Statistical Analysis Plan for Writing about Values

Durham University

INTERVENTION	Writing about Values
DEVELOPER	University of Sussex
EVALUATOR	Durham University
TRIAL REGISTRATION NUMBER	**ISRCTN79754465**
TRIAL STATISTICIAN	Stephen Gorard
TRIAL CHIEF INVESTIGATOR	Beng Huat See
SAP AUTHOR	Beng Huat See
SAP VERSION	
SAP VERSION DATE	6 November 2017
EEF DATE OF APPROVAL	
DEVELOPER DATE OF APPROVAL	

Introduction

"Writing about Values" is a one-year intervention comprising two phases: an initial pilot phase and the main trial. The main trial consists of two randomly controlled trials in the same schools. One trial involves Year 10 pupils and the other involves Year 11 pupils. The inclusion of the Y10 pupils enables evaluation of the long-term impact of the intervention – a year after the end of the intervention.

The trial is a double-blind experiment where both pupils and teachers are not told what the intervention involves. The writing exercises are completed during English lessons as part of the regular English class. The treatment and control pupils are given similar exercises with a slight variation.

The intervention is based on the hypothesis that students from some stigmatised groups are aware that they are the target of a negative stereotype regarding their academic performance (Steele 1997). The "Writing About Values" strategy has been employed to alleviate the effects of stereotype threat on low performing students, especially those from ethnic minority backgrounds (Oyserman et al. 2006; Cohen et al. 2006; Miyake et al. 2010) by getting them to write positive statements about themselves (Cohen & Sherman 2006). It is believed that this can help ameliorate the detrimental effects of stereotype threat on academic performance. For this trial the focus is on EverFSM pupils rather than ethnic minorities because EEF's focus is on disadvantaged pupils.

Study design

This is an efficacy trial running for two years involving two randomised trials: one with Y10 pupils for two years, and another with Y11 pupils for one year. However, delivery of the intervention stops at the end of the first year (July 2017). Evaluation of impact for Y11 will be undertaken at the end of the first year after the GCSE results, while impact evaluation for the Y10 will be at the end of the second year to test the sustained effect.

This trial is conducted as a double-blind experiment in that both pupils and teachers are not given information about the intervention apart from the fact that it involves a writing task.

Protocol changes

Following the pilot trial, a few changes have been made to the initial protocol:

- The non-attainment survey items have been revised and after testing the convergent and predictive validity of the scales as well as the internal consistency, it has been decided that the pupils' perceived self-efficacy will be the non-attainment outcome of interest. This will be measured using the subscales from the Motivated Strategies for Learning Questionnaire (Pintrich et al. 1993).

  The sub-group analysis will now include an analysis of self-efficacy as the non-attainment outcome

- Pupils will be individually randomised, stratifying by year group and FSM status only, but not by teaching class as originally planned.

- Due to the recent changes in the GCSE exams we anticipate that there may be a delay in getting hold of the results. It is decided that the unconfirmed GCSE scores will be used to ensure that results are available on time for analysis. The expected date of completion for Report 1 will now be February 2018 (instead of January 2018) and February 2019 (instead of January 2019) for Report 2.

Randomisation

Pupils will be individually randomised within school, stratified by year group and free school meal status. Randomisation is carried out after all participating schools have submitted the pupil data. A random number generator on Excel will be used for this process which will be conducted in the presence of colleagues in the School of Education. Year groups will be randomised separately, and FSM and non-FSM separately within each year. There will be, in effect, four randomisations. Schools will not be informed of the results of the randomisation as it is a double-blind experiment. Since the treatment and control pupils are given different writing tasks, the developers will be informed of the result of the randomisation immediately in order for the named exercise booklets to be printed on time.

Calculation of sample size

The sample size calculation is based on the assumption that there would be 25 schools and two year groups (Year 10 and Y11). Assuming an average of 5 forms in each year group, there will be 125 forms (25 X 5) per year group. Working on the assumption of an average of 30 pupils per form, there will be 3,750 pupils (30 X 125) for each year group or 7,500 overall. Randomising individual pupils to treatment conditions, there will be 1,875 pupils in each arm for each year group (or 3,750 per arm overall). Assuming around 25% of EverFSM-eligible pupils overall, this would mean around 470 EverFSM pupils per arm per year group (or 940 per arm of the trial).

Traditional power calculations are based on the approach of significance testing (Gorard et al. 2017), which is misleading. They are therefore not included here.

Instead, we calculate 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 and Gorard 2016). This number needed to disturb (NNTD) is calculated as 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). This approach allows for estimating ES and sample size using the formula as shown.

$$NNTD = ES*n$$

Therefore, $n = NNTD/ES$ and

$$ES = NNTD/n$$

This is a useful measure of the scale of the findings to chance (and their variability as represented by the standard deviation used to compute the 'effect' size), taking into account the scale of the study. It can then be extended to compare this sensitivity directly to other more substantial sources of error such as the number of missing values/cases. The number of cases actually missing a value can be subtracted from the NNTD to give an estimate of how large the 'effect' size would be even in the extreme situation that all missing cases had the "counterfactual" score hypothesised in the NNTD calculation. Here the 'counterfactual' score is one standard deviation away from the mean of the group with the largest number of cases. The standard deviation would be added if the mean of the smaller group (in scale) were smaller than the mean of the larger group, and subtracted if the mean of the smaller group was the largest. (Gorard et al. 2017).

Based on Gorard et al. 2016, NNTD of 50 can be considered a strong and secure finding. Using this as a working assumption, the number of cases needed in each group (assuming

equal size) to detect an 'effect' size of 0.2 (which is typical for an education intervention) will be 250 (or 50/0.2). This is assuming no attrition.

Assuming 25% of pupils are EverFSM (n=940), we would expect to detect an 'effect' size of 0.05, or 50/940 (ES=NNTD/n), for the EverFSM pupils. In reality 30 schools and 11,978 pupils were recruited (n=5953 treatment and n=6025 control). Of these 26% (or 3,131) had EverFSM status (or around 1,565 per arm). This will enable us to confidently detect an 'effect' size of +0.03.

The NNTD calculation concerns the security of a difference, and so is relevant to internal validity only. Issues such as clustering, concerned with whether the result may also occur among cases not in the RCT, are therefore irrelevant. In addition, as pupils are individually randomized within schools and analysis would be of all pupils in the two groups and not by schools, clustering effects, if there are any, should be evenly spread between the two groups across all schools.

Follow-up

One school pulled out of the intervention after randomisation, but it was agreed with the EEF and the developers that only schools that completed the baseline survey will be included in the trial and the analysis. This school was therefore not included in the evaluation.

Outcome measures

We propose using attainment 8 KS4 scores as the main attainment outcomes. We will use KS2 maths total marks and reading marks as the pre-intervention attainment measures.

Primary outcomes

- Attainment 8 KS4 scores for FSM pupils (based on EverFSM) after 1 year of treatment (for initial Y11 pupils).
- Attainment 8 KS4 scores for FSM pupils (based on EverFSM) after 2 years (one year after the end of the intervention) for pupils who received the intervention when they were in Y10.

Secondary outcomes

- Attainment 8 KS4 scores for All (EverFSM and non-EverFSM) pupils after 1 year of treatment (for initial Y11 pupils).
- Attainment 8 KS4 scores for All (EverFSM and non-EverFSM) pupils after 2 years (for pupils who received the intervention when they were in Y10).
- Attainment 8 KS4 scores for FSM pupils (based on current FSM status) after 1 year of treatment (for initial Y11 pupils).
- Attainment 8 KS4 scores for FSM pupils (based on current FSM status) after 2 years (one year after the end of the intervention) for pupils who received the intervention when they were in Y10.

Non-attainment outcomes

Pupils' perceived self-efficacy measured using the subscales from the Motivated Strategies for Learning Questionnaire (Pintrich et al. 1993).

Analysis

Primary intention-to-treat (ITT) analysis

The analyses for the impact evaluation will be based on the difference between groups in terms of their post-test mean scores for EverFSM pupils only. The differences will be expressed as effect sizes (Hedge's) and converted to progress in months. Given the number of cases per cell and the reported effect sizes, it will be possible for readers to construct a confidence interval.

Imbalance at baseline

To establish baseline equivalence we will use the 'effect' sizes for each measurement at the outset and also present the characteristics of schools in each group. To cater for any initial imbalances between groups we also present the gain scores analysis. For the benefit of readers we present the pre-, post- and gain scores regardless of imbalance.

Missing data

Dong and Lipsey (2011) demonstrated that any missing values can create bias, even if attrition is balanced between comparator groups. And where such attrition is not random (as is most often the case) it can bias the estimate of the treatment effect, and the bias can still be large even when advanced statistical methods like multiple imputations are used (Foster & Fang 2004; Puma et al. 2009). Such bias can distort the results of statistical significant tests and threaten the validity of any conclusion reached (Shadish, Cook & Campbell 2001; Campbell & Stanley 1963; Little & Rubin 1987).

Based on this, we should not use existing data to substitute for data that is missing, since we have little or no knowledge of the missing cases, and missing data/cases are seldom random. Doing so will increase the potential for bias. We therefore present differences in pre-test scores (KS2 Maths and Reading) between cases dropping out from both groups (where these are available).

In addition, we will report any missing data and compare the level of missing data to the number of hypothetical counterfactual cases needed to disturb the finding (Gorard et al 2017). The number of counterfactual cases will help determine whether the number of missing cases is large enough to alter/explain the findings (see explanation in section on Calculation of Sample Size).

Fidelity analysis

The fidelity to the intervention will be assessed by comparing the outcomes of pupils with the number of exercises completed (dosage). We will run a regression analysis using dosage (number of exercises completed) as the predictor.

The number of exercises will be used as a count variable in the analysis. This will be zero for all cases in the control group.

In addition, we will perform Complier Average Causal Effect (CACE) analysis to estimate the effects for the subgroup of treatment students who comply with their treatment assignment. Compliance is defined as completing the first writing task.

Secondary outcome analyses

Secondary outcome analyses will be comparisons of pre-, post- and gain score 'effect' sizes for:

- All pupils (i.e, both EverFSM and non-EverFSM) after 1 year
- All pupils (i.e. both EverFSM and non-EverFSM after 2 years
- Self-efficacy

Additional analyses

We will create two multivariate regressions, the first will use post-test scores (Atttainment 8 KS4 scores) as the dependent variables, and total prior test scores (KS2 maths and Reading) and membership of treatment group as predictors. The second model will also include year group and FSM status as predictors.

Effect size calculation

'Effect' sizes will generally be calculated as Hedges' *g* based on the difference between mean post-test (and gain scores) for each variable. We will not report 'confidence intervals' but an interested reader can compute them if they wish as we will report the number of cases per group, and the effect size for each comparison.

'Effect' sizes for categorical variables (self-efficacy) will be based on post-intervention odds ratios – or changes in odds where the groups are clearly unbalanced at the outset. All will be presented with the number of counterfactual cases needed to disturb the results.

Report tables

Executive Summary

Group	Effect size	Estimated months' progress	EEF security rating	EEF cost rating
Treatment vs. control EverFSM – Attainment 8 at GCSE for after 1 year (for Y11 cohort)				
Treatment vs. control Ever FSM– Attainment 8 at GCSE after 2 years (for Y10 cohort)				
Treatment vs. control – Attainment 8 at GCSE for All pupils after 1 year (Y11 cohort)				
Treatment vs. control – Attainment 8 at GCSE for All pupils after 2				

years (Y10 cohort)				
Treatment vs. control – Pupils' perceived self-efficacy				

References

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Chicago, IL: Rand McNally.

Dong, N. and Lipsey, M. (2011) *Biases in estimating treatment effects due to attrition in randomised controlled trials: A Simulation study*. SREE Conference, 2011.

Foster, M. E. and Fang, G. Y. (2004). Alternatives to Handling Attrition: An Illustration Using Data from the Fast Track Evaluation. *Evaluation Review, 28*:434-464.

Little, R. J. A., and Rubin, D. B. (1987) *Statistical analysis with missing data.* New York: Wiley

Mikaye, A., Kost-Smoth, L. E., Finkelstein, N. D., Pollock, S. J., Cohen, G. L. & Ilto, A. (2010). Reducing the gender achievement gap in college science: A classroom study of values affirmation. *Science*, 330, 1234-1237. doi:10.1126/science.1195996

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, 188-204. doi:10.1037/0022-3514.91.1.188

Pintrich, P.R., Smith, D.A.F., Garcia, T and Mckeachie, W.J. (1993) Reliability and predictive validity of the Motivated Strategies for Learning Questionnaire (Mslq). *Educational and Psychological Measurement*, 53, 3, 801-813.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001) *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton Mifflin.

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