Learning by Questions

Institute for Effective Education
Empowering educators with evidence

Retrieval practice in science

Protocol version two

Contents

Introduction

This amended protocol describes changes made to the original protocol to reflect two separate unforeseen circumstances (IEE, 2019). The first changes reflect logistical challenges initiating the evaluation across two schools. The second set of changes are due to the COVID-19 pandemic which led to school closures that prevented the planned post-testing (DfE, 2020).

Taken together, the changes reduce the security of the findings from the trial. However, in updating the protocol we have made every effort to minimise this by strengthening aspects of the evaluation that we can and minimising a widespread problem known as researcher degrees of freedom by pre-specifying our analysis in greater detail (Ioannidis, 2014).

During an evaluation, specifically when analysing data, researchers make many choices. There is seldom a single 'best' analysis, rather a family of analyses with some more appropriate than others. An informative example of this comes from a project where 29 teams analysed data about the awarding of red cards in football and players' skin colour (Silberzahn et al., 2018). The analysis teams made widely different choices, yet the variation could not be explained by their expertise, prior beliefs or the quality of their work assessed by peer review.

Wide-ranging evidence indicates that this array of choices – or analytic flexibility – can be exploited to bias research (Open Science Collaboration, 2015). A useful analogy is the 'garden of forking paths' whereby every choice creates a new fork with paths that may lead to different destinations (Rubin, 2017). Researchers can (un)consciously exploit the array of pathways open to them. They may explore many pathways, yet only report the minority leading to their favoured destination. When this is done deliberately, it is known as *p*-hacking (Gehlbach & Robinson, 2018). An interactive tool illustrating the problem is available from the data journalism website FiveThirtyEight (Aschwanden, 2015). The unreported flexibility plays a major role in the 'reproducibility crisis' affecting science (Munafò et al., 2017), which led the famous claim that 'most claimed research findings are false' (Ioannidis, 2005, p. 0696).

Developing a detailed protocol that pre-specifies the analysis is akin to planning the journey through the garden before leaving the house. After conducting the pre-specified analysis, researchers are free to 'explore the garden' more freely, but it becomes clearer that this analysis was exploratory – and therefore more tentative – than the pre-specified analysis.

While protocols have grown in use, adherence to protocols is often poor. A review of trials published in leading medical journals – where protocols are a requirement for publication – found that adherence to protocols was poor (Goldacre et al., 2019). Therefore, it is critical to adhere to the protocol and carefully document ®any changes to the protocol, which is the aim of this amended protocol.

This amended protocol was developed prior to accessing and analysing the data.

Phase one changes

We were forced to alter the design of the evaluation after developing the original protocol, but prior to pre-testing and randomisation. Curriculum changes in one of the intended schools meant that the intervention was no longer appropriate. Consequently, we changed the design of the study from a multi-site trial to a single site trial and increased the number of students participating in the single site. We adopted the same procedure for randomisation but conducted case matching of the four classes in a single school. We then proceeded to randomly allocate each of the case matched pairs using a coin toss as described in the original protocol.

The change in the design of the study delayed the timeline. This was compounded by other issues including a delay in receiving the materials for assessing resilience. As a result of the delay, the pupils in the study were in Year 11, rather than Year 10 as originally planned. We planned to run the intervention from April2019 to November 2020, but as described in the next section there were further changes to the evaluation.

We originally planned to measure pupils' resilience. We were unable to obtain the original scale to measure resilience so – with the support from colleagues at the Institute for Effective Education – opted for the Pearson Clinical Scale (Prince-Embury, 2008). Delays in obtaining this scale for the baseline measurement further delayed the evaluation so that we did not begin the evaluation until September 2019.

TABLE 1: SUMMARY OF PHASE 1 CHANGES

Dimension	Original protocol	Updated protocol
Number of sites (schools)	2	1
Timing	April 2019 – November 2019	September 2019 – April 2020
Measure of resilience	The Resilience Scale	Pearson clinical scale

Phase two changes – impact evaluation

The Covid-19 pandemic prevented the administration of the planned post-tests due to nationwide school closures (DfE, 2020). As the students involved were in Year 11, it was not possible to extend the evaluation into the following year. GCSE results could not be used either as the assessments were cancelled and replaced with teacher assessment, which would be conflated with treatment allocation (Ofqual, 2020). Therefore, we decided to use two end-of-topic assessments that students had already completed as the best option for securing an impact evaluation.

Further changes stemming from this decision, and the technical details are described below. Notably, we sought to mitigate the limitations of these changes and strengthen other aspects wherever possible. However, the opportunities for further data collection were very limited, especially from the participating students.

Impact evaluation

Outcome measures

We changed the primary outcome measure of pupil attainment to an end-of-topic test completed around six weeks after the intervention began. The new and original assessments attempt to measure the students' scientific knowledge, although the new measure focused on a narrower range of content. The assessment forms part of the usual routine in our school, comprised 29 marks and was marked and administered by class teachers. We included an additional secondary measure of students' scientific knowledge taken from a similar assessment administered three weeks into the intervention. Both of the assessments are available on the Open Science Framework (Boyce & Martell, 2020). Figure 1 shows the amended research design using design notation (Gorard, 2013).

$$O_1\ R\ X\ O_2\ X\ O_3$$

$$O_1\ R\quad\ O_2\quad\ O_3$$

Figure 1: design notation for the updated protocol where R indicates randomisation, X indicates the intervention, O_1 indicates the baseline assessment, O_3 is the primary outcome measure taken after six weeks of the intervention and O_2 is the secondary outcome taken after three weeks of the intervention

A clear limitation of the amended design is the reduction in the length of the intervention. This introduces three challenges. First, students will receive a lower dosage of the intervention than originally expected – this will make it harder to discern any difference between conditions. Second, shorter interventions are more susceptible to Hawthorne effects. Third, an 'implementation dip' is often associated with any new intervention – it is therefore plausible that this would explain any negative effect. Each of these issues will be considered in the analysis.

We originally intended to mark the assessments blind to treatment allocation to minimise bias (IES, 2017). This was not possible as the assessments had already been marked by the class teachers and returned to students to support their revision before we realised that we would need to use them as the outcome measure. Therefore, we were unable to quality assure the marking. Overall, we think the risk posed by this issue is relatively low. The nature of the assessments means that they are either right or wrong – there is very limited scope for professional judgement or systematic errors. Further, the formative purpose, and low-stakes nature, of the assessments means that there was limited incentive for teachers to subvert the marking.

We originally planned to undertake some sub-group analyses focusing on gender and disadvantage. However, given the reduction in the quality of the assessments as well as wider changes described from phase one, we have decided to not undertake these due to the low security of the findings. Further, we do not have data from the planned post-tests about resilience so these will also be abandoned. We considered remote administration, but we feared that the response rates would be too low to enable meaningful interpretation. Table 1 summarises the changes made to the outcome measures and associated assessments.

TABLE 2: SUMMARY OF THE CHANGES TO OUTCOME MEASURES MADE BETWEEN THE ORIGINAL AND UPDATED PROTOCOL

Outcome	Protocol version one	Protocol version two
Pupil attainment		
Primary measure	50 item multiple choice science test	29 item test about electromagnetism (6 weeks into the intervention)
Sub-group analyses	Pupil premium; non-pupil premium Boys; girls	None
Secondary measure	None	38 item test about ecology (3 weeks into the intervention)
Sub-group analyses	None	None
Pupil resilience		
Primary measure	The Resilience Scale – 25 items (Prince-Embury, 2008)	None
Sub-group analyses	None	None

Analyses

We will also make changes to strengthen our analysis. The statistical models will be the same across the primary and secondary outcome, except for the change in the post-test. Both models will use analysis of covariance (ANCOVA) and ordinary least squares to estimate the parameters. Using appropriate covariates to analyse an RCT has two benefits: first, it adjusts for any chance imbalance on important variables caused by randomisation; second, it improves the precision of the estimates (Bloom, Richburg-Hayes, & Black, 2007). Each model will include three co-variates: science pre-test scores; average point scores from KS2 SATS; and a dummy variable indicating treatment allocation. To facilitate ease of analysis, we will transform all data into z-scores (Connolly, Biggart, Miller, O'Hare, & Thurston, 2017).

The model that we will construct can be expressed using equation 1. Where Y_i is the predicted post-test score, b_0 is the intercept, b_1X_{1i} is the parameter and variable associated with the pre-test, b_2X_{2i} is the parameter and variable associated KS2 average point score, b_3X_{3i} is the parameter and variable associated with treatment allocation, and ε_i is the residual.

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + \varepsilon_i \#(1)$$

ANCOVA is an extension of the general linear model so all of the assumptions for this model apply (Field, 2018). There are two further assumptions specific to ANCOVA: independence of the covariate and treatment effect; and homogeneity of regression slopes. We will report appropriate assessments of each of these assumptions. Further, as we plan to include two measures of prior attainment, there is a risk of multicollinearity. Therefore, we will assess this risk and if the correlation is higher than 0.5, we will only include the measure with the least missing data.

Security of findings

The original protocol contained limited information about how we would assess the security of the findings. Given the anticipated reduction in security, we strengthened this section to facilitate transparent, critical interpretation of the findings. In addition to the procedures described below, we will also use the findings from the process evaluation to triangulate our findings (Evans, 2009).

The original protocol included a plan to conduct null hypothesis significance testing and report p-values. We will no longer do this for three reasons: first, the assumptions for significance testing, including random sampling, are not met (White & Gorard, 2017); second, *p*-values are widely misinterpreted and mis-reported (Ioannidis, 2019); third, we think that the trial is likely underpowered, which is exacerbated by the smaller than planned sample size and the reduction in treatment duration, which increases the risk of a type II error (Shadish, Cook, & Campbell, 2002). Taken together, we do not think that significance testing will add to our understanding of the results and we plan to use other, more appropriate, tools.

We will still report standard descriptive statistics as well as an effect size (Hedges' g) to communicate the magnitude of the differences between treatment conditions (Coe, 2017). We originally anticipated having multiple opportunities to catch up on missed assessments so we thought we could achieve virtually no attrition. The changes to data collection will have likely increased attrition. Therefore, we plan to include a new measure that seeks to communicate the magnitude and robustness of the finding in a single, intuitive number. We will calculate the number needed to disturb (NNTD), which is an estimate of the number of counterfactual cases that can be added to the smallest group before the difference between the groups is reversed, minus the number of missing cases (Gorard & Gorard, 2016a, 2016b). The NNTD provides a conservative estimate as it assumes that all attrition is counterfactual to the available data. Nonetheless, we think it will helpfully communicate the sensitivity of the findings.

If the assumptions of statistical models are not met, then they can lead to misleading conclusions. When assumptions of statistical models are violated, there are four broad remedies: trimming the data; winsorizing the data; applying robust correctional methods; and transforming the data (Field, 2018). If the assumptions already described are violated, then we will conduct a robust ANCOVA as a sensitivity analysis. We will only include a single measure of prior attainment (whichever has the least missing data) because the model only works with a single predictor variable (Field, 2018).

Handling missing data

We anticipate that there will be some missing data due to the changes in data collection processes. Missing data can have a profound effect on the security of research findings (Gorard, See, & Siddiqui, 2017). Therefore, we will transparently report all missing data. We will also conduct further sensitivity analyses using default replacement of missing data (Gorard, 2020).

Research questions

The forced changes to the impact evaluation described have subtly changed the original research questions. The questions stated in the original protocol were:

1. Is LbQ a more effective method of retrieval practice in science than more traditional SMART Connect activities in Year 10 pupils, as measured on attainment scores after six months?
2. Is LbQ a more effective method than more traditional SMART Connect activities of raising resilience of learners in Year 10 after six months?

The timing of outcome assessments necessitates a change to the original research question. Thus, the primary research question becomes:

1. Primary research question: Is LbQ a more effective method of retrieval practice in science than more traditional SMART Connect activities in Year 11 pupils, as measured on attainment scores after six weeks?
2. Secondary research question: Is LbQ a more effective method of retrieval practice in science than more traditional SMART Connect activities in Year 11 pupils, as measured on attainment scores after three weeks?

As already described, note that the research questions related to resilience, as well as all sub-group analyses have been removed.

Phase two changes – process evaluation

To complement the impact evaluation, we will conduct a process evaluation. Our original protocol contained relatively limited information about the process evaluation. Therefore, we have sought to strengthen this section so that it can help to triangulate our findings from the impact evaluation as well as in response to the increasingly recognised need to pre-specify process evaluations wherever possible (EEF, 2019).

Research questions

1. What is the perceived impact of the intervention?
2. Which is the most plausible?
3. Feasibility and barriers and enablers
4. Cost-benefit in terms of the workload of the intervention
5. The control condition – including different levels of dosage

Structure

We plan to conduct a multi-phase mixed methods process evaluation whereby data from each phase influences the subsequent phases (Humphrey et al., 2016b, 2016a). There will be three phases to the process evaluation:

1. Practitioner reflections and the development of an explicit logic model;
2. Semi-structured interview with the teachers leading the intervention; and
3. Integration with findings from the impact evaluation.

Phase 1: reflections and logic model development

A logic model provides a graphical representation of how an intervention leads to the intended outcomes (Coldwell & Maxwell, 2018). Appendix 1 summarises the logic model for the intervention, which was developed following the end of the intervention, but prior to data analysis. The model is informed by both the existing research as well as practitioner reflections. Four different plausible mechanisms were identified that could lead to improvements in pupil outcomes. Note, that these mechanisms are not mutually exclusive.

Phase 2: semi-structured interview with participating teachers

We will undertake a semi-structured interview with the teachers in the intervention condition. Appendix 2 summarises the planned questions for the interview. The results from the interview will be coded thematically.

Phase 3: integration with the impact evaluation

Finally, we will integrate the findings from the process evaluation with the results of the impact evaluation as well as the wider literature.

References

Aschwanden, C. (2015, August 18). Science isn't broken: it's just a hell of a lot harder than we give it credit for. Retrieved April 4, 2020, from https://fivethirtyeight.com/features/science-isnt-broken/#part1

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using Covariates to Improve Precision for Studies That Randomize Schools to Evaluate Educational Interventions. *Educational Evaluation and Policy Analysis*, *29*(1), 30–59. https://doi.org/10.3102/0162373707299550

Boyce, C., & Martell, T. (2020). Retrieval practice in science. https://doi.org/10.17605/OSF.IO/EHDP5

Coe, R. (2017). Effect size. In R. Coe, M. Waring, L. Hedges, & J. Arthur (Eds.), *Research methods and methodologies in education* (2nd ed., pp. 5–14). London: SAGE.

Coldwell, M., & Maxwell, B. (2018). Using evidence-informed logic models to bridge methods in educational evaluation. *Review of Education*, *6*(3), 267–300. https://doi.org/10.1002/rev3.3151

Connolly, P., Biggart, A., Miller, S., O'Hare, L., & Thurston, A. (2017). *Using randomised controlled trials in education* (1st ed.). London: SAGE.

DfE. (2020). Schools, colleges and early years settings to close. Retrieved April 4, 2020, from https://www.gov.uk/government/news/schools-colleges-and-early-years-settings-to-close

EEF. (2019). *Implementation and process evaluation guidance for EEF evaluations*. London. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Evaluation/Setting_up_an_Evaluation/IPE_guidance.pdf

Evans, M. (2009). Reliability and validity in qualitative research by teacher researchers. In E. Wilson (Ed.) (1st ed., pp. 112–124). SAGE.

Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). London: SAGE.

Gehlbach, H., & Robinson, C. D. (2018). Mitigating Illusory Results through Preregistration in Education. *Journal of Research on Educational Effectiveness*, *11*(2), 296–315. https://doi.org/10.1080/19345747.2017.1387950

Goldacre, B., Drysdale, H., Dale, A., Milosevic, I., Slade, E., Hartley, P., … Mahtani, K. R. (2019). COMPare: a prospective cohort study correcting and monitoring 58 misreported trials in real time. *Trials*, *20*(1), 118. https://doi.org/10.1186/s13063-019-3173-2

Gorard, S. (2013). *Research design: creating robust approaches for the social sciences* (1st ed.). London: SAGE.

Gorard, S. (2020). Handling missing data in numeric analyses. *International Journal of Social Research Methodology*, 1–10. https://doi.org/10.1080/13645579.2020.1729974

Gorard, S., & Gorard, J. (2016a). Explaining the number of counterfactual cases needed to disturb a finding: a reply to Kuha and Sturgis. *International Journal of Social Research Methodology*, *19*(4), 497–499. https://doi.org/10.1080/13645579.2015.1126494

Gorard, S., & Gorard, J. (2016b). What to do instead of significance testing? Calculating the 'number of counterfactual cases needed to disturb a finding.' *International Journal of Social Research Methodology*, *19*(4), 481–490. https://doi.org/10.1080/13645579.2015.1091235

Gorard, S., See, B. H., & Siddiqui, N. (2017). *The trials of evidence-based education* (1st ed.). Abingdon: Routledge.

Humphrey, N., Lendrum, A., Ashworth, E., Frearson, K., Buck, R., & Kerr, K. (2016a). *Implementation and process evaluation (IPE) for interventions in education settings: A synthesis of the literature*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Evaluation/Setting_up_an_Evaluation/IPE_Review_Final.pdf

Humphrey, N., Lendrum, A., Ashworth, E., Frearson, K., Buck, R., & Kerr, K. (2016b). *Implementation and process evaluation (IPE) for interventions in education settings: An introductory handbook*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Evaluation/Setting_up_an_Evaluation/IPE_Guidance_Final.pdf

IEE. (2019). *Protocol: retrieval practice in science*. York. Retrieved from https://the-iee.org.uk/wp-content/uploads/2019/07/IEE-LBQ-Protocol-Ashington.pdf

IES. (2017). *What Works Clearinghouse™ Standards Handbook (Version 4.0)*. Washington DC. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2014). How to Make More Published Research True. *PLoS Medicine*, *11*(10), e1001747. https://doi.org/10.1371/journal.pmed.1001747

Ioannidis, J. P. A. (2019). What Have We (Not) Learnt from Millions of Scientific Papers with *P* Values? *The American Statistician*, *73*(sup1), 20–25. https://doi.org/10.1080/00031305.2018.1447512

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., … Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 0021. https://doi.org/10.1038/s41562-016-0021

Ofqual. (2020). *Summer 2020 grades for GCSE, AS and A level, Extended Project Qualification and Advanced Extension Award in maths*. Coventry. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/877842/Summer_2020_grades_for_GCSE_AS_A_level_EPQ_AEA_in_maths_-_guidance_for_teachers_students_parents.pdf

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/SCIENCE.AAC4716

Prince-Embury, S. (2008). The Resiliency Scales for Children and Adolescents, Psychological Symptoms, and Clinical Status in Adolescents. *Canadian Journal of School Psychology*, *23*(1), 41–56. https://doi.org/10.1177/0829573508316592

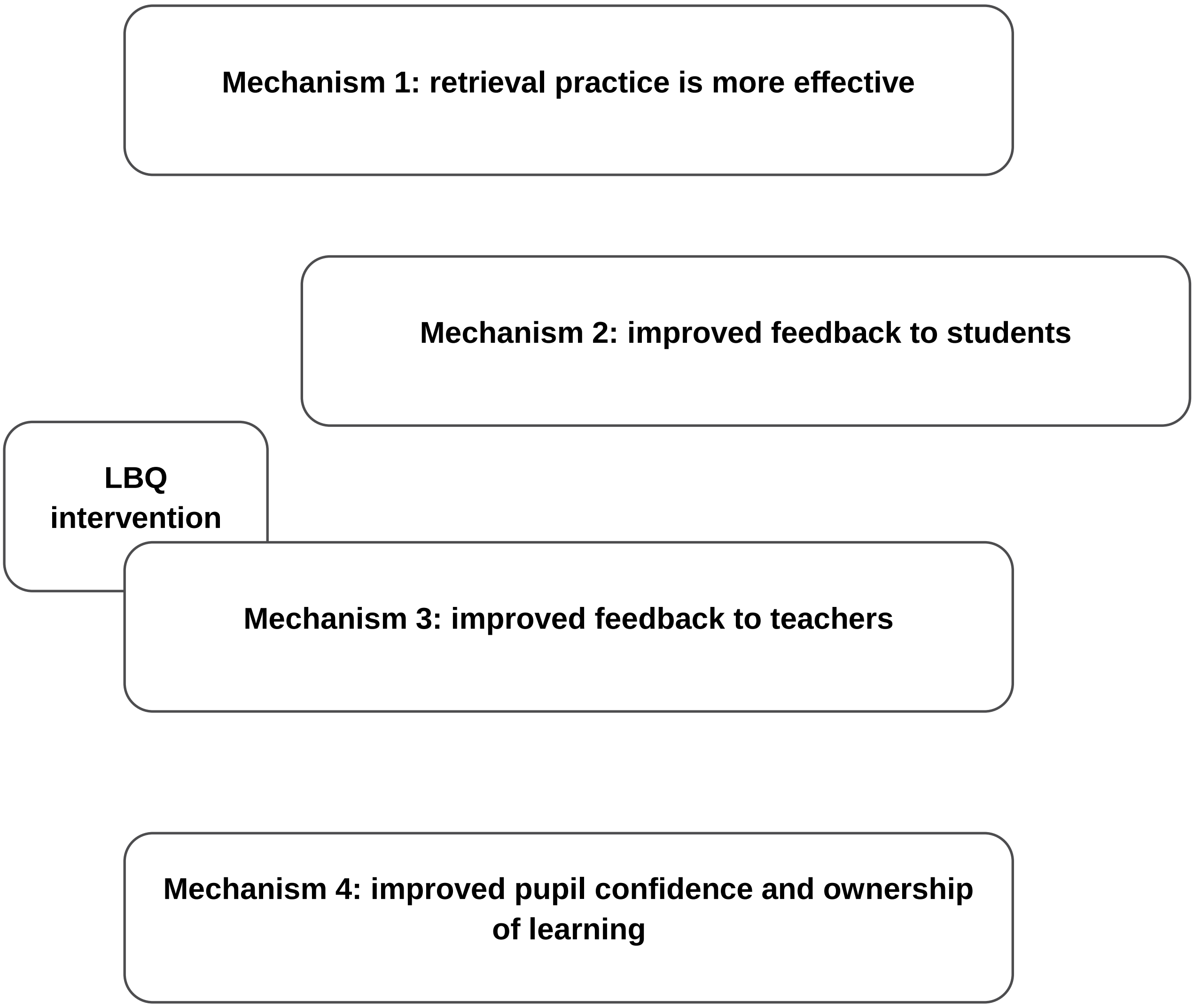
Rubin, M. (2017). An evaluation of four solutions to the forking paths problem: Adjusted alpha, preregistration, sensitivity analyses, and abandoning the Neyman-Pearson approach. *Review of General Psychology*, *21*(4), 321–329. https://doi.org/10.1037/gpr0000135

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalised causal inference*. Belmont: Houghton Miffin.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., … Nosek, B. A. (2018). Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

White, P., & Gorard, S. (2017). Against Inferential Statistics: How and why current statistics teaching gets it wrong. Retrieved from https://lra.le.ac.uk/handle/2381/37564

Appendix 1: logic model



Appendix 2: teacher interview questions

Technology

1. Did you feel you were adequately trained for LBQ technology?
2. Did you feel you were using LBQ to its full potential?
3. How reliable were the tablets?
4. How well were the students able to use the tablets?
5. Were the students efficient in entering the class, retrieving a tablet and getting on with the task?
6. How could the tablets be improved?

Appropriateness of intervention

1. Were the questions of an appropriate quality?
2. Was the literacy demand of the questions appropriate?
3. Did the question structure vary?
4. Were the questions challenging?
5. Were the questions relevant to the topic being studied?
6. Were they relevant to the exam board?

Perceived impact

1. How did you use the data generated from LBQ?
2. Did you find the approach to be efficient, both in terms of students and teacher time?
3. Did you find the intervention an effective alternative to traditional SMART retrieval practice?
4. Did you see any difference in perceived resilience of the students? i.e. did they try more than usual? More likely to answer the questions?
5. Did you see any change in academic achievement of the students?
6. Would you use LBQ again?

Contact us

+44 (0)1904 328166 | info@the-iee.org.uk
Ron Cooke Hub, University of York, York YO10 5GE
Twitter: @IEE_York | the-iee.org.uk/

The Institute for Effective Education (IEE) is an independent charity working to improve education for all children by promoting the use of evidence in education policy and practice.

Learning by Questions is a development company financed and owned by the Bowland Charitable Trust

The Institute for Effective Education is a charity registered in England, charity number 1168744