., 2016). It is important to note that obtaining qualifications, particularly in mathematics, has become much more crucial for all students since the 1970s. Hence, examination results may have improved because a greater proportion of students have been given the opportunity to sit the examinations, because these students have greater motivation to do well and because schools are influenced by accountability measures.

It could also be that tests which focus on a deeper level of reasoning, such as the CSMS tests, show a decline, whereas those, such as the national GCSE examination, involving more routine items and/or more coached performances show an improvement. Indeed, as we have discussed above, the CSMS tests were deliberately designed to test conceptual understanding rather than the ability to perform routine, procedural tasks. In addition, our own comparative analysis of mathematical textbooks in England (Hodgen et al., 2010) suggests that much less emphasis was placed on conceptual understanding in 2008–2009 than in the 1970s.

Another possible explanation is that, unlike GCSE examinations, the ICCAMS and original CSMS tests were administered without preparation or revision, whilst secondary education in England has become more highly focused on examination performance in recent years (Office for Standards in Education, 2012). It may be that an effect of the rise in prevalence of high-stakes testing between 1976–1977 and 2008–2009 is that low-stakes tests (such as ours) seem less worthy of effort by comparison. This might also explain why the number of unanswered questions is higher in 2008–2009 than 1976–1977. Unfortunately, the existing literature is not conclusive. Penk et al., (2014) show that some studies do find an association between test performance and motivation, whereas others do not. Some experimental studies do find relatively large effects for motivation, although these may be distorted by the effect of academic ability (Wise & DeMars, 2005). In addition, these large effects are recorded in designs that emphasise extreme differences in the stakes of tests or monetary incentives. Whilst we cannot rule out the possibility of a test motivation effect in the decline, the current evidence suggests that any such effect would be small at most, given that the test was low stakes at both administrations.

The decline may also be related to changes in the population of students in England's schools, particularly to changes in the proportion of ethnic minority students, students with English as an Additional Language (EAL) or students with Special Educational Needs (SEN). Unfortunately, data on students' ethnicity or EAL was not systematically collected in the 1970s (Khan, 1983), but it is generally accepted that the proportions of these students significantly increased over the period. However, the evidence that is available suggests that these factors have not contributed to the decline (and might, if anything, have reduced the decline over time). Strand (2015) finds that, over the period 1991 to 2006, the gap in educational attainment between ethnic minority and White British students has narrowed. In addition, Strand et al., (2015) examine the relationship between EAL and achievement between 1997 and 2013 and, whilst they identify an attainment gap in the early years, they find that, in mathematics, this is almost eliminated by age 11, and that, by age 16, EAL students outperform First Language English students.

One serious issue concerns the proportion of the lowest attaining students, those who fail to achieve Level 1 and are thus "at Level 0". In Algebra and Ratio, the proportion of these students has more than doubled to around 15% of the population. These students have difficulty with the very simplest items on the tests and thus do not fully grasp some of the core ideas in the primary curriculum. This may be partially reflected in the TIMSS results, which show no change between 1995 and 2007 in the proportion of Grade 8 students who do not achieve the low international benchmark, despite a significant rise in England's average attainment (Sturman et al., 2008). It is difficult to explain this; one possibility is the closing of many Special Schools and greater inclusivity of students with SEN within the mainstream sector. The Warnock Report into special educational needs records that 1.8% of the school population (ages 5–15) were in special schools or classes in 1977 (Warnock, 1978, p. 37). In 2007, 1.05% of students were in special schools, with a further 0.2% in pupil referral units (Department for Children Schools and Families, 2008). Hence, it is unlikely that this factor could account for the full size of the difference. Another possible explanation lies in the finding that the National Numeracy Strategy had the effect of depressing attainment at the lower end, perhaps because of the failure to address children's particular needs in attempting to provide equal access to the curriculum (Barnes et al., 2003; Brown et al., 2008).

The decline is also in contrast to England's performance in international surveys, although, as previously noted, these surveys enable reliable comparisons only over a much shorter period: back to 1995 for TIMSS and back to 2003 for PISA. One possible explanation for England's rises in TIMSS is that the English mathematics curriculum has become closer to the curriculum tested, particularly at primary (Brown, 2011; Burstein, 1992). It is also important to note that the CSMS tests do not test the whole curriculum and, indeed, do not test the entirety of algebra and multiplicative reasoning. Nevertheless, the topics tested are critical to further progression in mathematics.

Of course, it is possible that, whilst mathematical performance in England declined over the period as a whole, this decline may not have been monotonic; our findings would be consistent with some improvements over shorter periods. Indeed, the evidence at primary level does suggest a modest rise since 1995 (Tymms, 2004), although the OECD's (2013) survey of adult skills suggests a gradual decline over the period. Nevertheless, the issue of when the decline took place, and indeed whether the decline is associated with any particular reform initiatives, remains open.

A related issue is the increase in the proportion of blank responses to questions. In 1976–1977, missing responses were treated as incorrect, and so, for purposes of comparability, we have treated the 2008–2009 missing responses as incorrect. Brown et al.'s (2014) analysis of NAEP data indicates that this is a reasonable approach, because alternative methods (such as ignoring missing responses or using imputation) produce similar estimates for large samples. Analysis of the missing responses does suggest that the rise in missing responses is more than would be expected to arise purely from increased difficulty, although the effect is relatively small (Coe & Hodgen, 2017c). This does not appear to be the result of students having less time for the tests. One possible explanation is that an increased focus on examination technique has led to some students leaving a blank response to items that they are unsure about.

The fall in the proportion of those at the highest level of attainment is also of concern. Although this fall is statistically significant only for the Algebra test, the actual proportion of the current cohort at this level of attainment is worryingly low.

On the Decimals test, the effect of a rise in attainment focused in that section of the attainment range between the 15th and 60th percentile again has possible explanations. This could reflect greater focus on coaching students predicted to be around the Level 4 borderline in Year 6 and then the C/D borderline at GCSE, since these are key performance indicators for schools in England. While this does not explain why this feature is not present in any of the other curriculum areas, this could be attributed to the fact that these borderline students are more likely to have been coached in basic number than in the more formal and abstract topic of algebra. However, these differences may also occur because of cultural changes in student knowledge. There was probably more use and knowledge of fractions in 1970s society in England. The 1970s saw the advent of decimalised money and metrication, and also the rise of calculator and computer use, which employ decimal notation. These societal changes probably had the effect of enhancing knowledge about decimals in relation to knowledge about fractions. This would explain both improvements in the middle range for Decimals and the presence in that test of a greater proportion of items which are unchanged or improved compared to other areas.

7 Conclusion

In this study, we conducted a 'scaling out' replication of the CSMS study originally carried out in the 1970s in order to compare performance over time in key areas of lower secondary mathematics. One key finding of the study is to demonstrate how modern statistical methods can be used to carry out such a comparison, when the original data and statistical findings are no longer fully available.

It certainly seems counterintuitive that given the long list of major Government initiatives between 1985 and 2009 aimed at increasing attainment in mathematics there has not been any obvious positive effect on the understanding and application of two of the key areas in lower secondary. Of course, one might speculate that this list *in itself* explains why performance does not appear to have risen; the effectiveness of teachers and schools may be negatively affected by initiative overload (Perryman et al., 2011). Indeed, there is some evidence that higher performing countries are less prone to frequent external initiatives (Askew et al., 2010). The evidence presented here does suggest that government investment on initiatives is not sufficient on its own to increase mathematical attainment across the system and that it is also important to focus on the quality of reform initiatives. Hence, one implication of this study is that such initiatives should include research focused on building the evidence base on the efficacy of educational interventions (Coe, 2009).

A further implication of this study is that it is important to take steps to monitor standards of attainment over time. Frequently, the debate about educational performance in England and elsewhere has focused on examination standards (Anthony & Walshaw, 2007; Truss, 2013; Walport et al., 2010). However, we believe this focus to be somewhat misplaced, since the nature and purpose of the examination system changes significantly over time, and public, high-stakes examinations are not well-suited to monitoring standards over time. In the US, the NAEP-LTT program goes at least some way towards meeting this need, but in England there is currently no equivalent. If we want to know about system-wide changes in performance over time, we need an assessment program designed for this purpose.

A related implication is that there is a need for mathematics education to place greater focus on conceptual understanding (see also, Kilpatrick et al., 2001). Procedural understanding of mathematics is important, but conceptual undertstanding is critical to using mathematics in new and unfamiliar contexts.

Overall, our results are sobering. In England, over a 30-year period, despite huge investment in well-intentioned reform and widespread perception of improvement, student outcomes appear to have declined, at least in the key areas of mathematics that are the focus of this study. The most plausible interpretation of our results is that overall attainment in mathematics for 14 year olds in England has declined. This should be a salutary warning to anyone who thinks that systemic educational improvement can be decreed, imposed, bought or assumed: evidently it needs something much smarter than that.

Acknowledgements

This research was supported by Economic and Social Research Council (grant number RES-179-25-0009). We are very grateful to the anonymous schools and students who took part in the research.

References

Afrassa, T. M., & Keeves, J. P. (1999). Changes in students' mathematics achievement in Australian lower secondary schools over time. *International Education Journal, 1*(1), 1–21. https://www.aare.edu.au/data/publications/1997/afrat031.pdf.

Anthony, G., & Walshaw, M. (2007). *Effective pedagogy in mathematics/Pàngarau: Best evidence synthesis iteration*. Ministry of Education.

Aguilar, M. S. (2020). Replication Studies in Mathematics Education: What Kind of Questions Would Be Productive to Explore? *International Journal of Science and Mathematics Education, 18*(1), 37–50. https://doi.org/10.1007/s10763-020-10069-7.

Askew, M., Hodgen, J., Hossain, S., & Bretscher, N. (2010). *Values and variables: A review of mathematics education in high-performing countries*. The Nuffield Foundation.

Ball, S. J. (2013). *The education debate* (2nd ed.). Polity Press.

Ball, S. J., Maguire, M., & Braun, A. (2012). *How schools do policy: policy enactments in secondary schools*. Routledge.

Barber, M., & Mourshed, M. (2007). *How the world's best performing school systems come out on top*. McKinsey & Company.

Barnes, A., Venkatakrishnan, H., & Brown, M. (2003). *Strategy or straitjacket? Teachers' views on the English and mathematics strands of the Key Stage 3 National Strategy: Final Report*. ATL.

Behr, M. J., Harel, G., Post, T., & Lesh, R. (1992). Rational number, ratio and proportion. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 296–333). Macmillan.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental measurement in the human sciences* (2nd ed.). Lawrence Erlbaum.

Boylan, M., Maxwell, B., Wolstenholme, C., Jay, T., & Demack, S. (2018). The Mathematics Teacher Exchange and 'Mastery' in England: The Evidence for the Efficacy of Component Practices. *Education Sciences*, *8*(4), 202. https://www.mdpi.com/2227-7102/8/4/202.

Brooks, G., Foxman, D., & Gorman, T. (1995). *Standards in literacy and numeracy 1948–1994*. National Commission on Education.

Brown, G., Micklewright, J., Schnepf, S. V., & Waldmann, R. (2007). International surveys of educational achievement: how robust are the findings? *Journal of the Royal Statistical Society: Series A* (*Statistics in Society*), *170*(3), 623–646. https://doi.org/10.1111/j.1467-985X.2006.00439.x.

Brown, M. (1981). *Levels of understanding of number operations, place value and decimals in secondary school children*. (Unpublished PhD thesis), Chelsea College, University of London.

Brown, M. (1989). Graded Assessment and learning hierarchies in mathematics: an alternative view. *British Educational Research Journal*, *15*(2), 121–128. https://doi.org/10.1080/0141192890150202.

Brown, M. (1996). The context of the research: the evolution of the national curriculum for mathematics. In D. C. Johnson & A. Millett (Eds.), *Implementing the Mathematics National Curriculum: policy, politics and practice* (pp. 1–28). Paul Chapman Publishing.

Brown, M. (1999). One mathematics for all? In C. Hoyles, C. Morgan & G. Woodhouse (Eds.), *Rethinking the mathematics curriculum* (pp. 1–28). Falmer.

Brown, M. (2011). Going back or going forward? Tensions in the formulation of a new National Curriculum in mathematics. *Curriculum Journal*, *22*(2), 151–165. https://doi.org/10.1080/09585176.2011.574882.

Brown, M., Askew, M., Hodgen, J., Rhodes, V., Millett, A., Denvir, H., & Wiliam, D. (2008). Individual and cohort progression in learning numeracy ages 5–11: Results from the Leverhulme 5-year longitudinal study. In A. Dowker (Ed.), *Children's mathematical difficulties: psychology, neuroscience and education* (pp. 85–108). Elsevier.

Brown, N. J. S., Svetina, D., & Dai, S. (2014). *Impact of methods of scoring omitted responses on achievement gaps*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Philadelphia, PA.

Burstein, L. (Ed.). (1992). *The IEA Study of Mathematics III: student growth and classroom processes*. Pergamon Press.

Calvert, B. (1958). *Non-verbal Test D.H.* NFER.

Carraher, T. N. (1986). Rated addition: a correct additive solution for proportion problems. In L. Burton & C. Hoyles (Eds.), *Proceedings of the Tenth International Conference of the Psychology of Mathematics Education (PME10)* (pp. 247–252). University of London.

Cascella, C., Giberti, C., & Maffia, A. (2023, Onlinefirst). Percentages versus Rasch estimates: alternative methodological strategies for replication studies in mathematics education. *Research in Mathematics Education*, 1–19. https://doi.org/10.1080/14794802.2022.2154826.

Chowdry, H., & Sibieta, L. (2011). *Trends in education and schools spending*. Institute for Fiscal Studies.

Cockcroft, W. H. (1982). Mathematics counts. HMSO.

Coe, R. (2008). Comparability of GCSE examinations in different subjects: an application of the Rasch model. *Oxford Review of Education*, *34*(5), 609–636. https://doi.org/10.1080/03054980801970312.

Coe, R. (2009). School improvement: reality and illusion. *British Journal of Educational Studies*, *57*(4), 363–379. https://doi.org/10.1111/j.1467-8527.2009.00444.x.

Coe, R., & Hodgen, J. (2012). *Selection of the ICCAMS sample. Technical Report No. 01.*

Coe, R., & Hodgen, J. (2014). *Analysis of the performance of the ICCAMS tests and items. Technical Report No. 06.*

Coe, R., & Hodgen, J. (2017a). *Using Bootstrapping to investigate the precision and bias of estimates from the ICCAMS sample. Technical Report No. 04.*

Coe, R., & Hodgen, J. (2017b). *Bias arising from unmatched MidYIS scores. Technical Report No. 03.*

Coe, R., & Hodgen, J. (2017c). *Analysis of missing response bias. Technical Report No. 07.*

Coe, R., & Tymms, P. (2008). *Summary of research on changes in educational standards in the UK*. Institute of Directors.

Collis, K. F. (1975). *The Development of formal Reasoning: Report of a SSRC sponsored project carried out at the University of Nottingham School of Education during 1974*. University of Newcastle, New South Wales.

Department for Children Schools and Families. (2008). *Education and Training Statistics for the United Kingdom: 2008* http://www.education.gov.uk/rsgateway/DB/VOL/v000823/index.shtml.

DfEE. (1999). *The National Numeracy Strategy: Framework for teaching mathematics from Reception to Year 6*. Department for Education and Employment.

Education Endowment Foundation. (2013). *Recruitment and retention pack*. EEF.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman and Hall.

Færch, J. V., & Hodgen, J. (2023). Danish Students' Understanding of Fractions: A Replication Study. *Implementation and Replication Studies in Mathematics Education*, *3*(2), 200–242. https://doi.org/https://doi.org/10.1163/26670127-bja10014.

Fullan, M. (2000). The return of large-scale reform. *Journal of Educational Change*, *1*(1), 5–27. https://doi.org/10.1023/a:1010068703786.

Galton, M., Hargreaves, L., Comber, C., Wall, D., & Pell, A. (1999). *Inside the primary classroom: twenty years on*. RoutledgeFalmer.

Giberti, C., & Maffia, A. (2022). Primitive Model of Partitive Division: A Replication of the Fischbein et al., Study. *Implementation and Replication Studies in Mathematics Education*, *2*(2), 149–173. https://doi.org/10.1163/26670127-bja10007.

Goldin, C., & Katz, L. F. (2008). *The race between education and technology*. Harvard University Press.

Goldstein, H., & Heath, A. (Eds.). (2000). *Educational standards*. The British Academy.

Hanushek, E. A., Peterson, P. E., & Woessmann, L. (2010). *U.S. math performance in global perspective: how well does each state do at producing high-achieving students?* Harvard's Program on Education Policy & Governance (PEPG), Harvard Kennedy School.

Hart, K., Brown, M., Kerslake, D., Küchemann, D. E., & Ruddock, G. (1985). *Chelsea Diagnostic Mathematics Tests: Teacher's Guide*. NFER-Nelson.

Hart, K., Brown, M. L., Küchemann, D. E., Kerslake, D., Ruddock, G., & McCartney, M. (1981). *Children's understanding of mathematics: 11–16*. John Murray.

Hart, K. M. (1980). *Secondary school-children's understanding of ratio and proportion*. (Unpublished PhD thesis), Chelsea College, University of London.

Hart, K. M., & Johnson, D. C. (Eds.). (1983). *Secondary school children's understanding of mathematics: A report of the mathematics component of the concepts in secondary mathematics and science programme*. Centre for Science Education, Chelsea College.

Hiebert, J. (Ed.). (1986). *Conceptual and procedural knowledge: The case of mathematics*. Lawrence Erlbaum Associates.

Hill, H. C., & Shih, J. C. (2009). Examining the Quality of Statistical Mathematics Education Research. *Journal for Research in Mathematics Education*, *40*(3), 241–250. http://www.jstor.org/stable/40539336.

Hodgen, J., Brown, M., Coe, R., & Küchemann, D. (2012). Surveying lower secondary students' understandings of algebra and multiplicative reasoning: to what extent do particular errors and incorrect strategies indicate more sophisticated understandings? In J. C. Sung (Ed.), *Proceedings of the 12th International Congress on Mathematical Education (ICME-12)* (pp. 6572–6580). International Mathematics Union.

Hodgen, J., Foster, C., & Brown, M. (2022). Low attainment in mathematics: An analysis of 60 years of policy discourse in England. *The Curriculum Journal*, *33*(1), 5–24. https://doi.org/10.1002/curj.128.

Hodgen, J., Küchemann, D., & Brown, M. (2010). Textbooks for the teaching of algebra in lower secondary school: are they informed by research? *Pedagogies*, *5*(3), 187–201. https://doi.org/10.1080/1554480X.2013.739275.

Hüffmeier, J., Mazei, J., & Schultze, T. (2016). Reconceptualizing replication as a sequence of different studies: A replication typology. *Journal of Experimental Social Psychology*, *66*, 81–92. https://doi.org/https://doi.org/10.1016/j.jesp.2015.09.009.

Jerrim, J. (2013). The reliability of trends over time in international education test scores: is the performance of England's secondary school pupils really in relative decline? *Journal of Social Policy*, *42*(2), 259–279. https://doi.org/10.1017/S0047279412000827.

Jones, I., Wheadon, C., Humphries, S., & Inglis, M. (2016). Fifty years of A-level mathematics: have standards changed? *British Educational Research Journal*, 543–560. https://doi.org/10.1002/berj.3224.

Karplus, R., Karplus, E., & Wollman, W. (1972). *Intellectual development beyond elementary school IV: Ratio, the influence of cognitive style*. Lawrence Hall of Science, University of California, Berkeley.

Karplus, R., & Petersen, R. (1970). *Intellectual Development Beyond Elementary School II. Ratio, a Survey*. Science Curriculum Improvement Study, Lawrence Hall of Science, University of California, Berkley.

Kieren, T. E. (1976). On the mathematical, cognitive and instructional foundations of rational numbers. In A. R. Osborne & A. B. David (Eds.), *Number and measurement: papers from a research workshop* (pp. 101–144). ERIC Center for Science, Mathematics, and Environmental Education.

Kieren, T. E. (1988). Personal Knowledge of Rational Numbers: Its Intuitive and Formal Development. In J. Hiebert & M. J. Behr (Eds.), *Number Concepts and Operations in the Middle Grades*. Lawrence Erlbaum.

Kilpatrick, J. (2011). *Review of "U.S. Math Performance in Global Perspective: How Well Does Each State Do at Producing High-Achieving Students?"*. National Education Policy Center.

Kilpatrick, J., Swafford, J., & Findell, B. (Eds.). (2001). *Adding It Up: Helping Children Learn Mathematics* (*Prepared by the Mathematics Learning Study Committee, National Research Council*). The National Academies Press.

Khan, V. S. (1983). The "mother-tongue" of linguistic minorities in multicultural England. *Early Years*, *3*(2), 8–25. https://doi.org/10.1080/0957514830030205.

Kloosterman, P. (2010). Mathematics skills of 17-year-olds in the United States: 1978 to 2004. *Journal for Research in Mathematics Education*, *41*(1), 20–51. http://www.jstor.org/stable/40539363.

Küchemann, D. E. (1981). *The understanding of generalised arithmetic (algebra) by secondary school children.* (Unpublished PhD thesis), Chelsea College, University of London.

Lamon, S. J. (2007). Rational Numbers and Proportional Reasoning: Toward a Theoretical Framework for Research. In F. K. J. Lester (Ed.), *Second handbook of Research on mathematics teaching and learning* (pp. 629–667). Information Age Publishing.

Majewska, D., Rushton, N., & Shaw, S. (2022). *How did we get here? Timelines showing changes to maths education in England and the United States.* Cambridge Assessment.

Melhuish, K. (2018). Three conceptual replication studies in group theory. *Journal for Research in Mathematics Education, 49*(1), 9–38. https://doi.org/https://doi.org/10.5951/jresematheduc.49.1.0009.

Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics.* TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College / International Association for the Evaluation of Educational Achievement (IEA).

Nuthall, G. (1985). Standards of achievement in schools. *Economic Bulletin, 702.*

Oates, T. (2011). Could do better: using international comparisons to refine the National Curriculum in England. *Curriculum Journal, 22*(2), 121–150. https://doi.org/10.1080/09585176.2011.578908.

OECD. (2013). *OECD Skills Outlook 2013: First results from the Survey of Adult Skills.* OECD Publishing.

Office for Standards in Education. (2012). *Mathematics: Made to measure.* OFSTED.

Openshaw, R., & Walshaw, M. (2010). *Are our standards slipping? Debates over literacy and numeracy standards in New Zealand since 1945.* New Zealand Council for Educational Research.

Penk, C., Pöhlmann, C., & Roppelt, A. (2014). The role of test-taking motivation for students' performance in low-stakes assessments: an investigation of school-track-specific differences. *Large-scale Assessments in Education, 2*(1), 1–17. https://doi.org/10.1186/s40536-014-0005-4.

Perryman, J., Ball, S., Maguire, M., & Braun, A. (2011). Life in the pressure cooker — school league tables and English and mathematics teachers' responses to accountability in a results-driven era. *British Journal of Educational Studies, 59*(2), 179–195. https://doi.org/10.1080/00071005.2011.578568.

Piaget, J., Inhelder, B., & Szeminska, A. (1960). *The Child's Conception of Geometry.* Basic Books.

Rashid, S., & Brooks, G. (2010). *The levels of attainment in literacy and numeracy of 13- to 19-year-olds in England, 1948–2009: Research report.* National Research and Development Centre for Adult Literacy and Numeracy (NRDC).

Shayer, M., Ginsberg, D., & Coe, R. (2007). Thirty years on — a large anti-Flynn effect? The Piagetian test volume & heaviness norms 1975–2003. *British Journal of Educational Psychology*, *77*(1), 25–41. https://doi.org/10.1348/000709906X96987.

Strand, S. (2015). *Ethnicity, deprivation and educational achievement at age 16 in England: Trends over time. Annex to compendium of evidence on ethnic minority resilience to the effects of deprivation on attainment.* (*DfE RR439B*). Department for Education.

Strand, S., Malmberg, L., & Hall, J. (2015). *English as an additional language and educational achievement in England: An analysis of the National Pupil Database*. Educational Endowment Fund.

Sturman, L., Ruddock, G., Burge, B., Styles, B., Lin, Y., & Vappula, H. (2008). *England's Achievement in TIMSS 2007: National Report for England*. NFER.

Sumner, R. (1975). *Tests of attainment of mathematics in schools: project report*. NFER.

Truss, E. (2013). *Open lecture on A-level reforms*. Speech at the Institute of Education, London, 7th March. https://www.gov.uk/government/speeches/institute-of-education-open-lecture-on-a-level-reform.

Tymms, P. (2004). Are standards rising in English primary schools? *British Educational Research Journal*, *30*(4), 477–494. https://doi.org/10.1080/01411920420002371 94.

Tymms, P., & Coe, R. (2003). Celebration of the Success of Distributed Research with Schools: the CEM Centre, Durham. *British Educational Research Journal*, *29*(5), 639–653. https://doi.org/10.1080/0141192032000133686.

Vergnaud, G. (1983). Multiplicative structures. In R. Lesh & M. Landau (Eds.), *Acquisition of mathematics concepts and processes* (pp. 127–174). Academic Press.

Walport, M., Goodfellow, J., McLoughlin, F., Post, M., Sjøvoll, J., Taylor, M., & Waboso, D. (2010). *Science and mathematics secondary education for the 21st century: Report of the Science and Learning Expert Group*. Department for Business, Innovation and Skills.

Warnock, M. (1978). *The Warnock Report (1978): Special Educational Needs. Report by the Committee of Enquiry into the Education of Handicapped Children and Young People*. HMSO.

Wheater, R., Ager, R., Burge, B., & Sizmur, J. (2013). *Achievement of 15-year-olds in England: PISA 2012 National Report* (*OECD Programme for Student Assessment*) Department for Education.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: problems and potential solutions. *Educational Assessment*, *10*. https://doi.org/10.1207/s15326977ea1001_1.