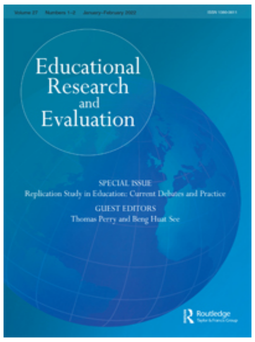

Educational Research and Evaluation

An International Journal on Theory and Practice

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/nere20

# A conceptual replication study of a self-affirmation intervention to improve the academic achievement of low-income pupils in England

Beng Huat See, Rebecca Morris, Stephen Gorard, Nadia Siddiqui, Matthew J. Easterbrook, Marlon Nieuwenhuis, Kerry Fox, Peter R. Harris & Robin Banerjee

**To cite this article:** Beng Huat See, Rebecca Morris, Stephen Gorard, Nadia Siddiqui, Matthew J. Easterbrook, Marlon Nieuwenhuis, Kerry Fox, Peter R. Harris & Robin Banerjee (2022) A conceptual replication study of a self-affirmation intervention to improve the academic achievement of low-income pupils in England, Educational Research and Evaluation, 27:1-2, 83-116, DOI: 10.1080/13803611.2021.2022317

**To link to this article:** https://doi.org/10.1080/13803611.2021.2022317

EDUCATIONAL RESEARCH AND EVALUATION
2022, VOL. 27, NOS. 1–2, 83–116
https://doi.org/10.1080/13803611.2021.2022317

Routledge
Taylor & Francis Group

OPEN ACCESS

# A conceptual replication study of a self-affirmation intervention to improve the academic achievement of low-income pupils in England

Beng Huat See [a], Rebecca Morris [b], Stephen Gorard [a], Nadia Siddiqui [a], Matthew J. Easterbrook [c], Marlon Nieuwenhuis [c,d], Kerry Fox [c,e], Peter R. Harris [c] and Robin Banerjee [c]

[a]School of Education, Durham University, Durham, UK; [b]Centre for Education Studies, University of Warwick, Coventry, UK; [c]School of Psychology, University of Sussex, Falmer, UK; [d]Faculty of Behavioural, Management and Social Science, University of Twente, Enschede, the Netherlands; [e]School of Humanities and Social Science, University of Brighton, Brighton, UK

**ABSTRACT**

This paper describes an independently evaluated randomised controlled trial of a self-affirmation intervention, replicating earlier studies, mostly conducted in the US with ethnic minority students. Self-affirmation theory suggests that some stigmatised groups, such as those from ethnic minority or poor families, face stereotype threats which undermine their academic performance. Engaging in value affirmation writing activities when such threats are most salient can give individuals a positive sense of value, negating harmful feelings, and fostering academic learning. The present study, involving 10,807 pupils aged 14 to 16 in England showed that the intervention can be successfully replicated with children from low socioeconomic backgrounds in England. The analysis showed positive effects for the intervention group. Pupils who completed more exercises also performed better. The findings are worth consideration given that it costs virtually nothing and does no harm.

**KEYWORDS**

Randomised controlled trial; self-affirmation; conceptual replication; stereotype threat

## Introduction

Replications are important for validating the trustworthiness of scientific findings, but they are rare in education, and while there has been an increase in the number of randomised controlled trials and meta-analyses in education, few have been fully replicated. In the UK, the Education Endowment Foundation (EEF), a What Works centre, has commissioned over 160 trials (10% of all known trials in education around the world since its inception in 2011 (EEF, 2018).

**CONTACT** Beng Huat See b.h.see@durham.ac.uk

Although some of the more promising interventions have been scaled up from efficacy trials to effectiveness trials, few of these are direct or even conceptual replications. In the US, Slavin (2018) noted that half of all programmes in the What Works Clearinghouse are single evaluations. Relying on the evidence of single studies to accept or reject a programme is premature (Morrison, 2019). Replications of single studies are needed to corroborate the initial findings, to overcome possible bias and errors in the original research, and to confirm and/or extend generalisability (Johnston & Pennypacker, 2009; Travers et al., 2016) to other contexts and populations. This is especially important for programmes that report positive results and have the potential to benefit pupil outcomes.

This new paper reports a conceptual replication of previous studies conducted in the US (Cohen et al., 2006; Sherman et al., 2013) to test the impact of self-affirming values on the academic attainment of low-income students in England. The new study replicates the conditions in terms of implementation and delivery as described in the initial studies from the US. It attempts to keep almost all the known conditions (outlined by Sherman et al., 2013) the same in terms of the timing, setting, and "stealth" with which the intervention is delivered. In this respect, this is a conceptual rather than a direct replication (Hunter, 2001; Morrison, 2019). Such conceptual replications are useful in addressing generalisability (Earp & Trafimow, 2015; Makel & Plucker, 2014; Morrison, 2019).

## Background

Closing the attainment gap between rich and poor students is a policy issue relevant to many education systems in the world. The relationship between socioeconomic status (SES) and students' academic attainment is well-established. Young people's PISA (Programme for International Student Assessment) outcomes for maths, science, and reading can be predicted to a considerable extent by their SES (Organisation for Economic Co-operation and Development, 2019). In England, an ambitious policy initiative was introduced in 2011, which gave schools an initial £625 million of extra funding to close the attainment gaps for disadvantaged children (Gorard et al., 2021). This increased to £2.45 billion in the 2015–2016 financial year (Department for Education [DfE], 2015). Alongside this Pupil Premium (PP) funding was the establishment of the EEF to evaluate and identify promising programmes that can raise the attainment of the poorest children. These linked approaches were one of the most important recent developments in education in England. The PP funding is for schools to use for programmes or interventions to support the academic development of disadvantaged children (mainly children who are eligible for free school meals, but also those who had been in care or with parents in the armed forces). In England, children are eligible for free school meals if they live in households on income-related benefits, such as Universal Credit, Child Tax

Credit, income-related employment and support allowance, income support, jobseeker's allowance, and asylum seeker support.

Evidence from studies conducted largely in the US has suggested that self-affirmation interventions can have positive and long-term results improving academic achievement, especially of those from ethnic minority backgrounds (Cohen et al., 2009; Good et al., 2003; Miyake et al., 2010; Oyserman et al., 2006; Sherman & Cohen, 2006; Steele, 1988; Wu et al., 2021). More recent studies showed that these effects persist through high school (Borman et al., 2018, 2021) and right up to college (Goyer et al., 2017). Treatment students were more likely to enrol in college and competitive colleges than their untreated peers.

In light of such findings, the EEF funded an evaluation of self-affirmation writing exercises, aimed at improving the academic attainment of disadvantaged students at Key Stage 4 (KS4), principally the GCSE (General Certificate of Secondary Education) assessment. GCSE is a national standardised exam taken by 15/16-year-olds at the end of their compulsory secondary education. The purpose of the trial was to see if similar results were produced for students in England who are eligible for free school meals (FSM), an indicator of low SES. The theory suggests that the intervention is effective only for groups that experience stereotype threat. The focus of this evaluation is to replicate the conditions of the implementation of the self-affirmation intervention as used in the original studies by Cohen et al. (2006, 2009).

According to the notion of "stereotype threat", students from some potentially stigmatised groups (e.g., students from disadvantaged backgrounds) are aware of the negative stereotype people have of them regarding their academic performance (Steele, 1997). This can (a) lead to anxiety about confirming this negative stereotype during school assessments, which can undermine performance, or (b) elicit a defence mechanism, known as "disidentification", in order to protect the self-concept from being devalued by the negative stereotype (S. J. Spencer et al., 2016). Disidentification results in academic achievement being discounted or devalued (Woodcock et al., 2012), and can reduce learning and motivation.

Self-affirming activities, such as writing positive statements about the values that are important to oneself, are believed to help protect students' self-worth and free up cognitive resources so that they can engage more effectively with their learning (Miyake et al., 2010; Oyserman et al., 2006; Sherman & Cohen, 2006; Steele, 1988). The theory is that such writing activities reinforce pupils' sense of value, alleviating negative feelings they may have about themselves. The advantage of this approach is that no stigma is attached to individual pupils and the cost of delivery is minimal, apart from the initial training of teachers and the costs of printing any exercise booklets and teacher manuals. If this approach is found to be effective in raising attainment for disadvantaged children, it could prove to be attractive as it is almost cost-free, simple to implement, and would appear to generate few, if any, contra-indications. However, one needs to be cautious in how the intervention is implemented and to whom it is applied. There is evidence that such an approach may be counterproductive for some groups where the factors affecting their academic performance are not psychological or social, or if it is not properly implemented (Binning & Browman, 2020; Easterbrook & Hadden, 2021; Walton & Yeager, 2020).

Most of the studies conducted on this so far have been based in the US. The results are mixed but promising, and suggest that the intervention is particularly effective in raising the attainment of ethnic minority groups (Cohen et al., 2009; Cohen & Sherman, 2014; Sherman et al., 2013). Cohen et al. (2006), for example, found that although there were no overall gains in grade point averages (GPAs) across four core academic subjects in both treatment and control groups, African American students in the treatment group improved their GPA score by 0.24 points, and the low-achieving African-American students by 0.41. The intervention also appeared to reduce the likelihood of grade retention for lower achieving African American students. A longitudinal experiment (Sherman et al., 2013) showed that a self-affirmation intervention also benefitted Latino American students. Borman et al. (2016) also reported a positive impact on minority pupils' standardised maths test scores, while Mikaye et al. (2010) showed the self-affirmation can help close the gender attainment gap.

However, other studies have shown no effects on either academic or other outcomes (Bratter et al., 2016; de Jong et al., 2016; Hanselman et al., 2017; Protzko & Aronson, 2016). There are subtle differences between these less promising studies and the ones by Cohen et al. (2006, 2009). These provide hints about the delicate nature of delivering the intervention. Such tweaks in the procedures of implementation from the original study can change the replication. For example, Simmons's (2011) study showed that students trained to use the self-affirmation strategy did not do better, and were no more psychologically engaged, than the control students. One important difference was that Simmons administered the intervention after the beginning of the term, whereas Cohen and his colleagues (Cohen et al., 2006, 2009; Cohen & Sherman, 2014) typically administered the intervention very close to the start of the term before students have the opportunity to experience negative stereotype influence. Cohen et al. (2006) and Miyake et al. (2010) also administered the intervention immediately before or after a threatening event and in the regular classroom, whereas Simmons administered the intervention in a different setting from their regular lesson (e.g., a cafeteria or another room). Also, students were offered monetary incentives to complete the post-measure, and this may have affected the apparent stereotype threat for participants. These differences could be important and suggest that the intervention is not simply about writing self-affirming statements. To be effective, these activities have to be carried out immediately prior to stressful events, such as before an exam, and as routinely as possible. This suggests that it is the conditions of delivery as much as the writing exercise that is the driver.

In the study by Protzko and Aronson (2016), the writing instructions were handed to students by researchers rather than teachers. The knowledge that it was a research exercise may have altered its impact. de Jong et al. (2016) also found no effects on school attainment of migrant children in the Netherlands despite close replication of the conditions of the earlier American studies (Cohen et al., 2009; Cohen & Sherman, 2014). One explanation is the cultural context. Unlike in the US, where the ethnic minority students are largely either African American or Latino American, those in the Netherlands are of Turkish or Moroccan descent. They were likely to be Muslims and often chose "religion" as an important value to reflect on in the self-affirming activity. Writing about their religion which has attracted negative media attention may have sometimes heightened their negative stereotype rather than reduced it. de Jong et al. also implied that cultural power distance (defined as the degree to which members of society accept their position in a hierarchical society) may explain why self-affirmation intervention may not work for certain cultural groups. Moroccan and Turkish students have a relatively high power distance (Hofstede et al., 2002), making it difficult for them to believe that they can change their situation. Other studies suggest that the writing exercise alone is not enough. A supportive classroom environment is needed for the intervention to have any impact (Dee, 2015).

These studies suggest that the effectiveness of the intervention can depend on how the intervention is delivered. First, according to Sherman et al. (2013), it is important that the intervention should be seen as part of a normal classroom activity, and not billed as a stress-reduction or academic performance enhancement exercise. Awareness of the intent of the activity could reduce its efficacy. Second, the intervention should be administered at a period when identity issues pose the biggest threat to the students. In the case of low-income and low-performing students, this threat is often associated with exams when students know that they will be judged by how well they perform in the exams (Hadden et al., 2020). Therefore, intervening just before students take their exams can help to break the recursive cycle of negative self-belief. Finally, Sherman et al. stressed that it is important to consider the social and psychological context within which the intervention takes place. In some contexts (e.g., very disadvantaged schools), the stereotype threat may be less important than other structural barriers in students' academic performance, and in other contexts the threat may contribute less to the performance of ethnic minority groups than to other groups. Therefore, depending on the context, intervening with self-affirming values may not work as well.

Most of the studies cited were conducted in the US and focused on African American or Latino American students, and the results for White students were less promising (Cohen et al., 2006; Sherman et al., 2013). Although there is often a close relationship between SES and ethnicity, this link is perhaps weaker in England than in the US. Where academic disparities exist in the UK, it is more often examined along the lines of SES. The underachievement of White working class boys in England, for example, is well documented (e.g., Demie & Lewis, 2011; Strand, 2014). In the UK, for the 2019 GCSE cohort, 24.7% of disadvantaged pupils achieved at least a grade 5 (considered a standard pass) in English and mathematics compared to 49.9% for non-disadvantaged pupils, a difference of 25.2 percentage points. The difference between the lowest performing major ethnic group (Black students) and the White majority students, on the other hand, was only 4.6% percentage points (DfE, 2019).

Although the interactions between SES and ethnicity are closely related, both in the US and UK, in the UK, SES is a stronger predictor of attainment than ethnicity at all levels at school (Brannon et al., 2017; DfE, 2019; Harackiewicz et al., 2016; Strand, 2014). Hadden et al. (2020) suggest that an important factor for this might be the stereotype threat. There is empirical evidence suggesting that early failure, such as experienced by low-SES children, can influence students' future performance and psychological state (Cohen et al., 2009; Fell & Hewstone, 2015). A review of psychology research reported that studies show "self-stereotyping" effects, in which the individual's perception of their membership of a particular group can influence their self-evaluation and psychological performance (Fell & Hewstone, 2015). Teachers' expectations of their students' academic potential can also result in self-fulfilling prophecies of students' academic performance (Rosenthal & Jacobson, 1968), known as the "Pygmalion Effect". And since students from low-income families tend to perform less well (on average) than their more privileged peers, teachers may underestimate their abilities. These are not conscious acts, but they can perpetuate students' feeling of inadequacy. A number of studies in the US have examined stereotype threat among students of different SES groups (e.g., Croizet & Claire, 1998; B. Spencer & Castano, 2007). Using experiments, researchers compared the test outcomes of low- and high-SES students when tests are presented as being diagnostic of intelligence. They found that low-SES students performed as well as other students when the test was not presented as a measure of their intellectual ability. This finding suggests that the disparities in performance between high and low SES observed in other studies might actually be the result of stereotype threat for low-SES students.

An earlier study conducted in England (Hadden et al., 2020) using a randomised controlled design involving 562 pupils showed that the self-affirmation approach works in raising the attainment of low-SES pupils, reducing the attainment gap by 62%. Our present study was the first large-scale independently evaluated randomised controlled trial of the self-affirmation theory conducted in the UK, replicating the conditions used in the earlier studies in the US in its implementation and delivery. The aim was to test if the intervention also benefits poor children in England who may be stigmatised by their group's low academic performance, by simulating the exact conditions of administration immediately prior to exams. The following section describes the intervention as was implemented in our study.

## The intervention

Throughout the trial, the intervention was referred to by its pseudonym, "Writing About Values" (WaV), to help mask the nature of the intervention somewhat. This is an important element of the intervention, as previous research has shown that knowledge of the purpose of the intervention can interfere with its efficacy (Yeager & Walton, 2011). Therefore, every effort was made to keep the primary intention of the writing activity from both teachers and students (as agreed with the ethics panel).

The intervention was proposed, developed, and implemented by a team of social and developmental psychologists at the University of Sussex, who adapted the workbooks, training materials, and teacher instruction sheets from those previously used in the US (e.g., Sherman et al., 2013). The evaluation was conducted by independent evaluators from Durham University using national assessments at KS2 (Key Stage 2 – a test taken at the end of primary school) and KS4 (Key Stage 4 – a test taken at the end of secondary education). While efforts were made to keep the intervention as close to those used in the US by Borman et al. (2016, 2018, 2021), Sherman et al. (2013), and Cohen et al. (2006, 2009), there were some slight variations.

The intervention comprised three writing activities, each lasting 10 to 15 min, in which all students wrote short essays during their regular English lessons. These writing exercises were presented in booklets that were placed in named envelopes and distributed to pupils individually. For the first writing task, the treatment group within each class wrote about values that were important to them, such as friendships and honesty. A list of values was provided for the pupils from which they could choose two or three to write about. Examples of such values included enjoying sports, being honest, and relationships with friends. The control pupils, on the other hand, wrote about values that might be important to other people. For the second writing activity, the treatment pupils wrote about things/people that mattered to them, while the control pupils wrote about things they did that morning. In the third writing exercise, treatment pupils selected from a list of values those that were important to them and were asked to write about what they would do to show that these were important to them, or how much they enjoyed doing them. These values could be about relationships with friends, having a sense of humour, being with family, and following government and politics. An example of the writing task for the treatment and control groups is available in Appendix 1 and Appendix 2.

Short scripted instructions (Appendix 3) were provided for teachers to introduce the task and to explain to pupils that the purpose of the activity was to get them to write freely. It was suggested that teachers say something like: "We will check to see if you've engaged with it properly, but it won't be marked. It is the process of writing about your own thoughts and ideas rather than me providing feedback. The exercises will be stored away". Pupils were told that there were no right or wrong answers, and that they did not need to focus on grammar or spelling; content and ideas were more important. Instructions were also available on the booklets. Pupils were encouraged not to talk to each other or look at their neighbours' writing task while writing. Teachers were given strict instructions to use the prescribed answers to pupils' queries about the purpose of the exercises.

These exercises were delivered by English-language teachers as a whole-class activity, as part of their regular English lessons, and collected in the envelopes by the teachers at the end of the session. Efforts were made to ensure that these exercises were delivered as naturally as possible, to avoid pupils linking them to a research project. The researchers who conducted a light observation of the delivery of the intervention in some classes were blind to treatment conditions, in that they would not know which were treatment or control pupils, as all the pupils were involved in a writing exercise that differed only in terms of content seen only by themselves.

In line with the theory of self-affirmation, the writing exercises were delivered during three crucial time points: once at the beginning of the academic year (before the experience of negative stereotype was established), and again before potentially stressful events: In this study these were the mock GCSEs and the actual GCSEs examinations later in the year. The timeline of the three writing activities is presented in Table 1. The original Cohen et al. (2006) paper used three to five exercises, Sherman et al. (2013) used four or five, and Borman et al. (2018) did "up to four" exercises. We opted to implement the interventions at the most obvious high stress times – the beginning of the year, before mock exams, and before exams – which is when the interventions are supposed to be most effective, while keeping the administrative load on schools to a minimum. Ultimately, the number of doses does not impact the effectiveness of the intervention (Wu et al., 2021). It is the timing of the intervention that is crucial.

**Table 1.** Timeline of delivery of writing activities.

| Date | Activity |
|---|---|
| **5 September 2016** | Delivery of first writing task |
| **December 2016/January 2017** | Delivery of second writing exercise before mock GCSEs |
| **May 2017** | Delivery of third writing exercise before GCSEs |

Note: GCSE = General Certificate of Secondary Education.

The baseline survey was conducted in the first two weeks of the term. The aim was for all exercises to be completed within the following two weeks of the term. More than two thirds of pupils completed them between September 19 and September 23 (Table 1), and just 5% completed them within the first week of October. Because the exercises are theorised to be most effective when they are delivered before a stressful event, we tied the second exercises to be implemented within the two weeks before the start of each school's mock GCSEs rather than a specific timeframe, and this varied across schools. Ninety percent of schools completed the second exercise before February 23. For the final exercises, the aim was to deliver them in the two weeks before the start of the GCSE examinations. Eighty percent of pupils completed them between April 24 and May 24, and 97% by the end of June.

To protect the integrity of the intervention, a number of strategies were employed to safeguard the precise nature and purpose of the intervention. Observation visits, for example, were kept to a minimum to avoid pupils linking it to a research project, and no schools were visited for both observation of the delivery of the exercises and the administration of the survey questionnaire. The presence of evaluators in the classroom was explained as part of a programme to observe how English was being taught. Feedback from pupils about the intervention was obtained only from the Year 11 pupils and only after their KS4 exam. Interviews with teachers were conducted only after the third writing exercise and only in very general terms about the writing activity itself, and not about the specifics or theory of the intervention. The briefings to teachers were presented in a very general way, and, although occasional reference was made to the evidence-based nature of the intervention, teachers were not given the detailed background of the intervention. Instead, the briefing focused on the delivery of the exercises, and how teachers should ensure that pupils were not aware that they were taking part in a research project. All material was available after the post-intervention tests.

### *Ethics*

There were no unusual ethical issues raised with the trial because the intervention was introduced by teachers voluntarily as part of normal classroom activity. The ethics committee at Durham University and the University of Sussex agreed that the obscuring of the precise nature of the intervention was justified by its intent, with only an extremely small chance of harm (based on prior studies). All pupils took part, and all wrote about values (with only the precise nature of the task varying). This writing about values was how the intervention was explained and introduced.

## Methods used in the evaluation

This was a 2-year, double-blind randomised controlled efficacy trial involving Year 10 (Y10; age 14–15) and Year 11 (Y11; age 15–16) pupils from 29 schools in the South East of England. Pupils (and their teachers) had no knowledge of whether they were in the treatment or control group. This was possible because control groups were given a writing activity as a placebo, which was also a viable alternative approach to writing about values. Teachers were not told what the intervention was, and students were not aware that it was a research activity.

Due to changes in the processes of access to the National Pupil Database (NPD) in the 2nd year of the trial, data from both cohorts could not be merged. The two cohorts of pupils were therefore evaluated, and their results are presented, separately. Analysis for the Y11 cohort was undertaken at the end of the year while that for the Y10 cohort was taken a year later when they sat for their KS4 assessments. The inclusion of Y10 allowed us to test the longer term impact of the intervention.

### *Research questions*

To assess the impact of the self-affirmation intervention, we posed the following research questions:

(1) What impact does the self-affirmation intervention have on the academic attainment of disadvantaged pupils, defined as those who were eligible for free school meals at some point in the last 6 years (EverFSM6), after 1 year of treatment? (This replicates the earlier studies from the US, but applies the intervention to disadvantaged pupils in England rather than potentially stigmatised groups, for example, ethnic minority pupils in the US.)
(2) Is the impact for EverFSM6 pupils sustained after 2 years (1 year after the end of the intervention)?
(3) Does the self-affirmation intervention have any impact on the general pupil population (including not EverFSM6)?
(4) Is the impact for all pupils (EverFSM6 and non-EverFSM6) sustained after 2 years?
(5) Is post hoc conceptual replication of the self-affirmation intervention feasible in the English context with low-SES pupils?

EverFSM6 was used as a measure of socioeconomic disadvantage because this was the definition underlying the distribution of Pupil Premium funding to schools.

### *Sample*

The trial was conducted in 29 secondary schools across the South East of England with a total of 5,619 Y11 and 5,188 Y10 pupils. The schools recruited were those not in "special measures" (i.e., at risk of failure), and which had a minimum of 10% of pupil population eligible for FSM. Pupils were individually randomised within schools, stratified first by year and then by FSM status to either the treatment or control conditions. This was to help ensure initial equivalence between the two groups.

Table 2 details the number of participants at randomisation and subsequently. The key figures are for the headline based on EverFSM6 pupils. There was no attrition from Year 10 pupils, with pre-intervention scores, and just under 10% from Year 11 pupils with pre-intervention scores. Cases were only missing if they could not be found on the NPD, and so it did not matter if they moved to another school in England after the randomisation. As this is an intention-to-treat design, all pupils in the original design were included in the analysis even if they were no longer in one of the 29 schools in the trial. It was not clear why a number of Year 11 pupils could not be found by the DfE. One possible reason would be mistakes in the unique identifiers provided by the schools, or pupils who had just arrived in the country and not been given an identifier. Pupils who had received their primary education overseas or in an independent school would also not have taken the KS2 exams. For the Y10 cohort, we excluded all pupils without KS2 scores. KS2 is a national exam that pupils take at the end of their primary school. We used KS2 test scores as the baseline assessment.

This sample size of 1,380 (EverFSM6) individually randomised cases in the intervention groups (the smallest cells) is traditionally large enough to detect an effect size of just over +0.1 (which is small for a typical education intervention). However, we do not use traditional power calculations as these are based on an erroneous assumption (Gorard et al., 2017). Instead, we calculated the sample size needed for any "effect" size to be considered secure by considering *a priori* the number of "counterfactual" cases needed to disturb a finding (Gorard, 2018). This number needed to disturb (NNTD) is the "effect" size multiplied by the number of cases in the smallest group in the comparison (i.e., the number of cases included in either the control or treatment group, whichever is smaller). Therefore, the smallest detectable effect size for any NNTD is NNTD divided by the size of the smallest group. Based on Gorard (2018), NNTD of 50 can be considered a very strong and secure finding. Using this as

**Table 2.** Participants by year group and treatment conditions ($N = 29$ schools).

| | Intervention pupils | Control pupils |
|---|---|---|
| Randomisation Year 10 all | 2,569 | 2,619 |
| Randomisation Year 11 all | 2,809 | 2,810 |
| Randomisation Year 10 EverFSM6 | 674 | 698 |
| Randomisation Year 11 EverFSM6 | 706 | 800 |
| Analysed Year 10 EverFSM6 | 674 | 698 |
| Analysed Year 11 EverFSM6 | 640 | 711 |

Note: EverFSM6 = pupils eligible for free school meals at some point in the last 6 years.

a working assumption, a sample of 1,380 might enable us to detect an effect size as little as 0.04 with considerable confidence.

### *Outcome measures*

To test the effect of the intervention on the academic attainment of disadvantaged groups of pupils, we used the KS4 Attainment 8 scores of pupils who were eligible for FSM at any point in the last 6 years (EverFSM6). Attainment 8 is used in England as a measure of students' academic performance in the last year of their compulsory secondary education. It is the student's average grade across their "best" eight subjects. Evaluation of impact for Year 11 was undertaken at the end of the first year following release of their results, whereas impact evaluation for the Year 10 cohort was completed a year later. This allowed us to see if the effect (if any) was maintained 1 year after the intervention.

The KS2 results for maths and English (national tests taken at the end of primary school) were used as a pre-test measure of pre-intervention equivalence. Impact analysis was based on the combined reading and English scores as per pre-trial protocol, agreed with the funder.

The attitudes theory suggests that the intervention is effective only for groups that experienced stereotype threat (e.g., pupils from disadvantaged backgrounds). To test this theory, we also analysed the attainment outcomes of the general pupil population. This includes both EverFSM6 and not EverFSM6 pupils.

### *Analysis*

Pupil attainment was analysed using intention-to-treat. This means that all pupils randomised to receive the intervention were included in the analysis regardless of whether they were known to receive the intervention or not. The impact of the intervention was measured as the difference between intervention and control groups in terms of the progress scores between average KS2 results for maths and reading and KS4 Attainment 8 outcomes. For comparability between phases, the test scores were converted to standardised *z* scores. The pre- and post-intervention differences are expressed as simple effect sizes (difference between means divided by their overall standard deviation). The advantage of using progress scores is that it addresses any initial imbalance in prior attainment created inadvertently by the randomisation. Significance tests and confidence intervals are not reported here as they are not relevant, because attrition means that the cases are no longer completely randomised. And, even with complete randomisation, such tests are still not appropriate because null hypothesis significant testing (NHST) states that assuming there is no difference between groups questions how likely are we to obtain the data to be as extreme as observed. The answer that most researchers want, is to the question: "How likely is it that there is a difference between groups, given the data?" Unfortunately, significant tests do not and cannot answer this question. Using such tests to judge if there is a difference between groups is therefore misleading. For further explanations, see Gorard (2021), Colquhoun (2014, 2016), Perezgonzalez (2015), and Pharoah et al. (2017).

There should be no issue of clustering, as randomisation was at the individual level within schools rather than at the school level. Analysis is of all pupils in the two groups and not by schools. The mean scores of all the pupils in the control group and treatment group in all schools or if computed class by class, would be the same as the mean scores of all treatment and control pupils in the whole trial, by definition. The issue of clustering has no relevance to the effect sizes presented in the paper. These remain exactly the same if they are computed for individuals, classes, schools, or subgroups like FSM-eligible, first, and then aggregated to treatment group, or as here just by computing for treatment groups. The latter is an easier way to compute them and easier for readers to follow. The idea of standard errors, like significance tests, is not used in this paper (see below). Standard errors would only apply to repeated samples of fully randomised cases. With no standard errors, there is no chance of them being underestimated due to clustering. There will be no difference in the effect sizes.

To account for missing cases or missing data, which can potentially bias the results (Dong & Lipsey, 2011; Foster & Fang, 2004; Little & Rubin, 1987; Puma et al., 2009; Shadish et al., 2001), we presented differences in pre-test scores (KS2 maths and reading) between cases dropping out from both groups (where these were available). Actually, this was not dropout as such; rather, it was that the DfE did not find later results in the NPD (see above). In addition, we also estimated how much these missing cases would skew the results if they were included. To do this, we first calculated the number of counterfactual cases needed to disturb the headline finding (NNTD, as above). The number of counterfactual cases determines whether the number of missing cases is large enough to alter/explain the findings (see the subsection on sample size above). The bigger this number is, the more stable is the substantive result, as this means it will take this number of counterfactual cases to reduce the effect size to zero.

### *Dosage and complier analysis*

Since not all pupils completed the three writing activities, we carried out two further analyses to test the impact of dosage. The first was a correlational analysis, comparing the outcomes of pupils with the number of exercises completed. This would be zero for all cases in the control group. Information about dosage was collected by the project delivery team who kept a log of the number of exercises completed by each pupil.

Further analysis was carried out to estimate the effects for the subgroup of treatment students who complied with their treatment assignment using the complier average causal effect (CACE) analysis. Compliance is defined as completion of the first writing exercise (according to the developers) because theoretically the first writing exercise is supposed to have the most impact (Cohen & Sherman, 2014; Garcia & Cohen, 2012) and is expected to trigger a recursive adaptive response to a threatening environment in a feedback loop. For example, if a student performs/behaves better as a result of the first activity, their self-confidence may improve, and their teacher may have higher expectations of them. This could lead to better performance, and the process perpetuates itself. The second and third exercises are meant to provide the boost to this process. It is more difficult to trigger a positive response later in the year once expectations set in. Therefore, it is important that pupils complete the first writing exercise.

CACE compares the average outcome of treatment pupils who complied with the control pupils who it is estimated would have complied if given the treatment (Dunn et al., 2005; Nicholl, n.d.). Table 3 illustrates how CACE is estimated. Given that we know the overall results for both groups (Cells F and K) and the mean scores for those in the treatment group who complied and who did not comply (Cells A to D), we can calculate the average outcome for those in the control group who would have complied if given the treatment (*x*). We assume that, because of randomisation, the proportion of compliers in both arms of the trial would be the same (on average), and the average outcome for those in the control group who did not comply (I) will be the same as the outcome of non-compliers in the treatment group (D). They are unaffected by the intervention.

The proportion in treatment group who complied will be A/E. The number who complied in the control group (Cell G) will be A/E*J. The number of non-compliers in the control group (Cell H) will be J-G. The average outcome for compliers in the control group (*x*) is thus ((J*K) − (H*I))/G.

## Process evaluation

We also carried out a light touch process evaluation to collect information about teachers' delivery of the intervention, staff and students' views of the intervention, and indications of any possible contamination or diffusion. This is not the

**Table 3.** Estimation of complier average causal effect.

| | Compliers | | Non-compliers | | Overall | |
|---|---|---|---|---|---|---|
| | *N* who completed first writing exercise | Mean | *N* who did not complete first writing exercise | Mean | *N* | Mean |
| Treatment | A | B | C | D | E | F |
| Control | G | *x* | H | I | J | K |

focus of this paper. The main method of data collection was classroom observations. These were as integrated and non-intrusive as possible, in order to minimise disruptions to classroom activities. The classroom observations were designed to see whether teachers kept to the scripts, the extent to which they adhered to the instructions for delivering the exercises, that the right pupils were given the correct writing exercise, and if there was any possibility that pupils could swap exercises with their classmates. Observation visits were made to classes in five schools. The number of visits was deliberately kept small to avoid alerting pupils to the research element of the writing activity.

We also had a number of informal conversations with teachers in schools to find out whether they had observed any changes in pupil behaviour and to gather their views on the writing exercises. However, because of the nature of the intervention and the restrictions in what teachers knew about the overall project, these conversations were limited in scope and focused predominantly on teachers' views of the writing tasks and the children's reactions to them. The process evaluation was also intended to find out indirectly if teachers and pupils had any knowledge of the intervention.

Due to the nature of the research, pupils were not interviewed while the trial was still running. A small number of Y11 pupils were contacted by their teachers after their GCSE exams via emails, through their parents inviting them to respond to a short questionnaire asking for their views on the writing activity.

## Impact results

To evaluate the impact of self-affirmation on the academic outcomes of EverFSM6 pupils, we compared the gain scores for the control and treatment pupils between KS2 and KS4 Attainment 8 for EverFSM6 pupils only. Since KS2 scores and Attainment 8 scores are not on the same metric, for fair comparability, we converted all to standardised *z* scores before analysis. The negative scores show that FSM pupils in general performed below the average for their cohort. Analysis was performed for pupils who had both pre-test scores for reading and maths and post-test scores.

Table 4 shows the number of pupils with pre-test scores and the number of missing pre-test scores in both intervention and control group. A total of 10,807 pupils (Year 11 = 5,619; Year 10 = 5,188) had post-test scores, but some were missing pre-test KS2 maths and some were missing KS2 reading scores. Impact analysis was conducted for 10,274 pupils (Y11 = 5,086; Y10 = 5,169) who had both KS2 maths and KS2 reading scores. The table shows that there is a very small difference between the two groups at pre-test, with the treatment group slightly ahead. Therefore, the gain score results are used for the headline findings.

**Table 4.** Comparison of pupils' baseline characteristics.

| Characteristics of pupils at randomisation | Intervention | | Control | | |
|---|---|---|---|---|---|
| | *n*/*N* (missing) | Percentage | *n*/*N* (missing) | Percentage | |
| Proportion of boys | | | | | |
| Year 11 | 1,430/2,823 (0) | 50.7 | 1,393/2,823 (0) | 49.3 | |
| Year 10 | 1,225/2,521 (0) | 49.0 | 1,296/2,521 (0) | 51.0 | |
| Proportion of pupils eligible for EverFSM6 | | | | | |
| Year 11 | 640/1,351 (0) | 47.4 | 711/1,351 (0) | 52.6 | |
| Year 10 | 674/1,372 (0) | 49.0 | 698/1,372 (0) | 51.0 | |
| Proportion of current FSM pupils | | | | | |
| Year 11 | 305/678 (15) | 45.0 | 373/678 (21) | 55.0 | |
| Year 10 | 307/635 (22) | 48.0 | 328/635 (25) | 52.0 | |
| Proportion of pupils with SEN | | | | | |
| Year 11 | 421/843 (0) | 49.9 | 422/843 (0) | 50.1 | |
| Year 10 | 401/806 (22) | 50.0 | 405/806 (25) | 50.0 | |
| Proportion of pupils whose first language is not English | | | | | |
| Year 11 | 314/652 (2) | 48.2 | 338/652 (9) | 51.8 | |
| Year 10 | 229/466 (14) | 49.0 | 237/466 (16) | 51.0 | |
| | | Raw means | | | |
| | Intervention | | Control | | |
| | *n*/*N* (missing) | Mean (*SD*) | *n*/*N* (missing) | Mean (*SD*) | Effect size |
| KS2 Maths (point scores) | | | | | |
| Year 11 | 2,621/5,619 (189) | 4.65 (0.79) | 2,620/5,619 (189) | 4.63 (0.80) | +0.03 |
| Year 10 | 2,569/5,188 (0) | 4.68 (0.78) | 2,619/5,188 (0) | 4.68 (0.81) | 0.00 |
| KS2 Reading (point scores) | | | | | |
| Year 11 | 2,547/5,619 (263) | 4.66 (0.74) | 2,539/5,619 (270) | 4.64 (0.78) | +0.03 |
| Year 10 | 2,569/5,188 (0) | 4.70 (0.77) | 2,619/5,188 (0) | 4.69 (0.79) | +0.01 |
| KS2 Reading and Maths combined | | | | | |
| Year 11 | 2,547/5,619 (263) | 9.31 (1.41) | 2,539/5,619 (270) | 9.28 (1.45) | +0.03 |
| Year 10 | 2,569/5,188 (0) | 9.38 (1.42) | 2,619/5,188 (0) | 9.37 (1.47) | +0.01 |
| EverFSM6 pupils | | | | | |
| KS2 Maths (point scores) | | | | | |
| Year 11 | 679/1,431 (27) | 4.45 (0.70) | 752/1,431 (48) | 4.50 (0.75) | −0.07 |
| Year 10 | 674/1,372 (0) | 4.40 (0.76) | 698/1,372/ (0) | 4.43 (0.80) | −0.04 |
| KS2 Reading (point scores) | | | | | |
| Year 11 | 640/1,351 (66) | 4.48 (0.69) | 711/1,351 (89) | 4.52 (0.68) | −0.05 |
| Year 10 | 674/1,372 (0) | 4.44 (0.83) | 698/1,372/ (0) | 4.44 (0.83) | 0.00 |
| KS2 Reading and Maths combined | | | | | |
| Year 11 | 640/1,351 (66) | 8.92 (1.26) | 711/1,351 (89) | 9.02 (1.31) | −0.11 |
| Year 10 | 674/1,372 (0) | 8.84 (1.45) | 698/1,372/ (0) | 8.85 (1.47) | −0.01 |
| FSM pupils | | | | | |
| KS2 Maths (point scores) | | | | | |
| Year 11 | 290/678 (15) | 4.31 (0.79) | 357/678 (16) | 4.32 (0.79) | −0.01 |
| Year 10 | 307/635 (0) | 4.32 (0.74) | 328/635 (0) | 4.32 (0.81) | 0.00 |
| KS2 Reading (point scores) | | | | | |
| Year 11 | 271/678 (34) | 4.35 (0.80) | 333/678 (40) | 4.33 (0.78) | +0.03 |
| Year 10 | 307/635 (0) | 4.39 (0.85) | 328/635 (0) | 4.33 (0.88) | +0.07 |
| KS2 Reading and Maths combined | | | | | |
| Year 11 | 271/678 (34) | 8.66 (1.47) | 333/678 (40) | 8.65 (1.46) | +0.01 |
| Year 10 | 307/635 (0) | 8.71 (1.45) | 328/635 (0) | 8.65 (1.52) | +0.04 |

Note: EverFSM6 = pupils eligible for free school meals (FSM) at some point in the last 6 years; SEN = special educational needs; KS2 = Key Stage 2.

Table 5 shows that both FSM-eligible groups made less than average progress between KS2 and KS4 (compared to the full cohort), but compared to the treatment group, the control group made even less progress. This suggests that the intervention may have a small influence in improving the performance of the EverFSM6 pupils. The effect sizes for both Year 10 and Year 11 cohorts are

**Table 5.** Comparison of pre-, post-, and standardised gain scores using Key Stage 2 (KS2) maths and KS2 reading combined as pre-test and Attainment 8 as post-test (EverFSM6 pupils only).

| | Pre-score mean | *SD* | ES | Post-score mean | *SD* | ES | Gain score | *SD* | ES |
|---|---|---|---|---|---|---|---|---|---|
| Treatment | −0.32 | 0.98 | | −0.42 | 0.93 | | −0.10 | 0.89 | |
| Year 11 | −0.37 | 1.01 | | −0.47 | 0.91 | | −0.10 | 0.85 | |
| Year 10 | | | | | | | | | |
| Control | −0.26 | 0.99 | | −0.39 | 0.94 | | −0.14 | 0.84 | |
| Year 11 | −0.36 | 1.02 | | −0.49 | 0.88 | | −0.13 | 0.84 | |
| Year 10 | | | | | | | | | |
| Overall | −0.29 | 0.99 | −0.06 | −0.41 | 0.94 | −0.03 | −0.12 | 0.87 | +0.05 |
| Year 11 | −0.37 | 1.01 | −0.01 | −0.48 | 0.89 | +0.02 | −0.11 | 0.85 | +0.04 |
| Year 10 | | | | | | | | | |

Note: EverFSM6 = pupils eligible for free school meals (FSM) at some point in the last 6 years. The scores are standardised *z* scores, not raw scores. The negative signs indicate that the students are performing worse (in relative terms) than the average for the cohort. The comparison is with the cohorts, not between the randomised groups. As expected, FSM-eligible pupils perform worse than their non-FSM peers.

positive (+0.05 for Y11, +0.04 for Y10). Even 1 year after the intervention ceased, a small positive “effect” had been sustained. Earlier field experiments (e.g., Cohen et al., 2009) suggest that these alterations in psychological states and performance provide the initial trajectory for a recursive process, and the changes in attributions and information processing it prompts can become self-reinforcing or self-sustaining over time.

We also calculated the number of counterfactual cases (i.e., number of cases with counterfactual results) that would be needed to eliminate the positive effect. For the Year 11 cohort, this number was 32 (0.05 multiplied by 640). This means that it would take approximately 32 missing cases with counterfactual scores (see the methods subsection Sample above) in the opposite direction for the findings to change. For the Year 10 cohort, the number of counterfactual cases was 27 (0.04 multiplied by 674). However, since there were no cases with pre-tests missing post-test scores, this means that the finding cannot be due to attrition, even in the worst-case scenario. Although the effects were small, they were therefore reasonably secure.

To examine whether self-affirmation had any impact on the general pupil population (not just FSM pupils), we compared the gain scores of treatment and control for all pupils. The analysis shows no differential benefit for either group, indicating that the intervention had no impact on the overall pupils’ Attainment 8 scores (−0.01 for Year 11 and 0.00 for Year 10) (Table 6). This is consistent with the theory that self-affirmation only works with pupils experiencing stereotype threat.

## *Dosage and complier analysis*

To test whether the number of exercises completed made a difference to the outcomes, we compared the number of exercises completed (dosage) with the gain scores as well as the Attainment 8 scores (as post-test only). The

**Table 6.** Comparison of pre-, post-, and standardised gain scores using Key Stage 2 (KS2) maths and KS2 reading combined as pre-test and Attainment 8 as post-test (all pupils).

| | Pre-score mean | *SD* | ES | Post-score mean | *SD* | ES | Gain score | *SD* | ES |
|---|---|---|---|---|---|---|---|---|---|
| Treatment | 0.005 | 1.005 | | 0.047 | 0.97 | | 0.042 | 0.80 | |
| Year 11 | 0.005 | 0.98 | | 0.015 | 0.99 | | 0.009 | 0.79 | |
| Year 10 | | | | | | | | | |
| Control | −0.005 | 0.995 | | 0.042 | 0.97 | | 0.047 | 0.81 | |
| Year 11 | −0.005 | 1.02 | | 0.004 | 1.00 | | 0.009 | 0.80 | |
| Year 10 | | | | | | | | | |
| Overall | 0.000 | 1.000 | +0.01 | 0.045 | 0.97 | +0.01 | 0.045 | 0.81 | −0.01 |
| Year 11 | 0.000 | 1.000 | +0.01 | 0.009 | 0.99 | +0.11 | 0.09 | 0.80 | 0.00 |
| Year 10 | | | | | | | | | |

number of exercises was treated as a continuous variable. For the control group this would be zero, as they did not complete the intended writing activity. Over 60% of the Year 10 intervention pupils completed all three writing exercises, while only 53% of the Year 11 cohort did (Table 7).

Correlation analysis shows a small positive relationship between number of exercises completed and the gains made between pre-test (KS2 scores) and post-test (Attainment 8). The results are similar whether using gain scores or Attainment 8 post scores (Table 8). The relationship is stronger for the Year 10 pupils (+0.36) than for the Year 11 pupils (+0.16). This may suggest that the lasting effect of the intervention is stronger, the more exercises that pupils complete.

For the impact evaluation, we use an intention-to-treat analysis, meaning that all those randomised to treatment are analysed as being in the treatment group, even if they did not receive the intervention. This helps to preserve the prognostic balance afforded by randomisation (McCoy, 2017) as those who adhere to the protocol differ in some ways from those who do not (Montori & Guyatt, 2001), and ensures an unbiased estimate of the efficacy of the intervention on the primary outcome at the level of adherence observed in the trial. However, in reality not all pupils who received the intervention complied with the intervention. Compliance is defined here as completion of the first writing task because it is deemed most impactful (see above). Complier analysis, therefore, is to see if pupils who complied with the intervention do better than those who did not.

We analysed the effect of compliance using the complier average causal effect analysis (CACE) based on the standardised gain scores and using the

**Table 7.** Number of exercises completed by intervention group.

| Number of exercises completed | Year 10 | Year 11 |
|---|---|---|
| 0 | 128 (5.0%) | 111 (4.4%) |
| 1 | 211 (8.2%) | 299 (11.8%) |
| 2 | 626 (24.2%) | 775 (30.4%) |
| 3 | 1,581 (61.5%) | 1,362 (53.5%) |
| Missing | 23 (0.0%) | |
| Total | 2,569 | 2,547 |

**Table 8.** Correlation between gain scores and number of exercises completed (EverFSM6 pupils only).

| | Gain scores using KS2 maths & reading combined | | GCSE Attainment 8 score | |
|---|---|---|---|---|
| Number of exercises completed | Year 11 | Year 10 | Year 11 | Year 10 |
| | +0.09 | +0.26 | +0.16 | +0.36 |

Note: EverFSM6 = pupils eligible for free school meals at some point in the last 6 years; KS2 = Key Stage 2; GCSE = General Certificate of Secondary Education.

overall standard deviation in Table 5 (0.89 for Year 11 and 0.85 for Year 10). The result shows a small positive effect size (Table 9), exactly the same as the impact evaluation result in Table 5.

## Process evaluation summary

The process evaluation was conducted primarily to ensure that the intervention was implemented with fidelity. It does not form part of the impact evaluation, although its findings may help explain the results. For example, if teachers or pupils became aware of the purpose of the intervention, or if the writing tasks were not delivered at the three crucial time points, it could affect the results. For this reason, we did not think it was necessary to do a frequency count of the number of teachers or pupils in each event mentioned below, but simply to capture the general views of staff and pupils.

### *Teacher briefings*

To ensure that the intervention was carried out as intended, and its covert nature was observed, we attended three teacher briefing sessions to understand how the intervention was explained to the teachers. These briefing sessions were hour-long meetings attended by Y10 and Y11 English-language teachers, during which the project developers presented a short introduction to the background to the project, including some reference to the evidence-

**Table 9.** Complier average causal effect based on completion of first writing task and standardised gain scores (EverFSM6).

| | Completed first writing task *N* | Mean | Did not complete first writing task *N* | Mean | Overall *N* | Mean | Effect size |
|---|---|---|---|---|---|---|---|
| Intervention | 526 | +0.01 | 114 | −0.60 | 640 | −0.10 | |
| Year 11 | 549 | −0.005 | 125 | −0.56 | 674 | −0.10 | |
| Year 10 | | | | | | | |
| Control | *583* | *−0.04* | *128* | *−0.60* | 711 | −0.14 | +0.05 |
| Year 11 | *569* | *−0.03* | *129* | *−0.56* | 698 | −0.13 | +0.04 |
| Year 10 | | | | | | | |

Note: EverFSM6 = pupils eligible for free school meals at some point in the last 6 years. The *N* in italics are based on there being the same proportion of compliers in the control group as in the treatment group, and the mean scores in italics are based on the non-compliers in the control group having the same mean as those in the treatment group.

based nature of the intervention and the success of similar trials in America. This information was provided in a very general way, so that teachers were not made aware of the full background to the intervention and the current project aims. The majority of teachers at the briefings seemed satisfied with the introduction to the trial that they were given. One teacher, however, did question the premise of the research, to which the project team simply repeated information from the introduction, thereby revealing no extra information. The focus of the briefing session was to explain to teachers how they should go about delivering the writing exercises in English classes. Teachers were clear about the task and the need to ensure that named envelopes were given to the right pupils.

### Classroom observations

We observed a total of 10 classes across five schools for the first writing exercise, which was delivered at the start of the academic year, one class for the second exercise, and a further five groups were observed across another two schools for the final writing exercise. The number of classes and schools we could observe depended on school's availability.

No evidence of diffusion in terms of pupils' swapping writing tasks was observed. In schools where we could not observe, teachers reported no issues with the administration of the writing task. However, there was one school where the pupils became suspicious, with one pupil saying, "this is so random and confusing, it's a conspiracy". Another questioned whether it was "some kind of social experiment". Other pupils noted that they had different questions from those of their neighbours. Although they were told to work individually, some students were observed talking to their peers. The "secret envelopes" also aroused some suspicion. One pupil commented that it was "very dodgy" (e.g., dishonest, unreliable), and there was vigorous questioning about who was going to read their work. The class teacher, an experienced head of department, stuck closely to the guidance provided and emphasised the whole-school nature of the project, and that this was something that other schools were doing as well. For details about the standard response that teachers were told to use, see Appendix 3.

One thing that most concerned pupils was whether their work would be marked and who was going to look at their work. Similar concerns were noted in all the schools we visited. Many were puzzled as to why their work would not be marked and whether they would receive feedback on their writing. In all such situations, the teachers explained to the pupils that their work would not be graded and no one would be looking at what they had written, and their writing would simply be stored away. The issue of spelling, punctuation, and grammar was also a source of discussion. Pupils had been informed, via the instructions, that they did not need to worry about technical accuracy in their writing, but that they should focus on content instead. This appeared to be contradictory to the usual advice they received from their English-language teachers, particularly in the lead-up to GCSEs, where they were preparing to be assessed on these skills. In all cases, the teachers responded to these queries as set out in the instructions from the sheet provided by the developers (Appendix 3). Teachers were not able to betray the true intention of the intervention because they had also not been told.

In a few instances, teachers attempted to link the writing exercise to the topic being studied in that term, so the writing activity seemed like a natural part of their curriculum. One teacher, for example, adeptly fitted in the writing exercise into the exam preparation on the play "An Inspector Calls" by explaining that they were now going to think about the values being displayed by some of the lead characters in the text.

### *Teachers' views*

There was also no evidence that teachers had knowledge of the real purpose of the intervention. Teachers generally believed that the intervention gave pupils the opportunity to write freely, without fear of making mistakes. When asked what they thought of the writing task, almost all the teachers interviewed said they could see the value in the activity, and one head of the English department suggested that pupils should be given more opportunity to express themselves. Some teachers suggested that the writing activity allowed pupils to be more creative. Most of the teachers we spoke to thought that the intervention was a "free writing" exercise and welcomed the opportunity, as they felt it provided a pleasant alternative to the very structured, exam-focused work that KS4 pupils usually undertake. A number of heads of the English department and English teachers commented that the opportunity for young people to write freely and be able to express their personal views was very important. One teacher commented that the children were so conditioned to focus on exams and meeting exam criteria that to do something different was refreshing and interesting. In another school, following the first writing task, one teacher said that being involved in the "Writing about Values" (WaV) project had made the faculty consider whether to teach more free and creative writing and to embed this within the KS4 curriculum. She felt that there could be opportunities to include WaV-style tasks within schemes of work, benefitting staff and students by making it a regular and expected part of English lessons. Tying the intervention in to the English lesson worked well.

### *Pupils' views*

To capture the views of pupils, a short questionnaire was sent out via emails through the school to pupils, after their GCSEs. As pupils had already left school after their GCSE, these emails were sent to their parents' email addresses.

Only six pupils responded, and these presented quite mixed perspectives. One pupil commented that it was “helpful to be encouraged to see things in a different way but at the same time a lot of people felt as though the time spent on the exercise could have been better used by working towards our GCSEs”. Another student felt that the writing task made them realise that there were “lots of things that I find valuable”, while another said that “doing something free and away from the prescribed GCSE was a relief”. Two students mentioned that there was considerable overlap between the values exercise and issues raised during their Religious Studies GCSE course. Given the small number of responses, we could not read much into these comments, but they did give us some food for thought.

### *What are the challenges teachers faced in delivering the intervention?*

One of the biggest challenges was in scheduling and delivering the written exercises around KS4 mock exams and actual GCSE exams. One school did not complete the third writing exercise, as the English teachers felt that the time was needed for revision before the GCSE. Despite several attempts to encourage them to complete the writing exercise, the school was simply unable to do it. Another school did not complete the second writing exercise. Apparently the exercise was lost in the school’s internal post, and it turned up 8 weeks later, by which time the third exercise was due, so it was not possible to fit in the second exercise before the final exercise. A small number of classes within some schools also did not complete at least one of the tasks. Ensuring that tasks were completed if students were absent from the original English lesson was another challenge mentioned by some teachers.

For the majority of the teachers, the task was not seen as too much of an imposition, as it took only 10 to 15 min and was delivered only three times in the year. It was simple, quick, and easy to deliver. Generally, the intervention fitted really well within the curriculum, and its covert nature made it easier to be seen as part of the regular English lessons. On the whole, the intervention appeared to have been delivered as intended. Teachers followed closely the guidance and instructions provided in the way that they handled pupils’ queries. Therefore, for effective implementation of the intervention, it is important that teachers are thoroughly briefed. In this trial, the developers gave very clear verbal and written instructions to ensure that teachers adhered to the protocol. Additional telephone and email briefings were offered to teachers.

## Conclusions

This trial shows that disadvantaged pupils who received the intervention made very slightly more progress between KS2 and KS4 than pupils who did not receive the intervention. In line with the theory set out in the literature review, the intervention shows no benefit for the general pupil population (i.e., including non-disadvantaged pupils). This is consistent with previous research, suggesting that the intervention can help to mitigate against the negative effect of being stereotyped for being a member of a group that is often performing poorly academically (Cohen et al., 2006; Hadden et al., 2020; Miyake et al., 2010). The effects, while positive, are small.

Previous evidence also suggests that the effects of the intervention could last for several years. We tested this with the Year 10 pupils a year after the intervention ended. The results show that the small positive effects of the intervention were sustained over 1 year.

There is no standard interpretation of effect sizes, and any effects must be considered in relation to costs, opportunity costs, and unintended outcomes. Given that the intervention takes under 20 min, it may be useful in its own right, as it is delivered three times a year and costs almost nothing, that is, there is hardly any opportunity cost for schools. Although the impact was small, the positive correlation between the number of exercises completed and the outcomes, plus the fact that the impact was sustained, all suggest that the intervention is worth considering, as there were currently no contra-indications or side effects. However, caution needs to be taken in deciding on the groups to which the intervention is administered. First, it does not benefit all children. Second, the evidence so far is that the intervention is beneficial only for groups in which the negative stereotype effect on their academic performance is psychological or social (Binning & Browman, 2020; Easterbrook & Hadden, 2021).

There is a problem that would need to be addressed if this intervention were to be rolled out more widely. As students and staff became more familiar with it, and more aware of the benign intentions, it may become less effective. Previous research suggests that knowledge of the intervention or purpose of the intervention can interfere with its efficacy (Yeager & Walton, 2011). Scaling it up effectively becomes a new project in itself. Replication of the study will no longer be tenable as awareness of the purpose of the activity could reduce its efficacy. This is the challenge of such an intervention.

What this study demonstrates is that post hoc conceptual replications are feasible with an intervention such as self-affirmation, where the intention and nature of the intervention have to be concealed from the participants, and its delivery is highly prescribed in order to maintain its integrity. The study closely adhered to the conditions of implementation in terms of its stealth, timing, and setting. In line with the theory of self-affirmation, the writing exercises were delivered once at the beginning of the academic year (before the experience of negative stereotype was established), and prior to a stressful event before the final mock GCSEs (for the Y10 pupils) and the actual GCSEs exams (for the Y11 pupils) in normal classroom conditions. The findings suggest (weakly) that the intervention works with disadvantaged pupils in

England just as it did with ethnic minority pupils in the US. In other words, the benefits of such value-affirming activities can be effectively generalised to other contexts outside the US with other groups facing stereotype threats. Consistent with the initial studies of Cohen et al. (2006, 2009) and Sherman et al. (2013), the intervention has no benefit for the general population.

In summary, we can say that the conceptual replication has worked in that it affirms the positive impact of the intervention on a disadvantaged group of pupils albeit in a different context.

## *Limitations*

As with any research, there are limitations and compromises. The characteristics of the pupils in the trial schools were broadly representative of secondary schools in England, although they had, on average, a higher proportion of disadvantaged pupils, including EverFSM6 and SEND (special educational needs or disability) pupils. This is not surprising, as the schools targeted were those with a higher than national average proportion of pupils eligible for free school meals. The trial schools also tended to have lower attainment, on average, for the same reason. They had a lower proportion of pupils achieving five A*–C at GCSE, or equivalent, compared to the national average. They were also more likely to have a higher proportion of White British pupils and a lower proportion of English as an additional language (EAL) pupils. Therefore, the results may not be as applicable to all other schools, such as those in the UK's capital city – London – or the Midlands conurbation in the centre of England, where the demographics may be different.

Another limitation is the use of EverFSM6 as a proxy for disadvantage. As shown by many studies (e.g., Gorard, 2012; Hobbs & Vignoles, 2010; Taylor, 2018), a snapshot FSM is not a reliable measure of disadvantage, for a number of reasons. There is also a big disparity between those who are long-term eligible and those who are temporarily eligible (Gorard, 2018). Short-term eligible pupils, while labelled disadvantaged, have higher average attainment than pupils with longer term eligibility (Gorard et al., 2021). The long-term FSM-eligible pupils are more clearly disadvantaged. Therefore, using EverFSM6 as a measure of SES may not accurately reflect the full impact of such an intervention, which is to address negative experiences associated with enduring membership of a disadvantaged group. Perhaps a more accurate measure would be parental income or occupational status, or permanent FSM status (pupils who have been eligible for FSM for most of their school life), but these figures were not available here.

Third, the agreed prior attainment used for the analysis was the combined results for reading and maths. Our analysis shows that the results vary slightly whether using reading (ES = +0.04) or maths (ES = +0.02), with the combined maths and reading showing a slightly bigger effect (ES = +0.05). We cannot be sure that the same effects would be achieved if different measures (e.g., English rather than reading) were used for both the prior attainment and the post-test.

There are also a number of differences between the current study and previous ones. One is that studies conducted in the US invariably use GPA as the outcome measure. Because GPA is a continuous assessment throughout the year, it provides students with almost continuous feedback on their performance, which could reinforce the cycle of adaptive potential through which recursive effects are thought to operate (Cohen & Sherman, 2014). There is no equivalent measure to GPA in the UK. The process through which the intervention operates could be different for GCSEs, which are high-stakes, stressful, and one-shot exams. This may have dampened the effect. Indeed, other studies that use test scores as the outcome, for example, Borman et al. (2016), have also shown effects that were small.

Unlike previous US studies where the final exercise was personalised, we were unable to personalise the values for the final exercise simply because we did not have the capacity with the very large sample. This could have diluted effects – a realistic limitation of scaling up.

## Discussion

Despite the increase in experimental studies in education in the past 2 decades, few studies have been replicated so far. Makel and Plucker (2014) noted that only 13% of around 16,000 studies in the top-100 education journals were replications. Of these, 28.5% were direct replications and the rest were conceptual replications. One reason for this could be that journals, or indeed reviewers, look for articles that are deemed "original" in terms of concepts and analysis. Replication studies tended to be viewed as lacking in originality and so not contributing to new ideas. We think that this widely held view is flawed and, as Makel and Plucker argued, this is a serious misunderstanding of science and creativity, privileging novelty over trustworthiness. Being able to verify the results of previous studies is a cornerstone of scientific rigour. However, this does not mean that every study needs to be replicated. But it is important that studies reporting positive results, or indeed negative or null outcomes, are replicated as they have the potential to influence student outcomes.

Initial studies on self-affirmation effects in education were conducted by researchers who are self-affirmation theorists. It is important that other researchers also conduct such tests to ensure that the findings are not open to the accusation that they come from those who have vested interests in the intervention. They were also conducted in the US with ethnic minority students. And where experiments were replicated showing different results (e.g., Bratter et al., 2016; de Jong et al., 2016; Hanselman et al., 2017; Protzko & Aronson, 2016; Simmons, 2011), this was because there had been changes in the way the intervention was implemented in terms of timing, setting, and cultural context. In Protzko and Aronson's (2016) and Simmons's (2011) studies, the research nature of the study was not concealed from the students. This could have compromised the integrity of the intervention. It is, therefore, essential to re-affirm the earlier positive findings replicating the conditions in terms of timing and settings.

Our study replicated these conditions changing only the study participants (FSM-eligible students in England instead of ethnic minority students in the US) to see if similar results can be generalised to other populations. Stereotype threats based on ethnicity or race could be more apparent because they are based on physical attributes, SES-based stereotypes could be less obvious to all observers. However, children from low-income families also experience stereotype threats through early failure in school, which can in turn influence their aspiration, attitudes, and behaviour. Empirical research in psychology shows that children living in poverty also self-stereotype, and can see themselves as "failures" (Fell & Hewstone, 2015), which can affect their performance at school. The theory is that the value-affirmation activities give pupils a sense of value, alleviating negative feelings associated with their perceptions of themselves. Initial effects might be that they feel less threatened and more confident, and this can affect peers' and teachers' expectations. The results of this study confirm that value-affirming activities (as replicated in this present study) might help to overcome stereotype threats of low-performing disadvantaged pupils in England. This corroborates the findings of earlier studies, which, for those who wish to apply this intervention, will give confidence in the efficacy of this approach in overcoming the detrimental effects of negative stereotype for disadvantaged pupils.

We can conclude that our study has successfully conceptually replicated earlier studies in the US, partly because of the good study design (e.g., randomised controlled trial) and large sample sizes in the earlier studies (Patil et al., 2016; Shadish et al., 2008; Steiner et al., 2019). Promising approaches from the EEF trials conducted so far generally have strong designs and involve large samples, and should be replicated before they are adopted more widely. To encourage such replication work, funders and government should require that research to inform policy and practice be directly replicated, preferably by an independent research team, different to the one which conducted the initial research. The moment has come in education research to demand such replications at a much larger scale.

While the effects are smaller than earlier studies, possibly due to muted effects because of the use of test scores rather than continuous teacher assessments and the fact that the final writing exercise was not personalised, this intervention is still worth considering given that it costs virtually nothing, does no harm, and could help reduce the poverty attainment gap.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Notes on contributors

*Beng Huat See* is a professor in the School of Education, Durham University, UK. Before that she was a research associate at the University of York and the University of Birmingham.

*Rebecca Morris* is an assistant professor at Centre for Education Studies, University of Warwick, UK. She previously worked at Durham University and the University of Birmingham.

*Stephen Gorard* is a professor in the School of Education, Durham University. He is also the Director of the Durham University Evidence Centre for Education (DECE). He has previously held chairs in social science and education at Cardiff, York, and Birmingham.

*Nadia Siddiqui* is an associate professor in the School of Education, Durham University.

*Matthew J. Easterbrook* is a lecturer in the School of Psychology, University of Sussex, UK. Before that he worked at both University of Sussex and Cardiff University.

*Marlon Nieuwenhuis* is currently employed as a postdoc at the Faculty of Behavioural, Management and Social Science, University of Twente, the Netherlands. From 2015 to 2018, she worked as a Research Fellow at Sussex University, UK.

*Kerry Fox* is a senior lecturer at the School of Humanities and Social Science, University of Brighton. Before that she was a research fellow at School of Psychology, University of Sussex.

*Peter Harris* is an Emeritus Professor in the School of Psychology, University of Sussex, UK. Before that he worked at the University of Sheffield, University of Hertfordshire, and University of Nottingham.

*Robin Banerjee* is Head of the School of Psychology, University of Sussex, UK. Before that he served as a British Academy Post-Doctoral Fellow and took up a position as Lecturer in Psychology in 2002, progressing to his current position as Professor of Developmental Psychology.

## ORCID

*Beng Huat See* https://orcid.org/0000-0001-7500-379X
*Rebecca Morris* https://orcid.org/0000-0002-1699-4172
*Stephen Gorard* https://orcid.org/0000-0002-9381-5991
*Nadia Siddiqui* https://orcid.org/0000-0003-4381-033X
*Matthew J. Easterbrook* https://orcid.org/0000-0002-9353-5957
*Marlon Nieuwenhuis* https://orcid.org/0000-0003-1930-6504
*Kerry Fox* https://orcid.org/0000-0002-5873-503X
*Peter R. Harris* https://orcid.org/0000-0003-4599-4929
*Robin Banerjee* https://orcid.org/0000-0002-4994-3611

## References

Binning, K. R., & Browman, A. S. (2020). Theoretical, ethical, and policy considerations for conducting social–psychological interventions to close educational achievement gaps. *Social Issues and Policy Review*, *14*(1), 182–216. https://doi.org/10.1111/sipr.12066

Borman, G. D., Choi, Y., & Hall, G. J. (2021). The impacts of a brief middle-school self-affirmation intervention help propel African American and Latino students through high school. *Journal of Educational Psychology*, *113*(3), 605–620. https://doi.org/10.1037/edu0000570

Borman, G. D., Grigg, J., & Hanselman, P. (2016). An effort to close achievement gaps at scale through self-affirmation. *Educational Evaluation and Policy Analysis*, *38*(1), 21–42. https://doi.org/10.3102/0162373715581709

Borman, G. D., Grigg, J., Rozek, C. S., Hanselman, P., & Dewey, N. A. (2018). Self-affirmation effects are produced by school context, student engagement with the intervention, and time: Lessons from a district-wide implementation. *Psychological Science*, *29*(11), 1773–1784. https://doi.org/10.1177/0956797618784016

Brannon, T. N., Higginbotham, G. D., & Henderson, K. (2017). Class advantages and disadvantages are not so Black and White: Intersectionality impacts rank and selves. *Current Opinion in Psychology*, *18*, 117–122. https://doi.org/10.1016/j.copsyc.2017.08.029

Bratter, J. L., Rowley, K. J., & Chukhray, I. (2016). Does a self-affirmation intervention reduce stereotype threat in black and Hispanic high schools? *Race and Social Problems*, *8*(4), 340–356. https://doi.org/10.1007/s12552-016-9187-4

Cohen, G. L., Garcia, J., Apfel, N., & Master, A. (2006, September 1). Reducing the racial achievement gap: A social-psychological intervention. *Science*, *313*(5791), 1307–1310. https://www.science.org/doi/10.1126/science.1128317

Cohen, G. L., Garcia, J., Purdie-Vaughns, V., Apfel, N., & Brzustoski, P. (2009, April 17). Recursive processes in self-affirmation: Intervening to close the minority achievement gap. *Science*, *324*(5925) 400–403. https://www.science.org/doi/10.1126/science.1170769

Cohen, G. L., & Sherman, D. K. (2014). The psychology of change: Self-affirmation and social psychological intervention. *Annual Review of Psychology*, *65*, 333–371. https://doi.org/10.1146/annurev-psych-010213-115137

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of *p*-values. *Royal Society Open Science*, *1*(3), Article 140216. https://doi.org/10.1098/rsos.140216

Colquhoun, D. (2016, October 11). The problem with p-values. *Aeon*. https://aeon.co/essays/it-s-time-for-science-to-abandon-the-term-statistically-significant

Croizet, J. C., & Claire, T. (1998). Extending the concept of stereotype threat to social class: The intellectual underperformance of students from low socioeconomic backgrounds. *Personality and Social Psychology Bulletin*, *24*(6), 588–594. https://doi.org/10.1177/0146167298246003

Dee, T. S. (2015). Social identity and achievement gaps: Evidence from an affirmation intervention. *Journal of Research on Educational Effectiveness*, *8*(2), 149–168. https://doi.org/10.1080/19345747.2014.906009

de Jong, E. M., Jellesma, F. C., Koomen, H. M. Y., & de Jong, P. F. (2016). A values-affirmation intervention does not benefit negatively stereotyped immigrant students in the Netherlands. *Frontiers in Psychology*, *7*, Article 691. https://doi.org/10.3389/fpsyg.2016.00691

Demie, F., & Lewis, K. (2011). White working class achievement: An ethnographic study of barriers to learning in schools. *Educational Studies*, *37*(3), 245–264. https://doi.org/10.1080/03055698.2010.506341

Department for Education. (2015). *2010 to 2015 government policy: Education of disadvantaged children* [Policy paper].

Department for Education. (2019). *Key stage 4 performance, 2019 (revised)*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/863815/2019_KS4_revised_text.pdf

Dong, N., & Lipsey, M. W. (2011). *Biases in estimating treatment effects due to attrition in randomized controlled trials: A simulation study*. https://files.eric.ed.gov/fulltext/ED517992.pdf

Dunn, G., Maracy, M., & Tomenson, B. (2005). Estimating treatment effects from randomized clinical trials with noncompliance and loss to follow-up: The role of instrumental variable methods. *Statistical Methods in Medical Research*, *14*(4), 369–395. https://doi.org/10.1191/0962280205sm403oa

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*, Article 621. https://doi.org/10.3389/fpsyg.2015.00621

Easterbrook, M. J., & Hadden, I. R. (2021). Tackling educational inequalities with social psychology: Identities, contexts, and interventions. *Social Issues and Policy Review*, *15*(1), 180–236. https://doi.org/10.1111/sipr.12070

Education Endowment Foundation. (2018). *Annual report*. https://educationendowmentfoundation.org.uk/public/files/Annual_Reports/EEF_-_2018_Annual_Report.pdf

Fell, B., & Hewstone, M. (2015). *Psychological perspectives on poverty*. Joseph Rowntree Foundation.

Foster, E. M., & Fang, G. Y. (2004). Alternatives to handling attrition: An illustration using data from the fast track evaluation. *Evaluation Review*, *28*(5), 434–464. https://doi.org/10.1177/0193841X04264662

Garcia, J., & Cohen, G. L. (2012). A psychological approach to educational intervention. In E. Shafir (Ed.), *The behavioral foundations of public policy* (pp. 329–347). Princeton University Press. https://ed.stanford.edu/sites/default/files/a_social_psychological_approach_to_educational_intervention_0.pdf

Good, C., Aronson, J., & Inzlicht, M. (2003). Improving adolescents' standardized test performance: An intervention to reduce the effects of stereotype threat. *Journal of Applied Developmental Psychology*, *24*(6), 645–662. https://doi.org/10.1016/j.appdev.2003.09.002

Gorard, S. (2012). Who is eligible for free school meals? Characterising free school meals as a measure of disadvantage in England. *British Educational Research Journal*, *38*(6), 1003–1017. https://doi.org/10.1080/01411926.2011.608118

Gorard, S. (2018). *Education policy: Evidence of equity and effectiveness*. Policy Press.

Gorard, S. (2021). *How to make sense of statistics*. SAGE Publications.

Gorard, S., See, B. H., & Siddiqui, N. (2017). *The trials of evidence-based education: The promises, opportunities and problems of trials in education*. Routledge.

Gorard, S., Siddiqui, N., & See, B. H. (2021). Assessing the impact of Pupil Premium funding on primary school segregation and attainment. *Research Papers in Education*. Advance online publication. https://doi.org/10.1080/02671522.2021.1907775

Goyer, J. P., Garcia, J., Purdie-Vaughns, V., Binning, K. R., Cook, J. E., Reeves, S. L., Apfel, N., Taborsky-Barba, S., Sherman, D. K., & Cohen, G. L. (2017). Self-affirmation facilitates minority middle schoolers' progress along college trajectories. *Proceedings of the National Academy of Sciences*, *114*(29), 7594–7599. https://doi.org/10.1073/pnas.1617923114

Hadden, I. R., Easterbrook, M. J., Nieuwenhuis, M., Fox, K. J., & Dolan, P. (2020). Self-affirmation reduces the socioeconomic attainment gap in schools in England. *British Journal of Educational Psychology*, *90*(2), 517–536. https://doi.org/10.1111/bjep.12291

Hanselman, P., Rozek, C. S., Grigg, J., & Borman, G. D. (2017). New evidence on self-affirmation effects and theorized sources of heterogeneity from large-scale replications. *Journal of Educational Psychology*, *109*(3), 405–424. https://doi.org/10.1037/edu0000141

Harackiewicz, J. M., Canning, E. A., Tibbetts, Y., Priniski, S. J., & Hyde, J. S. (2016). Closing achievement gaps with a utility-value intervention: Disentangling race and social class. *Journal of Personality and Social Psychology*, *111*(5), 745–765. https://doi.org/10.1037/pspp0000075

Hobbs, G., & Vignoles, A. (2010). Is children's free school meal "eligibility" a good proxy for family income? *British Educational Research Journal*, *36*(4), 673–690. https://doi.org/10.1080/01411920903083111

Hofstede, G. J., Pedersen, P. B., & Hofstede, G. (2002). *Exploring culture: Exercises, stories and synthetic cultures*. Intercultural Press.

Hunter, J. E. (2001). The desperate need for replications. *Journal of Consumer Research*, *28*(1), 149–158. https://doi.org/10.1086/321953

Johnston, J. M., & Pennypacker, H. S. (2009). *Strategies and tactics of behavioral research* (3rd ed.). Routledge.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. John Wiley & Sons.

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, *43*(6), 304–316. https://doi.org/10.3102/0013189X14545513

McCoy, C. E. (2017). Understanding the intention-to-treat principle in randomized controlled trials. *Western Journal of Emergency Medicine*, *18*(6), 1075–1078. https://doi.org/10.5811/westjem.2017.8.35985

Miyake, A., Kost-Smith, L. E., Finkelstein, N. D., Pollock, S. J., Cohen, G. L., & Ito, T. A. (2010, November 26). Reducing the gender achievement gap in college science: A classroom study of values affirmation. *Science*, *330*(6008), 1234–1237. https://www.science.org/doi/10.1126/science.1195996

Montori, V. M., & Guyatt, G. H. (2001). Intention-to-treat principle. *Canadian Medical Association Journal*, *165*(10),1339–1341. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC81628/

Morrison, K. (2019). Realizing the promises of replication studies in education. *Educational Research and Evaluation*, *25*(7–8), 412–441. https://doi.org/10.1080/13803611.2020.1838300

Nicholl, J. (n.d.). *Complier average causal effect analysis*. https://www.sheffield.ac.uk/polopoly_fs/1.418711!/file/JNicholls.pdf

Organisation for Economic Co-operation and Development. (2019). *PISA 2018 results (Volume II): Where all students can succeed*. https://doi.org/10.1787/b5fd1b8f-en

Oyserman, D., Bybee, D., & Terry, K. (2006). Possible selves and academic outcomes: How and when possible selves impel action. *Journal of Personality and Social Psychology*, *91*(1), 188–204. https://doi.org/10.1037/0022-3514.91.1.188

Patil, P., Peng, R. D., & Leek, J. T. (2016). What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspectives on Psychological Science*, *11*(4), 539–544. https://doi.org/10.1177/1745691616646366

Perezgonzalez, J. D. (2015). The meaning of significance in data testing. *Frontiers in Psychology*, *6*, Article 1293. https://doi.org/10.3389/fpsyg.2015.01293

Pharoah, P. D. P., Jones, M. R., & Kar, S. (2017, August 24). *P-values and confidence intervals: Not fit for purpose?* bioRxiv. https://doi.org/10.1101/180117

Protzko, J., & Aronson, J. (2016). Context moderates affirmation effects on the ethnic achievement gap. *Social Psychological and Personality Science*, *7*(6), 500–507. https://doi.org/10.1177/1948550616646426

Puma, M. J., Olsen, R. B., Bell, S. H., & Price, C. (2009). *What to do when data are missing in group randomized controlled trials* (NCEE 2009-0049). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Rosenthal, R., & Jacobson, L. F. (1968). Teacher expectations for the disadvantaged. *Scientific American*, *218*(4), 19–23. https://www.jstor.org/stable/24926197

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, *103*(484), 1334–1344. https://doi.org/10.1198/016214508000000733

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Sherman, D. K., & Cohen, G. L. (2006). The psychology of self-defense: Self-affirmation theory. *Advances in Experimental Social Psychology*, *38*, 183–242. https://doi.org/10.1016/S0065-2601(06)38004-5

Sherman, D. K., Hartson, K. A., Binning, K. R., Purdie-Vaughns, V., Garcia, J., Taborsky-Barba, S., Tomassetti, S., Nussbaum, A. D., & Cohen, G. L. (2013). Deflecting the trajectory and changing the narrative: How self-affirmation affects academic performance and motivation under identity threat. *Journal of Personality and Social Psychology*, *104*(4), 591–618. https://doi.org/10.1037/a0031495

Simmons, C. M. (2011). *Reducing stereotype threat in academically at-risk African-Americans students: A self-affirmation intervention* [Doctoral dissertation, UC Berkeley]. https://digitalassets.lib.berkeley.edu/etd/ucb/text/Simmons_berkeley_0028E_11998.pdf

Slavin, R. (2018, February 28). What kinds of studies are likely to replicate? *Robert Slavin's Blog*. https://robertslavinsblog.wordpress.com/2018/02/28/what-kinds-of-studies-are-likely-to-replicate/

Spencer, B., & Castano, E. (2007). Social class is dead. Long live social class! Stereotype threat among low socioeconomic status individuals. *Social Justice Research*, *20*(4), 418–432. https://doi.org/10.1007/s11211-007-0047-7

Spencer, S. J., Logel, C., & Davies, P. G. (2016). Stereotype threat. *Annual Review of Psychology*, *67*, 415–437. https://doi.org/10.1146/annurev-psych-073115-103235

Steele, C. M. (1988). The psychology of self-affirmation: Sustaining the integrity of the self. *Advances in Experimental Social Psychology*, *21*, 261–302. https://doi.org/10.1016/S0065-2601(08)60229-4

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, *52*(6), 613–629. https://doi.org/10.1037/0003-066X.52.6.613

Steiner, P. M., Wong, V. C., & Anglin, K. (2019). A causal replication framework for designing and assessing replication efforts. *Zeitschrift für Psychologie*, *227*(4), 280–292. https://doi.org/10.1027/2151-2604/a000385

Strand, S. (2014). Ethnicity, gender, social class and achievement gaps at age 16: Intersectionality and "getting it" for the White working class. *Research Papers in Education*, *29*(2), 131–171. https://doi.org/10.1080/02671522.2013.767370

Taylor, C. (2018). The reliability of free school meal eligibility as a measure of socio-economic disadvantage: Evidence from the Millennium Cohort Study in Wales. *British Journal of Educational Studies*, *66*(1), 29–51. https://doi.org/10.1080/00071005.2017.1330464

Travers, J. C., Cook, B. G., Therrien, W. J., & Coyne, M. D. (2016). Replication research and special education. *Remedial and Special Education*, *37*(4), 195–204. https://doi.org/10.1177/0741932516648462

Walton, G. M., & Yeager, D. S. (2020). Seed and soil: Psychological affordances in contexts help to explain where wise interventions succeed or fail. *Current Directions in Psychological Science*, *29*(3), 219–226. https://doi.org/10.1177/0963721420904453

Woodcock, A., Hernandez, P. R., Estrada, M., & Schultz, P. W. (2012). The consequences of chronic stereotype threat: Domain disidentification and abandonment. *Journal of Personality and Social Psychology*, *103*(4), 635–646. https://doi.org/10.1037/a0029120

Wu, Z., Spreckelsen, T. F., & Cohen, G. L. (2021). A meta-analysis of the effect of values affirmation on academic achievement. *Journal of Social Issues*, *77*(3), 702–750. https://doi.org/10.1111/josi.12415

Yeager, D. S., & Walton, G. M. (2011). Social-psychological interventions in education: They're not magic. *Review of Educational Research*, *81*(2), 267–301. https://doi.org/10.3102/0034654311405999

## Appendix 1. Second writing exercise for control group

Name:
Date:
English teacher:
Writing about your life

People begin their days in many different ways. Sometimes it can be interesting to think about the way we begin our own day.

In the space below, please write about what you did this morning before you started school. What time did you get up? How long did it take to get ready? Did you eat or drink anything? How did you get to school? What did you pass on the way to school?

Try to start with the very first thing you did this morning, then describe what happened afterwards.

Focus on writing down what happened, and don't worry about spelling, grammar, or how well written it is, or how much you can write.

*Please turn over*

## Appendix 2. Second writing exercise for treatment group

Name:
Date:
English teacher:
Writing about your life

There are a lot of things that are important to people—things that make their lives better, more important, or special.

For example, some people find being honest important because other people can trust them. Some other people find their family important because they love and value them. Other people find being good at sport important because it makes them feel good to play well.

In the space below, please write about what **you** find important in **your** life. How important is it to you? Why is it important to you? What does it mean to you to have it in your life?

Focus on your thoughts and feelings, and don't worry about spelling, grammar, or how well written it is, or how much you can write.

*Please turn over*

## Appendix 3. Teacher information sheet

**Writing about Values exercise – Instructions for teachers**

You will receive a box of envelopes with the writing exercises, sorted by class, with your pupils' names on the front.

**What to do:**

☑ Ensure the class is settled. Introduce the exercise as you would any other in-class exercise using your own words, but please ensure you cover the 10 numbered points below:

1. For the first part of today's lesson, we're going to be doing something a bit different - a free-expression exercise.
2. I'm going to hand you out an envelope with your name on.
3. DO NOT open them until I tell you.

☑ Then, give each envelope to the corresponding pupil, but do not let them open them yet. If a pupil's envelope is missing, please write their name on one of the blank envelopes and use that. Now please cover the following points:

4. Read the instructions carefully so you know what to do
5. There are no right or wrong answers
6. The exercise is a chance for you to spend some time writing about your own thoughts and ideas; it's about the process of doing the activity rather than me providing feedback so it's not going to be marked
7. You don't need to focus on spelling or grammar
8. It takes about 10-15 minutes
9. Work individually and silently
10. If you have a question, raise your hand and I will come over to your desk

☑ If you would normally do so, you can now check for questions. Ensure pupils are silent and then ask them to begin. Please make sure the pupils complete the exercise individually. If a pupil has a question, approach them at their desk and talk to them quietly, using the FAQs below where possible.

☑ Give pupils 10-15 minutes of writing time to complete the exercise. If a pupil finishes earlier, please encourage them to go back over their work. After about 10 minutes, please say something like "You have a couple of minutes left to finish up, don't worry if you can't quite finish it". It doesn't matter if some take longer than others.

☑ Have the **pupils put their completed exercise back into the envelopes and collect them.** Please fill out the cover sheet at the back and give everything to your school contact at the end of the day. Please do *not* refer back to the exercise in class once it is completed.

☑ **If any pupils are absent,** please give the exercise to them when they are next in your class (within 2 weeks of original exercise date) and write the date that they completed the exercise on the envelope.

Suggested responses to **frequently asked questions** from pupils:

- **Why are we doing this?** — Pupils in other schools have found that spending some time thinking and writing about their own thoughts really helpful and we are keen to try them out. Everyone in Y10 and Y11 is doing the exercise (If a pupil refuses, please accept this and note it on your cover sheet).
- **Will I get marked on this?/Who will read this?** — I will check to see if you've engaged with it properly, but it won't be marked. The exercises will be stored away.
- **What are you going to do with what I write?** — This is about the process of writing and giving you the chance to write your own ideas, so it won't be marked. We'll collect them up and store them away.
- **Why do we get envelopes?** — You're writing about your own personal thoughts and ideas, so it's important that they are private.

- **Why do I have different questions from him/her?** — Everyone's got their own task but there's not enough time for everyone to do them all, some people have different ones.
- **Is this for the whole school?** — All Y10 and pupils will be doing this at some point.
- **Does spelling/grammar matter?** — No, just focus on writing down your thoughts.
- **Can I write about a value that's not on the list?** — For now, just choose one on the list.
- **Is this part of the study/research?** — This is an exercise that our school is trying out this year. (If possible, address this question individually at their desk)