Do staff and pupils believe that structured retrieval practice in KS4 Science prepares the pupils for their end of year and GCSE examinations?

Andrew Reeve

MSc in Learning and Teaching, 2024

DECLARATION BY THE CANDIDATE AS AUTHOR OF THE DISSERTATION

1. I understand that I am the owner of this dissertation and that the copyright rests with me unless I specifically transfer it to another person.

2. I allow the Department to deposit on my behalf a copy of this dissertation in the Oxford University Research Archive ('ORA') where it shall be freely available online for use in accordance with ORA's Terms and Conditions of Use [https://ora.ox.ac.uk/terms_of_use].

3. I understand that this dissertation should not contain material that can be used to personally identify individuals or specific groups of individuals (unless permission has been obtained from the individuals) and that such material should be removed before this dissertation is deposited in ORA.

4. I agree to be bound by the terms of the ORA Grant of Non-exclusive Licence [https://ora.ox.ac.uk/deposit_agreements] and I warrant that to the best of my knowledge, making my thesis available on the internet will not infringe copyright or any other rights of any other person or party, nor contain defamatory material.

5. I agree that my dissertation shall be available for download in ORA in accordance with paragraphs 2, 3 and 4 above.

Signed [an electronic signature is sufficient]:	A P Reeve
Date:	17.09.24

Do staff and pupils believe that structured retrieval practice in KS4 Science prepares the pupils for their end of year and GCSE examinations?

1. Abstract

Retrieval Practice is a widely used pedagogy in schools across the United Kingdom. It is promoted by OFSTED and the Department for Education and often mandated by Senior Leaders within schools. Research shows that Retrieval or the "Testing Effect" enables pupils to develop their long-term memory, which subsequently leads to greater outcomes in formal tests and examinations in the future than when compared to re-reading. Although there is much research which suggests this to be the case, much is produced in laboratory conditions and not within the classroom. There is also much less research into the perceptions of both pupils and teaching staff as to whether they find this form of pedagogy supportive of their examination outcomes.

This research therefore investigates the perceptions of both pupils and teaching staff within an academic department to the introduction of a retrieval intervention at the start of most lessons in Year 10 and Year 11. It examines the use of cues and feedback as well as analysing if pupils perceive that they overly rely on re-reading during their revision.

2. Introduction

The retrieving of information from memory is a powerful way to enhance learning (Carpenter 2023). It is considered an approach to develop long term memory (Hostetter et al. 2019) and supports the development of the transfer of knowledge to new situations (Agarwal et al. 2017). Originally termed the “testing effect” (Roediger & Karpicke, 2006) it is defined as a pedagogical method in which pupils are tested in the form of low stakes tests on a regular basis, encouraging them to recall previous knowledge that has been covered in the course or programme of study. Roediger & Karpicke (2006) concluded that retrieval led to a greater retention of knowledge compared to simply re-reading notes during the process of revision. Therefore, introducing regular testing of knowledge in a structured manner can not only be used for assessment purposes by teachers, but also to develop the learning potential of the pupils too.

Retrieval Practice has become a widely used pedagogical method embraced and indeed, at times, enforced by Senior Leaders in schools across the United Kingdom. It is a method which is supported by the Department for Education and OFSTED. (OFSTED, 2019) Retrieval Practice forms part of the National Professional Qualification for Headship (NPQH) resource material, in the wider context of promoting Rosenshine’s Principles (Rosenshine, B. 2010; DfE, 2020). Effective retrieval is often cited as a positive pedagogical method in OFSTED reports, with their own research highlighting how it leads to the retention of knowledge for pupils.

As a newly appointed Headteacher, in a school that did not have a teaching and learning policy, I was keen to explore different pedagogical methods that we might adopt. The value-added score at GCSE was low in the last academic year and there needs to be an overall improvement in this in the next few years. Although there are several factors causing the low progress scores, the need for greater discussion of pedagogy and the sharing of good practice is an essential aspect of the school's improvement plan. The Head of Science was a keen advocate of evidenced based learning and we collaborated, researching various aspects of learning and teaching which might be adopted as a trial in the Science Department. Both the Head of Science and I are affiliated to the Chartered College of teaching; I was appointed a fellow in 2020, and we discussed various topics that have been written in their termly magazine together. The Head of Science had completed the National Professional Qualification for Senior Leaders (NQPSL) qualification for aspiring senior leaders, and I have recently completed the NPQH. As part of the specification of these qualifications is Rosenshine's Principles of Instruction and we were drawn to piloting structured retrieval practice in lessons. Although Rosenshine's research is not restricted to retrieval, the concept of recall is a common thread through his Principles of Instruction and appears to be effective pedagogy in the classroom (Cowan & Farrell, 2023).

I felt it important that the whole of the science department agreed to this trial, I did not want to enforce it upon them but wished for them to willing agree and buy into the trial. At a meeting in January 2024, there was an overwhelming willingness by teaching staff within the department to embrace this trial. There was a keen desire to develop learning and teaching within the department, share resources and depending on the trial consider recommending retrieval practice to be adopted across the whole school

curriculum. The teaching team were keen to focus on one Key Stage only, to limit any perceived increases in workload, and they selected Key Stage 4, as they believed that the student voice from this age group would be more informed. The age group was their choice, and I felt strongly that they should be empowered as part of this collaborative project and agreed to that.

During my first year in post, I began to work on the development of a teaching and learning strategy, working alongside members of the senior leadership team as well as with the cooperation of middle leaders in the school. A discussion was started which placed teaching and learning as a renewed priority in the school. In the Autumn Term the Self Evaluation Statement (SES) was written outlining the academic priorities for the academic year with the development of a teaching and learning strategy as a primary focus. I am keen on researched based learning and members of the senior team, and I read widely on different pedagogical concepts, including lesson study, starters, managing behaviour for learning and, of course, Rosenshine's Principles of Instruction.

What became apparent in my initial reading was the wealth of literature on the merits of retrieval by pupil outcome, but little in the way of measuring the pupils' own perceptions of this learning and teaching method. This research therefore focuses on measuring pupils' perception of the intervention, asking them to judge its usefulness and considering whether it prepares them for their GCSE and end of year examinations. It also seeks to understand teachers' perceptions too; is this a useful pedagogical intervention or a distraction within the curriculum. My research considers the preferred learning styles of pupils and enquires as to whether they use retrieval within their revision or whether they rely on re-reading strategies. This study is also designed to

consider pupils perceptions of being supported by hints and clues during their retrieval tasks as well as understanding their perception of how important effective feedback is in aiding their learning. The wider reading will allow for appreciation of the effectiveness of retrieval practices in developing memory recall within pupils, with consideration of additional variables such as the complexity of the retrieval and the durations between retrievals; known also as inter-leaving. Gaining the pupils perceptions, along with those of staff, will allow any new pedagogical methods to be evaluated thoroughly.

The results of this research project will be shared with middle leaders and then the whole staff as we drive forward our shared teaching and learning strategy over the next few academic years. I intend to develop teaching and learning cluster groups within the school, encouraging teachers to share good practice and try out new pedagogical ideas in class. I am also keen that, at all times, the views of pupils are sought and then listened to.

3. Literature Review

3.1 Introduction to Retrieval

The literature on the topic of testing within schools is wide ranging and not a relatively new area of research. Large-scale testing studies were taking place at the beginning of the twentieth century Gates (1917). Gates explored how the testing of recitation within a Music curriculum enhanced the outcomes of the pupils. He acknowledged that testing "recitation results in greater activity and expenditure of energy than reading" (Gates, 1917, p87) but in an evaluation linked to the human brain he concluded that recitation, or testing, stimulated, or created a "nervous impulse" which "flows along the most frequently traversed pathway" (Gates, 1917, p94). It would be decades before such research into the firing of pathways in the brain would become more widely researched. McDaniel and several writers engaged in a series of research projects in the late 1980s to investigate how retrieval and retesting enhances learning by developing memory paths and traces within the brain (McDaniel & Masson, 1985; McDaniel et al, 1989).

The 1980's and 1990's saw a revival of research into testing and memory development. Many of the studies were linked to university undergraduates (Glover, 1990; Cuddy & Jacoby, 1992) where the analysis of spacing repetitions in testing were seen to be beneficial to memory development. At times, the studies were less convincing, with limitations leading to conclusions that could, at best, state that retrieval "may have special processing properties" (Glover, 1990, p54)

But it was from the beginning of the twenty first century that the literature of a possible "testing effect" began to significantly expand and from 2010 onwards that this research began to be more grounded in classroom practice within schools rather than through

laboratory-based research within universities. This applied research has increasingly led to the term “Retrieval Practice” being used instead of “The Testing Effect” such that “retrieval practice” became a more commonly used search word than “The Testing Effect” on Google from 2019. (Agarwal, 2023, p3.) This change in the popularity of the search term is underpinned that the nature of the expanding research with testing or retrieval being seen as far more beneficial than simply using testing as an assessment device. It is now increasingly seen to be a learning strategy in a far wider sense that simple assessment monitoring and recording.

A seminal paper was published in 2006 by Henry Roediger III and Jeffrey Karpicke, two American educational researchers. This influential paper, based on research into the learning habits of 120 Washington University undergraduates, concluded that restudying a topic leads to greater short-term memory retention that “free recall resting” but, in the longer-term restudying leads to a higher incidence of forgetting concepts compared to those students subjected to testing. (Roediger & Karpicke, 2006) Roediger & Karpicke defined the short term as 5 minutes and the longer term as one week, indeed the three testing intervals were 5 minutes, 2 days and a week. The paper was set in controlled conditions, with two experiments, testing the retention of two passages of text over time. The results were significant, showing that over the duration of a week, students who undertook repeated testing retained 71% of the information while those who had undertaken the rereading pedagogical method could only recall 40% of the material. (Roediger & Karpicke, 2006, p 255). Their research did, however, show that restudying produced greater recall in the short term.

Both researchers continued to write on the concept of retrieval after this initial paper. Roediger, in a paper based on testing university students in controlled conditions, concluded that retrieval produces greater levels of retention than studying. (Roediger III & Butler, 2010). The notion of retrieval being used for wider educational purposes rather than it simply being an assessment tool was an area that saw an exponential growth in research in the first twenty years of the twenty first century. Although the "testing effect" had been explored a hundred years earlier, the amount of research was limited. Retrieval practice was found to promote meaningful learning compared to other pedagogical methods such as restudying. Karpicke & Blunt (2011) conducted a test involving 120 undergraduate students using laboratory-controlled methods. Here the paper investigated the relative effectiveness of both retrieval practice verses students designing concept maps as an aid for memory retention. The results were significant with "101 of the 120 students (84%) performing better in the final test after using retrieval practice" (Karpicke & Blunt, 2011, p3). What is powerful with this report is that 75% of the students believed that the use of concept mapping would lead to better academic outcomes compared to the use of retrieval.

Although the papers at this time tended to focus on university level education, clearly a more limiting factor when wishing to explore retrieval in schools, there was the beginning of a body of research that was exploring in class outcomes. Previously, much of the work focused on verbal memory testing, with the use of a non-western language at times being used in tests. University controlled testing of translation word pairs in Swahili and English have been adopted in several tests in order to explore the importance of retrieval (Nelson & Dunlosky, 1994; Pyc & Rawson, 2009).

Having gained a historical context for retrieval practice, I was keen to understand further the benefits and disadvantages of testing, both low and high stakes, within the curriculum.

3.2 Testing in Education

"Tests are a necessary evil" (Minear et al, 2018, p.1474). They are used extensively in educational settings as a means of measuring student performance. Tests of all descriptions are used within educational settings to assess the level of students' knowledge (Storm et al, 2010). Teachers use testing scores to ascertain the progress made by students and this can inform their future planning of lessons. Testing also defines the outcomes of students; it allows them entry into a school or course, determines their streaming or setting group, provides access to special education or exam access arrangements and determines their final outcomes from public examination success.

Yet, although testing is used extensively by educational establishments in this manner, it is fraught with limitations and unintended consequences. Testing can only provide a "partial lens on what a child knows and is able to do" (Dutro & Selland, 2012, p. 343). It can exacerbate social and cultural inequalities, those taking school or college entrance assessments are often limited by racial, socio-economic or political inequalities (Sororzano, 2019). This then creates a cyclical and inter-generational cultural of inequality within schools and society as a whole.

There are many concerns relating to the use of testing and the increased levels of mental health anxiety within young people today (Putwain, 2008; Howard, 2020). Wellbeing of young people has become a prominent public policy issue in our society.

The outcome in public examinations can deepen these anxieties, reducing self-worth and belief in yourself (Jerrim, 2022). High-stakes testing can contribute to school-related stresses amongst young people. Failure in high-stakes tests may create psychological distress, enhance drop-out rates within schools and colleges and directly correlate with the incidence of graduating later in their educational careers (Beck et al, 2024).

Test failures and outcomes are documented in research, but it is also useful to consider the build up to a test in the mind of a student. Adam et al (2017) researched the levels of cortisol in children ages 8-16 years old as they approached a high-stakes test. Their analysis showed that there were heightened levels of cortisol within children during examination weeks, with significantly heightened levels within the body during the morning of the actual test. This indicated significant physiological stress levels within the body which potentially translate into depressed test outcome results.

However, high-stakes testing can lead to better channels of communication between school and home, especially related to those children who are classed as more vulnerable. The results of such testing allow parents to seek answers to more appropriate questions, enables the collaboration between staff to support individual children and helps to channel limited resources to those areas within the curriculum where the largest progress gains can be made (Ysseldyke, 2004).

Whereas high-stakes testing invariably is undertaken for summative assessment purposes, low-stakes testing provides an opportunity to use the data for formative assessment, providing pupils with data informed feedback as well as supporting teachers in future planning (Adesope et al, 2017). The use of low-stakes testing on a regular basis develops the formation of memory traces, in other words a long-term

change in the brain which has developed memory, and allows for the retrieval of knowledge and enhanced memory ability (Hostetter et al, 2019). Low-stakes testing in class often provides immediate feedback opportunities, through a blend of peer marking and teacher led feedback. Bangert-Drowns et al (1991) found that pupils undertaking a number of low-stakes tests, obtained higher scores than those completing fewer, less regular, high-stakes tests. Several researchers have concurred with this, while researching the duration between retrievals and the number of optimal retrieval testing points. That said, their claim that "everyone agrees that repeated questions produce a large improvement on examination scores" is rather generic and simplistic and needs far greater evaluation (Bangert-Drowns et al, 1991, p. 90). A more balanced approach is presented by Wise et al (2024) who reflected on how the effectiveness of low-stakes testing relates to several factors. In particular, they considered the timing within the school day, concluding that disengagement rates fluctuate during the day, with heightened disengagement after lunch. Although their study did not consider high-stakes testing, which would have been a useful comparison, it mirrors general dips in engagement immediately after lunch which is evidenced by many in the profession.

Hinze & Rapp (2014) further considered the impact of anxiety with both high-stakes formal testing and low-stakes retrieval quizzes and tasks. Those suffering from high levels of test anxiety tend to retain less information than those who do not. High anxiety is associated with poor examination performance. Their conclusion addressed how even though low-stakes retrieval testing is a "powerful tool for supporting learning" (Hinze & Rapp, 2014, p. 605) its effectiveness is somewhat reduced by the presence of anxiety levels within learners.

It is clear from the literature that testing in school has the potential to develop memory retention and recall, but it can also cause heightened levels of anxiety and stress. Therefore, I was keen to read more widely around how often the retrieval testing should take place, so that recall is developed without the need for unnecessary and excessive testing taking place.

3.3 Duration and Number of Retrievals

There is a growing body of research which considers the duration and number of retrievals being undertaken. The foundation of retrieval is that repeatedly testing yourself or others improves the retention of knowledge from working memory (Wiklund-Hornqvist, 2020). This general conclusion is supported by a growing body of research (Roediger III & Butler, 2011; Roediger III & Karpicke, 2006; Rowland, 2014) who write of the developments within the testing effect.

Roediger III & Karpicke (2006) clearly concluded that retrieval was more beneficial when a period of time, in their study one week, has elapsed. In their laboratory-controlled experiment, the recall after 5 minutes was higher for repeated restudying than for repeated testing; with 83% successful recall compared to 73%. Yet, after a retrieval duration of one week this was reversed with recall rates of 40% and 61% respectively (Roediger III & Karpicke, 2006, p. 255). Roediger III & Butler (2010) developed this further, asking a central question on the number of retrievals needed to make a meaningful difference to recall and retention levels. Their overall conclusion was that the impact of long-term retention following retrieval is optimised when the retrieval occurs multiple times, with relatively long intervals between them and each time it is followed up by feedback. Their research, a controlled experiment on Swahili-English

word association might be criticised as lacking classroom-based activity, but developed the work by Roediger III & Karpicke several years earlier. A similar word association experiment, translating Swahili into Swedish took place with 50 pupils in Year 9 in a Swedish secondary school (Wiklund-Hornqvist et al, 2020). Retrievals on a daily basis, over a period of a week, showed a continuous increase in the proportion of correct recall from less than 30% before the intervention to over 85% after it. Their research, an Umea University, Department of Psychology, research paper, provided a strong case for retrieval, linking the findings to how the anterior hippocampus develops linkages with successive attempts at a task.

Similar research experiments studying undergraduate students, concluded similar results. Storm et al (2010) investigated the recall from a 200-word text on Antarctica, suggesting that expanding the retrieval practice reduced the level error when recalling information. A similar study, but involving adults, examined colour-image association, in an experiment involving an online observation test. The results were more subdued in favour of retrieval, but did suggest that the benefits were only evident when the pattern of retrieval testing occurred for more than three occasions (Brendan et al, 2019). Tse & Pu (2012) similarly found subdued benefits of retrieval practice, which they attributed to the small gap of seven days between learning the concept and the retrieval of the information. Their findings were later echoed by Agarwal et al (2016), during experimental testing of 166 undergraduate students at Washington University. In a controlled general knowledge experiment, their results concluded that retrieval with longer lags between learning and testing led to more positive recall outcomes. These papers were, however, countered by de Lima et al (2020) who researched the outcomes of retrieval within language impaired populations, reaching a conclusion that although

longer lag times benefit recall rates, the greater repetition of the retrieval "may lead to better performance at short intervals" (de Lima et al, 2020, p.1088).

It can therefore be concluded that there is not universal acceptance of both premises relating to the positive correlation between time lags, number of retrievals and more positive outcomes on recall. Vaughn, Hausman & Kornell (2017), present a convincing case that there are clear educational benefits of retrieval, but "more time spent in retrieval mode does not confer additional benefits in final memory performance" (Vaughn et al, 2017, p. 313). Their experiments involving adults aged 19 to 69, involved general trivia questions. However, their self-reflection relating to the limited usefulness of one highly-controlled experiment, applied to the high levels of variability in the average classroom, is much to their credit.

3.4 Feedback and the use of cues

The fidelity of the retrieval task and how the teacher uses it, will have an impact its success. One area to consider is the impact of using feedback during the retrievals to correct incorrect answers and provide further explanation. Leggett et al (2018) write that there "is little doubt that retrieval practice should be used by teachers and students in some cases, particularly when timely feedback can be given" (Leggett, 2018, p. 760). Feedback from the teacher or a peer, during a learning activity, especially where the correct response is provided to the pupil, will allow for self-correction of errors and a higher incidence of correct answers in latter retrieval tests (Roediger & Butler, 2010). This is beneficial to the pupil as a failed retrieval attempt may not lead to later recall (Kornell & Bjork, 2007). One premise is that the effect of retreival of long-term memory is stronger when the pupil is provided with feedback (Kang et al, 2007; Hostetter et al,

2019), and it reduces the chance of failed retrievals leading to incorrect information being propagated into the future (Storm et al, 2010).

Yet, despite the widespread acceptance of effective feedback being important in enhancing progress in pupils learning (Bangert-Drowns et al., 2023; Carpenter,2023; Leggett et al., 2018; McDermott & Roediger III, 2011; Roediger III; Butler, 2011 & Storm et al., 2010), there is not a consensus within the research on feedback during retrieval. Adesope et al, (2017) acknowledged their surprise in their findings that feedback did not improve the strength of the retrieval testing effects during their experiments, despite the "plethora of findings that testing plus feedback is more beneficial than testing without feedback" (Adesope et al, 2017, p. 686). However, in their meta-analysis they accepted the need for further research to consider the type of feedback being provided with more focused analysis on the frequency, speed and medium of the feedback. Other research has resulted in the same unexpected results too (Smith et al. 2016). Kornell et al, (2011) undertook word pairing experiments involving a cohort of American adults. Their premise was that feedback following retrieval would lead to greater outcomes compared to retrieval on its own. Yet, their results suggested that this was not the case. However, they provided no rationale as to why this is the case, which would be useful for developing pedagogy in the classroom. Slightly more helpful for pedagogical development is the study by Agarwal et al (2017). Again, the premise had been made that retrieval with feedback would lead to enhanced rates of recall. Yet, their results highlighted their surprise that the feedback did not increase the recall beyond that of retrieval testing without it. However, they did conclude that retrieval with feedback was disproportionately supportive to recall outcomes for those pupils with lower working memory capacity when there was a regular two-day retention interval. Their

conclusions therefore highlight the use of retrieval practice with feedback to somewhat "level the playing field for lower capacity students" (Agarwal et al, 2017, p.770).

If feedback is the intervention following the retrieval then cues are the intervention during a retrieval task to support the progress of learning. Cues within retrieval have been the focus on various papers. In a word pairing experiment involving university students, it was noted that the recall outcomes from retrieval were greater in the group which were provided with cues and undertook a cued-recall test at the end (Pyc & Rawson, 2009). Similar conclusions were supported by the others researching the support provided by teachers during retrieval (Rowland & DeLosh, 2015; de Lima et al., 2020)

Although interesting, the results are more challenging to translate into secondary school classrooms with little in the way of controlled conditions, and where accreditation towards the final course outcome is not derived partially from volunteering to take part in the experiment. A more helpful study involved a retrieval task for 89 pupils as part of Year 9 Geography lessons in Brisbane, Australia; a study more closely linked to mine. The conclusion was balanced; that the effectiveness of cues depends on their relevance and the fidelity of providing them, but they had the potential to increase recall (Leggett et al. 2018). The study also reviewed the use of feedback with the Year 9 pupils too; finding that the implementation of feedback increases the effectiveness of the retrieval practice.

Leggett et al. (2018) also reflected on the results by Carpenter et al. (2016) which argued that retrieval practice benefitted those pupils with higher ability more than

those with a lower academic ability. Yet, the results from Leggett et al countered this. This was accredited to effective feedback being provided.

3.5 Re-reading and Re-Studying

"During self-regulated learning, learners across different age groups are less inclined to use retrieval practice than restudy" (Hui et al., 2021, p. 433). Yet, the evidence from research over a sustained number of years mostly concludes that retrieval practice tasks are more beneficial to the pupil in terms of later recall ability (Roediger III & Karpicke, 2006). There may be several personal reasons for this choice of revision technique. First, there may be the perception by pupils that retrieval practice tasks, as part of a self-regulated approach to revision, are more time consuming and that the outcomes do not necessarily lead to higher recall outcomes in the examination. If this is the case then there is a need for further study skills education which explains the benefits of retrieval, ensuring the pupils understand that the perceived increase in effort through engaging in self-regulated retrieval will lead to enhanced outcomes later on. Although their research was at a time of remote learning during COVID, which reduces the relevance to classroom practice, the results from Hui et al. (2021) clearly show that retrieval is more powerful that re-studying and leads to greater rates of pupil recall. Other researchers also conclude that pupils prefer restudying and underuse retrieval in their own self-regulated learning (Toppino, 2018; Dirkx, 2019) often neglected to restudy a topic once they believe that they know it well. In laboratory experiments involving word pairing tasks, university students were provided with an opportunity to restudy words they were not sure about. Once they had successfully learnt the word, a large proportion of the cohort elected not to restudy it again. Yet, in the final testing it was evident that participants had over-estimated their long-term memory ability (Toppino et al., 2018,

p.1166). The research also investigated how learners tend to learn and revise the topics which they perceive to be more simplistic first, although a reason for this was not provided. This is replicated by others, who write “our hypothesis is that students believed, wrongly, that there was little to gain from restudying an answer once they could recall it, even if the answer might otherwise be forgotten (Kornell & Bjork, 2007, p.221).

Yet, there is a counter argument to this, that learners believe that better outcomes will be derived from putting in more effort during self-regulated revision, both at home and at school or college and that a perception may be that more time taken to restudy will lead to better outcomes (Baars et al., 2020). Blasiman, Dunlosky & Rawson (2016) write succinctly on this area, considering how self-regulated study habits are often ineffective. Their study of university students in the United States concluded that “on average, students appear to be relying on ineffective study strategies” (Basiman et al., 2016, p. 789). However, they found that students did differentiate between the use of re-studying involving self-regulated tasks, and simply rereading the textbook. With the latter, there was the perception amongst pupils that simple rereading was less effective in terms of later recall outcomes. This leads us to the general conclusion that learning to learn, and pupils’ understanding of study skills strategies, is crucial in order to maximise outcomes in class work and in examination results. As with many other countries, schools the United Kingdom spend a significant proportion of time in preparing pupils in class for tests, embracing retrieval activities and active learning, yet, do not allocate sufficient time to teaching the importance of effective self-regulated study (Voskamp et al., 2020). This lack of training can lead to errors in judgement as has been stated earlier. Kornell & Bjork (2007) also refer to resilience and mental health too, implying

that pupils may stop studying a topic if they are finding it challenging and cannot succeed at it. In other words, it is less damaging to their mental health if their perception is that they have given up rather than failed at the task. This lack of resilience is debated considerably in education, yet often there are not easy answers and solutions to improving it.

It is also worth dwelling on re-reading as a strategy at this point too, comparing it to the effects on recall linked to retrieval. Rereading is a common form of self-regulated study and it is often set in a homework task as a way of preparing for an upcoming test or examination. Rereading a textbook is one of the most popular ways to gain information on a topic and, according to Callender & McDaniel (2008), it leads to improved memory in the longer term. Yet their research failed to highlight the impact of rereading on different ability levels which would have been more relevant to classroom pedagogy. Research shows that compared to sitting a test without any form of study, a learner who undertakes a form of rereading will benefit from doing so, and rereading on two occasions will further benefit recall later on (Rawson & Kintsch, 2005). Their study highlighted the usefulness of regular and distributed rereading throughout a programme of study. Indeed, immediate rereading of a text or article appears not to have such an impact on recall enhancement (Callender & McDaniel, 2008). A large majority of learners prefer rereading when compared to testing as part of their self-regulated study strategy (Karpicke et al., 2009; Kuhbandner & Emmerdinger, 2019) although Kuhbandner & Emmerdidinger acknowledge that the confusion between pupil's perception of the words rereading and restudying creates a limitation within the analysis. Moreover, rereading is often used by pupils as an initial study strategy, before moving onto use some form of self-testing later on.

Although metacomprehension accuracy can improve with rereading there is some belief that it might diminish in its effectiveness with age (Stine-Morrow et al. 2004), whereas other research highlights how successive rereading attempts can create diminishing outcomes over time (Kuhbandner & Emmerdiger, 2019).

3.6 Laboratory Controlled vs Classroom Practice – is retrieval used correctly?

The early research papers, particularly from the turn on the century, tend to involve research projects based on laboratory-controlled testing as well as the investigation of the learning habits of undergraduate students at university. As Carpenter writes "Nearly everything we know about the testing effect has been derived from the way that adult college students learn information in relatively well-controlled environments" (Carpenter et al. 2009, p.761). Although this certainly has validity and can be used as a baseline within classroom research, it also acts as a limitation when considering classroom pedagogy.

In the classroom, testing is often used as a tool for measuring learning and it is only relatively recently that research-based pedagogy has supported the idea that they can be also used for long-term memory development. Whereas high-stakes testing for assessment creates uncertainty and potential heightened stress for the pupil, the use of low-stakes testing tends not to. School-based practitioners have adopted low-stakes retrieval activities, developing active learning pedagogies which are potentially enjoyable for the pupil.

These are also "impressive from the standpoint of the minimal adjustments required to incorporate them into the classroom" (McDaniel et al. 2011, p.410), any form of new pedagogical development will require some adaptation in lesson planning, which in turn

may potentially impact teacher workloads. Such low-stakes testing may involve online quizzes, such as Kahoot or Quizlet, or the use flashcards and matching games. Often these are marked automatically which has the additional benefit of allowing the teacher to manage their workload too.

Agarwal et al. (2012) conducted a collaborative research paper involving a teacher, principal and a university researcher. Their focus; the use of retrieval and its impact on over 1400 students within an American middle school. Their research aim was to extend the laboratory-based research on retrieval and explore if its benefits extend to all pupils within a school setting. Experiments involved comparing recall rates for pupils following quizzing or rereading. The results highlighted the benefits of low-stakes testing, with the incidence of recall being 91% for pupils who undertook quizzes compared to 83% for those who undertook rereading for their revision. (Agarwal et al. 2012, p. 438). Their research took place over a period of several months and involved different experiments, including one with a sample of 139 science pupils, who completed an end of semester test having undertaken quizzes or not, depending on their group. The results showed that the use of quizzes in the lead up to the test had beneficial recall results. Overall, at the end of the experiments, one of the most powerful outcomes was the level of collaboration that then followed, between leaders and teachers in different schools, designing pedagogies to enhance learning. There is, of course, an ethical consideration here, as those pupils in the control rereading groups might be considered to be disadvantaged and there needs to be some consideration for narrowing the outcome gap for them in the future.

Two further studies of interest relate to retrieval within the Science Departments at US middle schools. I was particularly drawn to these as the research being undertaken in this paper relates to science teaching too. Rowley & MCrudden (2019) tested 39 pupils within Year 7 and Year 8 classrooms in New Zealand. Their results confirmed the laboratory-based experiments previously researched by Roediger III, Karpicke and others; those pupils undertaking retrieval testing had enhanced recall rates in subsequent tests. Their research was, however, limited by a small sample size, caused by high drop-out rates and an initial high incidence of not providing consent. An earlier study, focusing on both science and social studies classes, highlighted the effectiveness of low-stakes retrieval testing on the development of retention, and the improvement in pupil's metacognition generally (McDaniel et al. 2010)

3.7 Retrieval in the classroom: A Pre and Post Covid Comparison

A recent blog on the Education Endowment Foundation website, entitled "retrieval practice: A Common Good or just Common Place? (Pearce, 2021) examined the effectiveness of class-based retrieval on the longer-term increase in examination grades. Although there is a general tendency to concur that retrieval is beneficial to the pupil, there is also a belief that it might only develop and improve the lower-level skills of knowledge and potentially application due to the common practice of using it to recall knowledge. Therefore, the skills of analysis and evaluation; skills which are often allocated more weighting in assessment criteria, are not affected as much.

The fidelity of delivering the retrieval task is also vital. Any form of pedagogy, however effective it might be, is only as good as the teacher who is delivering it. Failure to motivate the pupils or deliver the lesson in an engaging or relevant manner will limit

outcomes, particularly when teaching those with lower levels of self-regulation. Effective teaching essentially requires more than a practitioner with expert subject knowledge, it also requires an understanding of how to facilitate learning too (Anderson et al, 1991).

In recent years there has been a growth in the materials published for teachers on retrieval, ranging from lesson planning to resources. Authors such as Tom Sherrington and Kate Jones have published books providing support to teachers on this topic (Sherrington, 2019; Jones, 2019) There have also been a number of academic studies written on differing methods of retrieval than can be embedded into curriculum planning, all designed to build memory storage. Desy et al. (2018) focused on this last point of memory strength, considering the methods that could be adopted in the classroom. Relating their research to Bjork & Bjork's new theory of disuse, they investigated how embracing retrieval within the classroom can develop memory and reduce the incidence of forgetting information related to the subject (Bjork & Bjork, 1992). Furthermore, they explored how effective feedback extended this memory development, although, they expressed perceived benefits for minimising this feedback for more able learners. This is, of course, an example of adaptive teaching, a concept promoted by educators and praised by the Department for Education and OFSTED (DfE, 2022; OFSTED, 2019).

Two papers of note investigated different types of studies within the classroom (Carpenter, 2023; Bae et al. 2019) Their dates allow for some simplistic form of comparison of retrieval methods, before and after the COVID-19 global pandemic. Writing originally in 2017, Bae et al. highlighted that the most common form of retrieval

practice is “free-recall” where pupils write down as much as they can without notes or prompts. They also refer to the high incidence of multiple-choice retrieval tests as well as informal quizzes. Their conclusions draw attention to the desire for practice quizzes linked to free recall as an effective retrieval task, although the paper is limited in its scope as it only considered short retrieval durations over a week. Moving forwards four years, the study by Carpenter (2023) examined five types of retrieval activities. The strength of this study is that it highlighted the effectiveness of retrieval within a “authentic class environment” (Carpenter, 2023, p.96). Interestingly, the time difference between 2017 when Bae submitted the study and 2023 when Carpenter produced the paper, saw a rapid growth in online education and gamification. Online educational quizzes such as Kahoot, Quizit and Quizlet had become far more embedded as commonly used pedagogy in the classroom. The period of forced online learning, in both 2020 and again in the early months of 2021 had seen considerable growth in both the mindset of teachers and pupils towards the use of gamified learning. Therefore, Carpenter’s study is set in this post-Covid educational landscape. It acknowledges the need for flexibility within teaching as well as the need for pupils to enhance their level of self-regulation. Interestingly, the results of the paper were linked to outcome levels following the final examination – yet there was no consideration of the pupil’s perceptions of the retrieval activity and this was promoted as an academic priority for future studies.

Throughout this research the discussions relating to retrieval tended to place the research findings into one group. As my setting is a non-selective school, I was keen to understand how retrieval might benefit those of differing academic abilities. The next section of my literature review focused on this.

3.8 Difficulty Levels within Retrieval

In general, evidence suggests that pupils, indeed all humans, will tend to study the easiest and topics first during periods of self-regulated learning (Toppino, 2018). There is also the tendency to give up after trying a topic several times and failing to fully understand it. Although this is not the focus of research here, it does create an interesting question for educators; how do we encourage pupils to learn the more challenging topics as a priority? Baars et al. (2020) suggests that pupils believe that they are more likely to remember those items which are easily remembered compared to those that are more challenging and require more time during revision. Although a generalisation, there is a tendency for pupils to spend a greater proportion of time revising the easiest topics during their self-regulated study, at times focusing on those areas that potentially do not carry so much weight in the assessment criteria (Bjork, Dunlosky & Kornell, 2013). This can significantly limit educational outcomes in high stakes testing.

The cost of this failure in education is profound, those failing high-stakes tests might face serious adverse consequences, including failure to move to their next phase of education, which might incur financial costs as well as detrimental consequences for their mental health (Beck et al. (2024). It might to be the case that the size of the testing-effect with retrieval is correlated directly with the difficulty of the task and therefore the level of mental effort that has been used (Rowland, 2014). In other words, the impact of retrieval on recall is more powerful when the difficulty of the retrieval task is greater.

If this is the case, then it therefore translates into a need to educate pupils to challenge themselves more during self-regulated study and to undertake activities during revision where they are presented with challenging revision tasks and tests of a relatively difficult standard (Kornell & Bjork, 2007). In the classroom, this translates into providing pupils with retrieval tasks that present some degree of challenge, but where there is not systemic failure to complete the task. It is this premise of successful, yet challenging, retrieval leading to more successful retrieval recall in a later time period (Pyc & Rawson, 2009). The Retrieval Effort Hypothesis suggests that as a pedagogical tool, retrieval practice will benefit memory for more challenging and difficult retrievals compared to retrieving information on easier topic areas (Klier & Buratto, 2023). All of this need balancing with the need to ensure education is accessible to all pupils.

Several papers examine how the use of effective feedback, following a more challenging retrieval, lead to heightened levels of recall in the next time period. This approach to spacing the retrieval with feedback, counters the damaging effect of getting an answer wrong and then holding onto this incorrect information (Storm et al. 2010; Leggett et al. 2018; Leggett et al. 2018). The same argument can be applied to the use of cues within retrieval attempts, allowing the partially failed retrieval to be more successfully attempted at the next testing period (Grimaldi & Karpicke, 2012). However, this is not universally accepted by all. The recalling of incorrect information at an earlier retrieval moment can lead to those wrong responses being learnt and recalled later on. Although, this argument failed to consider how feedback might limit future error (Storm et al. 2010).

4. Research Questions

The review of the literature provided me with a far greater understanding of retrieval practice. It has provided a lens for the intervention to take place. Discussions with the science teaching team have been based around many of the findings of previous papers. Over a period of three weeks in the early Spring Term we discussed some of the theoretical studies, both in departmental meetings and more informally. We discussed the use of clues and structured support during retrievals, feedback after a testing moment as well as the type of questions which might be asked. Colleagues were keen be able to set their own retrieval tasks rather than having them imposed upon them. However, they did also agree that the retrievals should be a mixture of questions relating to more recent topics and those topics from previous academic terms, too.

The critical analysis of the literature has highlighted a lack of research which investigates the pupils' perceptions of retrieval practice in lessons. I was keen to understand the perceptions of both the pupils and those of staff as part of a full evaluation of the intervention, particularly if the use of retrieval is to be encouraged across all academic departments within the school. I was also keen to engage the pupils in this intervention as the level of pupil voice in the school is very limited. There is an urgent need to develop the amount of pupil voice, both in terms of developing pedagogical practice, but also in preparing for inspection. A pupil council is being established this academic year and teachers are being encouraged to seek feedback from pupils at least once a term. This provided another reason, linked to the organisational priorities of the school, why pupil perceptions were a central part of this paper.

The literature review indicates that there is a benefit, in terms of enhanced attainment, in introducing retrieval practice into the curriculum. Although it would be useful to test this in this paper, I have purposely selected to focus on pupil and teachers' perceptions. First, the timing of the intervention in the late spring and summer terms is precisely the time when pupils are increasing their self-regulated revision at home and when colleagues are engaging in lunchtime inventions. Therefore, it would be challenging to differentiate the outcomes from this intervention to the impacts of these changes too. Secondly, the duration of the intervention of 6-8 weeks is limited and would be challenging to attribute enhanced outcomes to it. Finally, as previously mentioned, gaining pupils' perceptions is a key priority for the school and this intervention is a good method of promoting this to others on the teaching staff.

It is hoped that the intervention will allow for the development of pupil agency and being involved in "fruitful dialogues about teaching and learning" (Karlsen & Ohna, 2021, p.11). The engagement of pupils in their own learning also supports the development of positive pupil-teacher relationships, trust and a feeling of respect, which over time will lead to the enhancement in outcomes through enhanced learning and teaching and greater levels of pupil engagement and motivation (Marsh, 2012).

This paper therefore investigates the following research questions:

i. Do pupils perceive retrieval testing to be helpful in preparing them for their exams.
ii. Do pupils and staff perceive hints and cues to be supportive.
iii. How important do pupils and staff perceive feedback following retrieval to be.
iv. Do pupils perceive that they overly rely on re-reading during revision.

Throughout the study, these questions will be considered in a variety of ways. First, over time, comparing responses before and after the intervention. Secondly in terms of year group and gender, and finally in terms of the pupil's ability within the science curriculum.

I asked the final question (iv) to gain an initial understanding of how pupils revise at home. Although not a significant part of this study, I am keen to understand how well the school prepares the pupils for self-regulated study and whether there needs to be a review of the study skills sessions which are currently provided.

5. Methodology – including limitations and ethics

The intervention involved the introduction of retrieval practice into Science lessons over the period from March 2024 to May 2024. The retrieval intervention was introduced at the start of most lessons. In both Year 10 and Year 11 there are four lessons of one hour every week for each of the three sciences. Therefore, the pupils have 12 hours of science altogether. During the collaboration meetings, the staff felt that introducing retrieval at the beginning of each lesson would be counterproductive and repetitive but introducing it for 3 of the 4 lessons would be supportive. The teachers were asked to spend no more than 5 minutes on the retrieval testing each lesson which would replace the usual Do It Now activity. The staff were asked to provide a mixture of questions, some testing topics studied recently, and some testing topics studied last term or even last academic year. There were to be no more than 5 retrieval questions and ideally 3 questions at the start of each lesson. The format was devolved to the individual teacher. Before the intervention began in March 2024, the Head of Science discussed various retrieval practice approaches, some using digital technology, and then shared some resources with the staff in the department using TEAMS.

The intervention drew heavily on papers read in the literature review. Sana & Yan (2022) investigated the use of retrieval practice within science education using a sample of 155 secondary pupils in Canada. Here retrieval practice tests were issued weekly for four weeks covering topics recently studied and also from previous semesters. This idea of mixing questions in each retrieval test to include both recent topics and those from longer ago was taken from this study. Agarwal et al. (2012) undertook retrieval testing in a five-year applied study in an American school, and in one of their experiments they

introduced short, low stakes testing on a regular basis for one semester. This concept of short low stakes testing on a regular basis was a central aspect of my study. Carpenter et al. (2009) undertook a similar study of pupils taking History in High School. Although their methodology involved laboratory-based testing, the use of regular testing review and testing of different aspects of the course, supported the work of Agarwal et al. and Sana & Yan.

At the end of the intervention period, those pupils in Year 11 embarked on their study leave while those in Year 10 continued and prepared for their end of year examinations after May Half-Term. Staff were, of course, free to continue embracing aspects of retrieval practice with their classes after the end of the formal intervention period.

For the collection of the data, this paper adopted a mixed method methodology, embracing both quantitative and more qualitative data. Traditionally, these two branches of research have been separated with purists believing that their chosen method of research should not be mixed. The third paradigm of mixed method research is relatively more recent and has gained more credibility over time (Burke Johnson & Onwuegbuzie, 2004). Triangulation is presented as a benefit of the mixing of research methods allowing researchers to gain a deeper understanding of their more quantitative results (Mason, 2006). It provides a method to help with the study of complex concepts, allowing various sources of data to be explored. (Lakshmi, 2019). The discipline of mixed methods research continues to evolve (Crawford & Tan, 2019) with issues such as effective integration of both qualitative and quantitative data, methods of coding and the integrity of the data all being examined. A 2019 meta-analysis of 1055 teacher education articles between 2010-2016 found that 25.6% used mixed methods,

46.5% were qualitatively based and 27.9% used quantitative methods. “The trend over time suggests a decline in percentages of qualitative-only studies and an increase in mixed methods studies” (Crawford & Tan, 2019, p793).

My research involved a quantitative questionnaire survey conducted with Year 10 and Year 11 pupils taking Science at GCSE (n=63) and semi-structured interviews with both teachers of science (n=5) and pupils from each year group (n=6).

The questionnaire was distributed after the CUREC approval in March and the results analysed and then again in Early-May prior to study leave for Year 11 students. For clarification the same questionnaire was issued on both occasions. This therefore allowed for a time-series analysis to be conducted.

Pupils in Key Stage 4 were selected for this intervention by the staff in the science department. As previously mentioned, this was due to their belief that pupils in Year 10 and Year 11 might be able to provide more informed feedback than those in Key Stage 3. The use of one key stage only also was designed to reduce the workload on staff in introducing the intervention into their lessons. The results of each year group could be then compared. A further study might examine those pupils currently in Key Stage Three who have not considered GCSE options yet. Here it would be useful to explore whether significant differences emerged in their perception of retrieval within lessons.

5.1 Digital Questionnaire

The questionnaire was produced and distributed using Microsoft Forms. Settings were put in place to allow only one response from each pupil and to prevent pupils from other year groups as well as external respondents commenting. The form was distributed using an electronic homework application which only became accessible at

the start of the lesson in question. The teacher then directed the pupils to the application and asked that they might consider completing it. The estimated time for completing the activity was six minutes and it replaced the usual retrieval activity during the lesson. Across Year 10 and Year 11 there are 70 pupils, of which 30 are in Year 10 and 40 in Year 11. The overall cohort is split into 26 males and 44 females. Therefore, with 63 completed questionnaires, the response rate was 90% which was due to the pupils being able to opt out of the questionnaire in line with the CUREC guidelines. The second round of the questionnaire provided 64 completed questionnaires.

The questionnaire comprised of three sections, with 29 questions or responses. The first section outlined the survey's rationale and reinforced to the pupils that there was no compulsion in completing it. Based on a 90% response rate, despite full attendance, it can be concluded that this was clearly communicated. It also requested personal details from the pupil such as their year group at school, whether they were male, female or non-binary, whether they were taking the Triple Award or Combined Science and finally their initials. It was made clear to the cohort that their initials would be redacted from any write up on the results at a later stage, ensuring that the findings would be completely anonymous when submitted to the university for assessment.

The rationale for this first section of the questionnaire was to allow for statistical analysis to take place, looking at any significant comparisons in the data when sorted by year group, gender and academic ability in science. Further development and understanding of the statistical methods will be examined later in this section.

Section Two of the questionnaire was designed to examine the perceptions of the pupils surrounding their styles of revision; both at home and in school. It was designed to

investigate whether pupils prefer re-reading or testing themselves, whether they use applications such as Post It Notes or making summary lists and writing key points down. Two questions were related to how challenging revision is for the pupil, both in terms of organisation and how they cope generally with revision. These two questions (Question 8 and Question 14) can then be triangulated with the findings in section Three to examine whether those who find revision more challenging have more of a prevalence to finding retrieval helpful. The remaining questions in Section Two examine whether the pupils perceive more effort to lead to greater academic achievement.

The final section: Section Three, asked questions specifically related to the pupils' perceptions of retrieval practice in their Science lessons. A range of questions examined their perception of being tested on a regular basis, revision activities at the start of the lesson and whether they preferred to learn new material instead. Question 28 enquired about the effectiveness of the use of cues or clues and Question 29 asked about their preferences regarding being tested on more recent material.

The questionnaire drew on several papers which tackled learning habits and retrieval practice. Letterman and Morris (2013) researched pupil perceptions of homework activities; their questionnaire style was influential in my initial design. They used a 5-point Likert Scale and questioned the pupils on the time allocations devoted to homework and whether they felt it was supportive of their educational outcomes. Likewise, Hui et al. (2021) used a Likert scale when investigating why students do (or do not) choose retrieval practice. They attributed numbers to the choices and were therefore able to undertake some quantitative analysis explaining.

I was keen to understand how pupils revised at home and the extent to which self-testing and online testing formed part of their revision. The paper by Bartoszewki & Gurung (2015) provided an opportunity to review a variety of revision techniques and their links to examination outcomes. Here pupils used a 4-point Likert Scale to self-evaluate the perceived usefulness of differing styles of learning. Although, useful in helping to formulate section two of the questionnaire, the paper was limited by the disproportionate number of females in the sample as well as the large adult age range of 18 to 42 years old. Kuhbander & Emmerdinger (2019) and Hartwig & Dunlosky (2011) researched the use of retrieval practice in class, although both papers were in a university context. Indeed, a common limitation of many papers is that the context setting is in Higher Education and not secondary schools. However, these papers influenced the design of section three of the questionnaire with a focus on the perceptions of retrieval practice as well as a particular comparison with re-reading. This last point was an area that I was keen to explore even further during the pupil interviews.

Each of the questions addressed one of the four research questions in this paper. These are summarised in the figure 1 below.

Question number	Question	Research Question
6	I have devised a revision timetable for my exams	IV
7	I often re-read my notes during revision	IV
8	I find revising at home easy	IV
9	I highlight my notes during revision	IV
10	I use aids like Post It Notes	IV
11	I write the main points into a short summary	IV
12	I test myself by asking questions about the topic	IV
13	I make lists of key points	IV
14	Testing myself is a large part of my revision strategy	I and IV
15	I struggle with my organisation when it comes to revising on my own	I and IV

16	The harder you revise the better the outcome	IV
17	I make lists of key vocabulary and points	IV
18	I tend to revise the easier parts of the specification first	IV
19	Once I feel that I understand a concept I tend not to spend time revising it again	IV
21	I like my teacher testing us on previous work at the start of the lesson	I
22	The revision at the start of the lesson is a good use of time	I
23	I would prefer that the revision activities were set as homework rather than be at the start of the lesson	I
24	I prefer learning new things to revising in class	I
25	I would prefer that the teacher provides us with information about previous topics rather than testing us at the start of the lesson	I
26	Retrieval testing at the start of the lesson is a waste of time	I
27	I find being tested on topics we covered a while ago more beneficial than being tested on recent topics	I
28	I find it more helpful when we get feedback after the retrieval activity	III
29	I find it helpful when we are given cues (or clues) about the answers needed	II
30	I much prefer being tested on very recent topics	I
31	Undertaking retrieval activities at the start of each lesson is repetitive	I
32	I think that the testing of topics in class is preparing me for the examination	I

Figure 1: Allocation of survey questions

All of the questions in Section Two and Section Three were formatted in a Likert scale format. A five-point Likert Scale was selected with the choices being, left to right; strongly agree, agree, neutral, disagree and strongly disagree. The mid-scale choice was placed as neutral rather than “neither agree or disagree” In Likert’s original analysis, he used a mid-point, although since 1932, therefore have been studies where there is not one too. In their study, Nadler et al (2015) compared the use of 4-point and 5-point Likert scales and in particular the use of the words “Neither agree or disagree” Their finding suggest that the word “neither” attracts a disproportionate number of respondents compared to wording such as “no opinion” or “neutral”. It was on the basis

of this research that the wording “neutral” was used. There have been additional papers examining whether the 5-point Likert scale is indeed wide enough, such as Gunderman & Chan (2013) who examined a far wider 13-point Likert scale. However, although this potentially improves the accuracy of the results it does create confusion amongst respondents and elongates the time to complete the survey. Knapp (1990) elected for a 11-point scale; however, a consensus appears that a 5-point scale or more leads to more accurate results (Gaviao et al, 2023; Awang et al, 2016; Wu & Leung, 2017) which led to this paper adopting a 5-point scale for its analysis methodology.

There are several heavily debated limitations with the use of Likert Scales. Originally introduced in 1932 and used primarily in the social sciences, they are ordinal in their nature and are therefore not developed for parametric data (Carifio & Perla, 2008). Wu & Leung (2017) outline how applying numbers to the scale so that Strongly Agree is 5 through to the Strongly Disagree being 1, essentially counters the principles of an ordinal measure.

There are also psychological limitations to the use Likert Scales. These focus on three elements: central tendency, the culture of reading from the left and the unreliability of respondents to maintain consistency over a period of time.

Central Tendency relates to the preference by most respondents to score towards the middle of the scale and not to use the extremes (Douven, 2017). Likert Scales provide an example within research where this is common and therefore within a 5-point scale the incidence of the use of Strongly Agree and Strongly Disagree will be somewhat lower than the other three. Western culture also is based on a reading process where one reads from left to right. This creates a bias known as Pseudoneglect, where one pays

more attention to the items on the left of a scale than the right. In their paper testing university students' satisfaction rates, Nicholls et al (2006) test for directional bias within the Likert Scale. Their research found that the Likert Scale in descending order with Strongly Agree (5) on the left to Strongly Disagree (1) on the right, produced greater satisfaction levels amongst the cohort than when it was repeated using an ascending scale. Reasons for this are unclear, although potentially linked to a more dominant right hemisphere within the brain or the cultural norm within the western world of reading left to right.

Part of this current research is to examine the perceptions of retrieval by pupils over the course of the intervention. Yet, one limitation of repeat responses to Likert Scales is a lack of consistency and therefore reliability by respondents in general. In tests where the Likert Scale survey is immediately repeated, it has been found that, on average, only 57% of the cohort write the same response to each question within a 5-point Likert Scale. This drops further to 47% in a 7-Point Scale (Dolnicar, 2021). Although this paper is not testing for this, and therefore not compensating and accounting for it in the results, it is worthy of noting at this stage.

5.2 Pupil Interviews

Pupil interviews were conducted with pupils from Year 10 and Year 11, with a rationale to triangulate the results of the survey and gain an understanding as to reasons underpinning some of the results. Interviews were conducted in a semi-structured manner, with six central questions relating to both revision techniques and the perception of retrieval within lessons. The core questions are listed in the appendix to this paper. The notion of semi-structured interviews allows for the interviewer to

prepare a list of questions but allow for the organic flexibility of direction within the interview, linked to the responses by the participant (Brown & Danaher, 2017). Brown & Danaher's (2019) CHE model, based on Connectivity, Humanness and Empathy were adopted to ensure that the participant was comfortable and relaxed during the time that the interview was conducted. Ensuring an initial rapport was the participant was vital, ensuring that trust was established so that the pupils could comment freely and openly. The interviewees were all given an opportunity to withdraw from the interview, both before via an email and verbally at the start of the interview process. This was in line with the CUREC ethical agreement. The interviews took place in my office in comfortable seating, and they were all offered water and a snack. The use of first names placed them at ease and we talked briefly about something related to their context, such as their music practice or a recent sporting event. This related well to the Humanness principle, creating informality and removing barriers and distance between interviewee and interviewer.

Finally, there was a keen focus on Empathy, the third CHE principle, ensuring that rapport was developed throughout. It was important to provide my full attention to the interviewee and display emotional intelligence by actively listening to their responses.

Questions were developed using the guidance suggested by Jacob& Furgerson (2012). The first questions tended to be related to the interviewee's study habits and the use of terms such as "tell me about …" encouraged responses and encourages the interviewee to begin talking. The interview was also deliberately kept to 10 minutes, first to encourage pupils to agree to taking part as well as ensuring the opportunity cost in terms of time was minimised for them.

Interviews generally involve a power dynamic (Kvale, 2006) and although there was considerable effort to ensure that this was minimised through the use of the CHE principles, this cannot be completely countered. My position, as Headteacher, created an imbalance in power, even if in terms of pupil perception only, which compounded the inequalities relating to adult – child interviews as outlined by Kutrovatz (2017). Although hard to quantify, the responses may have been provided to meet with perceived approval by the school. Although, it was clearly stressed that their responses needed to be honest, this must be considered as a limitation.

The choice of questions (Figure 2) in the pupil interviews were carefully selected to gain a deeper understanding as to the reasons why pupils selected the variables in the Likert Scales. As with the questionnaire, the questions were influenced by previous studies (Hartwig & Dunlosky, 2011; Kuhbandner & Emmerdinger, 2019). The lack of research into pupil perceptions of retrieval practice presented a limiting variable but the style of questioning was taken from these.

1. How do you find revising at home for exams? (I)
2. What about testing yourself? (IV)
3. What about re-reading your notes. Do you do this in revision? (IV)
4. And what do you think about the retrieval exercises at the start of the lessons in Science? (I)
5. What are your thoughts on the frequency of the retrieval activities? (I)
6. Finally, what do you think about hints and cues? (II and III)

Figure 2: Core Pupil Interview Questions with research question in brackets

5.3 Teacher Interviews

Interviews were conducted with five teachers in the Science Department. The staff covered all three Sciences; namely Biology, Chemistry and Physics. Staff were asked whether they wished to undertake the interview, in line with the CUREC guidelines of

the university. I was particularly keen, being in the position of Headteacher, to check that they were genuinely happy to meet and provided them with the opportunity to opt out of the interview both at the start of it and in the lead up to it using an email from myself. However, it is important to acknowledge that the imbalance of power between Head and teacher might lead to a slight distortion in the data.

The interviews were preceded by significant collaboration within the department linked to retrieval. It had formed a focus within their meetings when discussing teaching and learning. As a group, they are perceived to be collaborative, with shared resources, and a culture of mutually supporting each other. This development of collaboration within the department mirrors the growing culture of collaboration within the United Kingdom educational sector (Vangrieken, 2015).

The interviews were based on a semi-structured model. The paper drew upon several previous papers to support in the forming of questions. The rationale was to understand perception of teachers as to the importance of retrieval and to also link that to teacher motivation and workload levels. With changes in school or departmental pedagogy, as introduced by this retrieval intervention, there is an implied expectation that teachers will adapt their way of teaching and this potentially increases the level of workload (Bakkenes, 2009). This research was keen to explore whether this increased workload was perceived as being educationally worthwhile.

Russo & Hopkins (2022) used online surveys to question primary teachers about the ability of Key Stage 2 pupils to retrieve mathematical information. Although linked to an earlier educational phase, it provided a useful starting point in developing the interview structure. A more general educational questionnaire was written by Morehead et al

(2015), gleaning the perceptions of teachers on pupils learning habits. Finally, Agarwal (2023) produced an informative personal review of Science Communication and Retrieval, outlining how information relating to retrieval practice might be disseminated to colleagues in schools and colleges.

Each of the teachers was given a number from #1 to #5 so that their responses could not be identified in any way during the research write up. The responses were then used to triangulate against both the interview responses of the pupils and the two survey results of

Year 10 and Year 11 pupils.

1. What have you been doing in the Department?
2. And do you undertake RP at the start of a lesson? (I)
3. So not just a quiz? (I)
4. Does it benefit the pupils? (I)
5. Are you saying that failing RP can be detrimental? (I)
6. What is your view on providing cues? (II)
7. What are your views on providing feedback (III)
8. And what about your view of the length of duration? (I)
9. What do you thing the pupils' perception of RP is? (I)
10. And is this common place? (I)
11. So, does RP lead to an increase in workload? (I)
12. And do your Do It Now's all have RP in them? (I)
13. Do you think it is therefore worthwhile? (I)

Figure 3: Core Interview Questions with teaching staff with research question in brackets

5.4 Ethical Considerations

Throughout this research I used the *Ethical Guidelines for Educational Research* which is part of the British Educational Research Association (2018). I completed the Research Integrity: Introductory Core Course and submitted the completion certificate to my supervisor. (see appendix 1) and began the CUREC process and submitted it to the

university after discussions with my supervisor and after having gained permission from the Chair of Governors to undertake the research in school.

I ensured that the opportunity and ability of both pupils and staff to withdraw from the intended research be made extremely transparent. Being the Headteacher at school, I was extremely conscious of the power dynamics at play. Written consent was sourced from all interviewees and an opt-out form was provided for all parents prior to the questionnaire being distributed to the pupils. In addition, the start of the questionnaire provided the pupils with an introduction to me as a researcher at the University of Oxford and a further opportunity for them to withdraw. It is noted that a small number of pupils exercised this right to withdraw at this stage, which suggests that the communication of their ethical rights was effective.

In summary, a considerable amount of thought was devoted to trying to compensate for my leadership position in the school. Figure 4 below provides an overview of these at each stage of the data collection.

Research Method	Ethical Consideration	How this research addressed this consideration
Explanation of the research to the Science Department	Staff might feel obliged to undertake the intervention as it might be deemed as a compulsory school teaching and learning policy. This is particularly due to my position as Head.	I met with the Head of Science, who is keen on pedagogy and active within the Chartered College of Teaching. We discussed the intervention and they were keen to embrace this too. We then spoke to the team at a departmental meeting. The Head of Science had prepped this up before and there was general approval for the intervention. At the meeting I made it absolutely clear that there was no compulsion to become

		involved in this research and they must not feel pressured at all. However, all agreed willingly.
Online Pupil Questionnaire	The pupils feel obliged to complete the survey due to my position at school. The is a concern regarding whether an individual pupil might be identified from the questionnaire at the write up stage.	As part of the CUREC process, I ensured that there was informed consent at all stages (BERA, 2018). Pupils were informed that although they provided their initials for the purpose of research analysis, these would be replaced with a number at the write up stage.
Pupil Interviews	Pupils feel obliged to take part in the interview process. There is a perceived dominance of staff, felt by the pupils taking part (Kvale, 2006., Kutrovatz, 2017).	The interviews took place in a very relaxed manner. Although power inequalities were considered during this research, there was a lot of focus on settling the pupil into the start of the interview by talking about something related to them at school or home, such as a goal that they scored in a sporting fixture or a recent family vacation. Child friendly methods were adopted throughout (Kutrovatz, 2017).
Staff Interviews	The same concerns regarding power inequality were considered (Kvale, 2006) as a valid limitation of this research. Colleagues were concerned that their identity was not available at the write up stage.	It was stressed to the member of teaching staff that they did not have to take part in the interview, both prior to the interview taking place and, again, at the start of it. I encouraged them to be honest and not to simply say what they thought I wanted them to say. Each member of staff was assigned a number from #1 to #5. This number rather than their name was used at the write up stage of the research.

Figure 4: A summary of the ethical considerations during the collection of data.

6. Results and Findings

6.1 Research Question 1: Do pupils perceive retrieval testing to be helpful in preparing them for their exams?

The final question on the student questionnaire asked the pupils to indicate the extent to which the retrieval testing had supported their preparation for the end of year or GCSE examinations. As with other questions, the results were collected in March and then again in May.

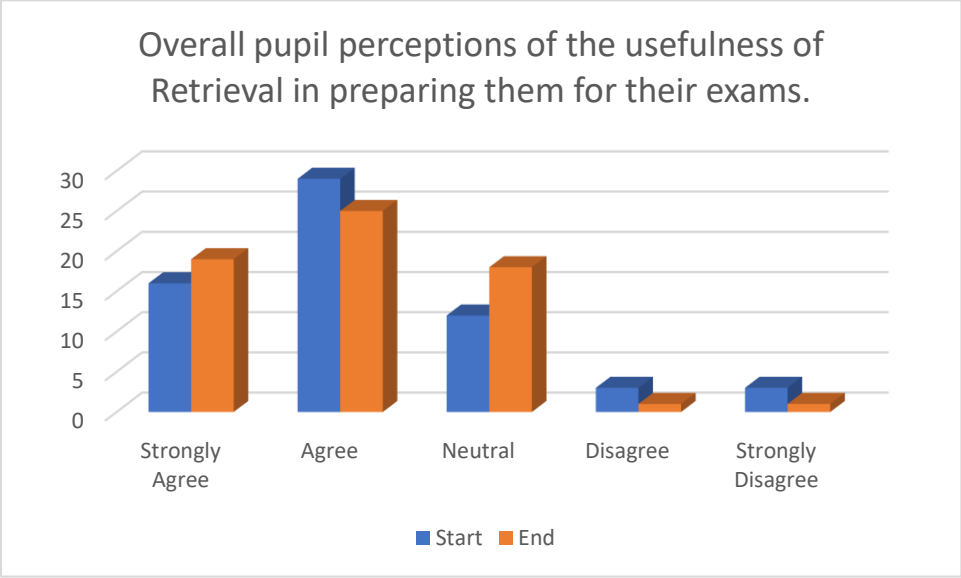

Figure 5

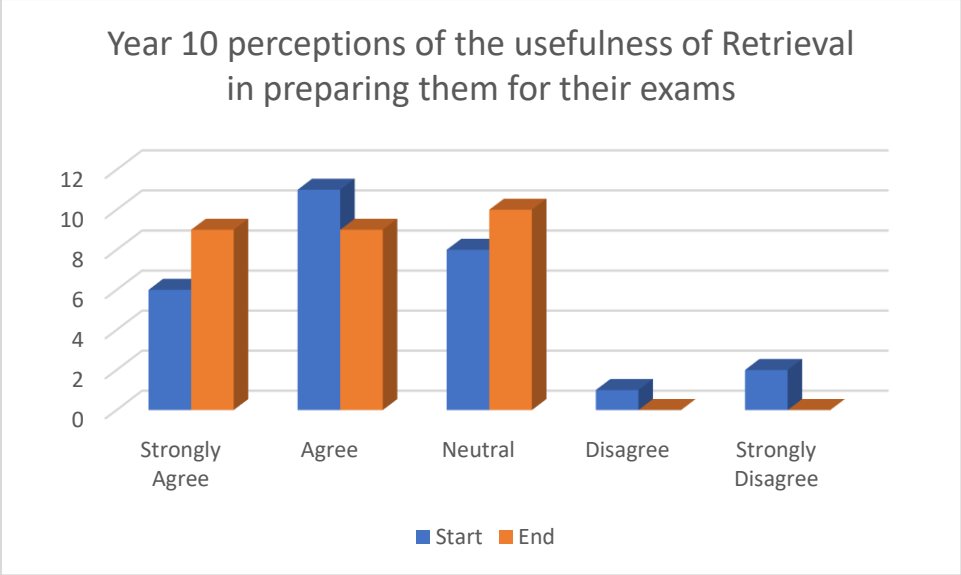

Figure 6

The results for those in Year 10 were dramatic. At the start of the retrieval intervention 4% of the sample disagreed that retrieval would be supportive in preparing them for their end of year school examinations and 7% strongly disagreed, whereas this had been completely removed 8 weeks later. Those strongly agreeing with the statement increased from 21% to 32%. The overall percentage either strongly agreeing or agreeing increased from 61% to 64% while the neutral percentage increased by 7 percentage points as those with previous negative feelings were amalgamated into it.

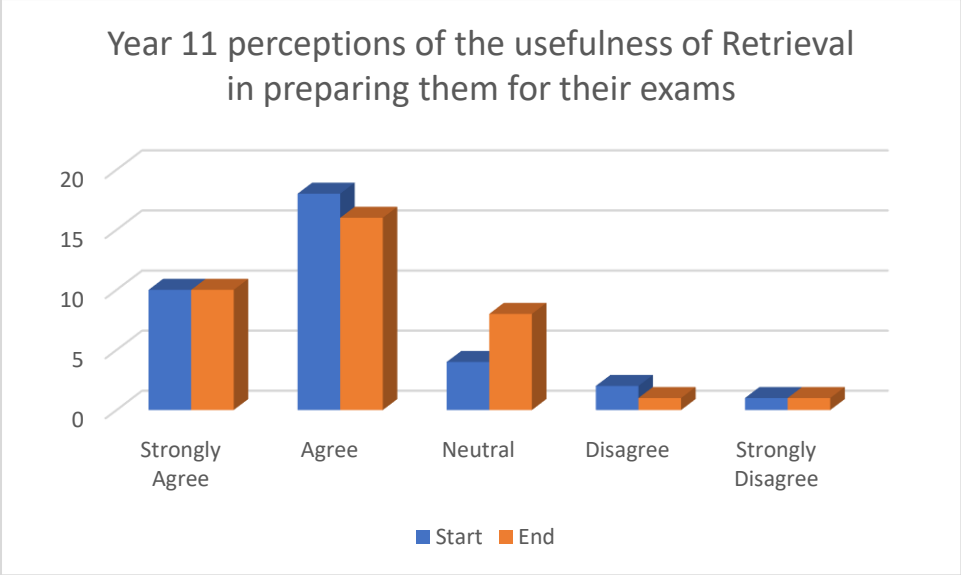

Figure 7

Year 11, on the other hand saw far more subdued and mixed results in terms of the pupil's perception that retrieval was preparing them for their GCSE examinations. In fact, the percentage that strongly agreed at the start of the intervention was 29% and it had fallen to 28% of the sample 8 weeks later on at the end of the intervention. The percentage that had agreed at the start was a significant reduction, falling from 51% to 44% following the intervention taking place. That said, the results did not show those strongly disagreeing increasing and the percentage disagreeing with the statement declined from 6% to 3% of the sample. It was those with "neutral" feelings towards the retrieval in class intervention that saw the largest increase from 11% to 22% overall. Of course, one needs to acknowledge the different sample sizes of 28 pupils responding in Year 10 and 36 pupils in Year 11, and the differences that one pupil can therefore make to the percentage changes, but nevertheless, the results for Year 11 provided some significant questions to explore during the pupil interviews.

Analysing this question in terms of gender provides interesting results too.

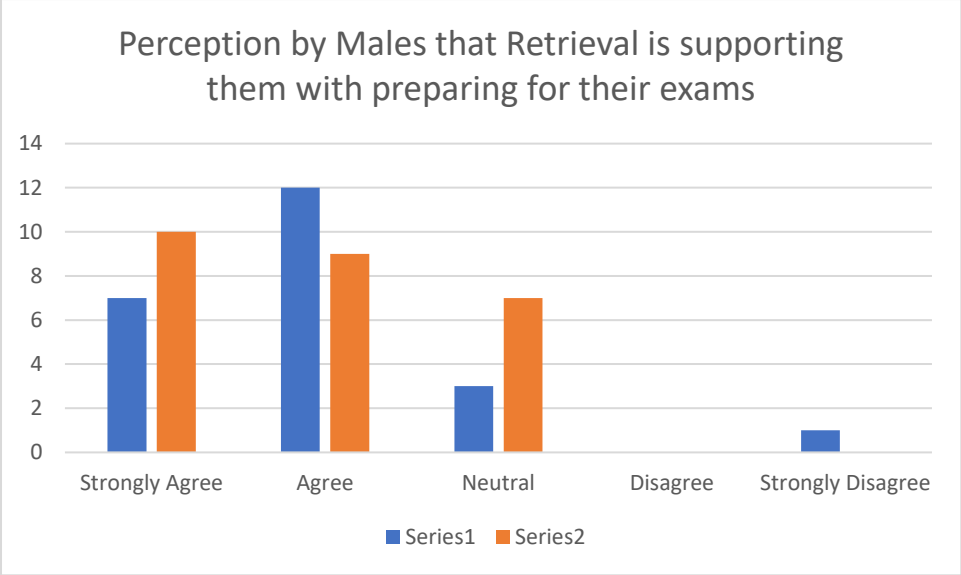

Figure 8

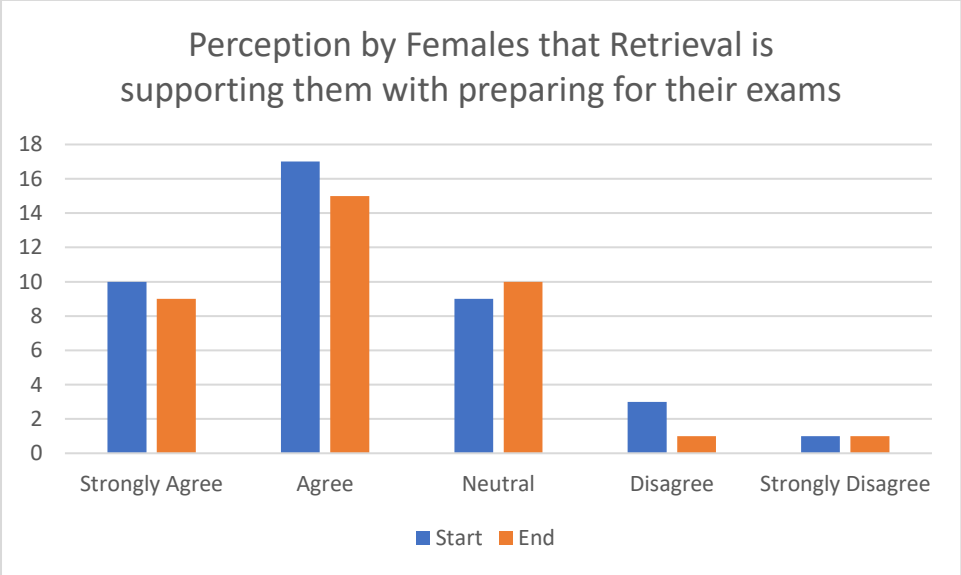

Figure 9

82% of the males in the sample believed that retrieval testing was going to prepare them well for their examinations, either GCSE or end of Year 10, and only 13% took a neutral stance. Whereas females recorded a positive agree or strongly agree of 68%, 23% of the them took a more neutral stance and 11% tended to display a negative response to the question.

After 8 weeks of the intervention occurring the responses to the question were collected again. The percentage of males agreeing or strongly agreeing with the statement had dropped to 73% with the remaining 27% taking a neutral stance. One positive in favour of retrieval was the reduction in those the very small percentage that had previously disagreed. Female pupils agreeing or disagreeing remained at a similar level, with 67% responding this way compared to 68% previously. There was, however, a slight reduction in those disagreeing with retrieval being helpful in preparing them for their examinations which saw the neutral stance response increase by 5 percentage points.

The results to this final, most holistic question, on their perception to retrieval do not indicate that the intervention was a great success, and it is therefore important to begin to understand this in more depth, both in terms of considering the qualitative responses as well as narrowing the remit and examining the factors that influenced it. During the interviews with the pupils, we tackled the overall effectiveness of retrieval. Some responses were hugely supportive of it, such as the comment made by Pupil #3 (Y10) who said "I really like it, it is really helpful as it makes me keep on thinking about a thing. I think that you remember more if you keep on going back and testing yourself on it". This positivity towards retrieval was also shared by Pupil #4 (Y10) who commented that "it is quite helpful, if you want to go a lot of time without testing then you forget it…. So, we study Chapter 3 and then forget Chapter 1. What the Do It Now's do is to make to remember things…. What it also does is to make tests in class a bit easier too". Yet, the support towards retrieval was not universal, which supports the questionnaire results and helps to provide some reasons for the quantitative results. Pupil #5 (Y11) was far more muted suggesting that "ummm… I think that it is useful, but I sometimes

switch off in class. Well, the teacher can be boring, and it goes on. They are not the best way to start every class, they need to be different, and they get really boring". Pupil #1 (Y11) also was more balanced too, indicating that "so, umm, perhaps…. Sometimes it is good and then I have some negative things to say too. It can often slow down the lesson"

These responses tend to suggest that the manner in which the retrieval is delivered in the classroom, along with the frequency of it, will have a large impact upon the way in which it is perceived by the pupils. The fidelity of the delivery is vitally important and there is a strong relationship between the quality of delivery and implementation and pupil outcomes. High quality implementation requires careful planning, to ensure that the low-stakes testing has validity, engages the pupils and encourages them to be motivated. The intervention used retrieval as the start to each lesson, and this needs more reflection. Was it the case that it was simply too frequent or perhaps the perception of limited delivery created a perceived feeling of disappointment?

One of the survey questions ask the pupils about their perception of its repetitiveness, with statement "Retrieval at the beginning of every lesson is repetitive" The overall responses, although not significant, tend to suggest that this is the perceived view and that over time that view of its repetitiveness continued to grow. The results tend to follow a normal bell-curve shape, which is common with Likert scaling and partially driven by central-tendency bias. The only slight anomaly are the two additional pupils who strongly disagree with the statement compared to the beginning of the intervention.

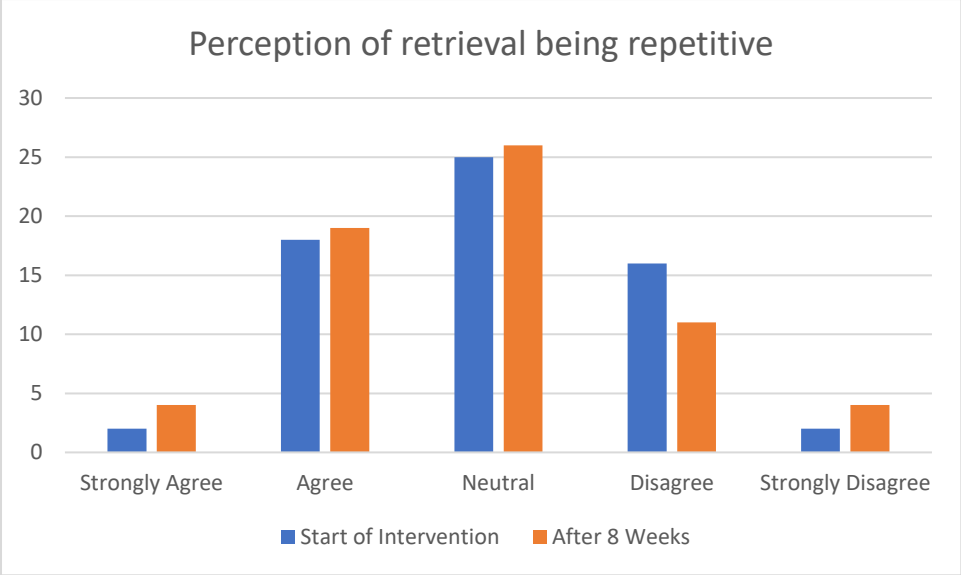

Figure 10

These responses in the pupil interviews linked to repetitiveness tend to indicate that it this perception is linked to the quality of the implementation. Pupil #1 comments were particularly telling “So, if they are the same each time then yes, some teachers need to do other things too” suggesting that a lack of variety, perhaps due to a lack of planning, is limiting their motivation. This was also echoed by Pupil #3 who commented that “some of them need mixing up a bit. Some are a bit, like, the same I think”. But they did acknowledge that repetition might be actually beneficial for retrieval and long-term memory development. In fact, the pupils all appeared to acknowledge that a certain level of repetition was beneficial in the longer-term, but they wanted to see low-stakes retrieval testing being carried out in a variety of ways too.

The research then triangulated these comments with the feedback from the teachers, gathered from the teacher interviews. What is immediately clear is that there is a mismatch of perceptions between the staff and the pupils. Each of the 5 teachers perceived that they devoted time to the retrieval activity, leading to a variety of

approaches being adopted. This polarisation of perceptions between staff and pupils was stark and is worthy of further evaluation in a future time period. When I asked Teacher #1 if their retrieval tasks were always a quiz they were rather resolute in their reply “No, not at all, otherwise it would get really samey. I use things like… a variety … blank mind maps, filling in diagrams…. It is important to have different types of retrieval tasks” However, conscious that staff may have been saying things they thought I wanted to hear, being the Head of the school, there might be some slight indication of the justification of the pupils’ perceptions in a few of the comments. For example, Teacher #5 was unsure at first in their reply “ummmm, I think so, I give them different types of tasks and tell them to have a go”, Teacher #3 potentially indicated some slight repetition in their reply "often, I have a question which says state three things” and Teacher #2 accepted that they “use mini-whiteboards for about 95% of all retrievals”

The triangulation might suggest that the misalignment of the perceptions is the result of more limited fidelity of implementation. If so, this would need addressing in the future should the trial intervention be continued.

Other final piece of analysis relating to this question was to investigate whether there was a difference in perception linked to the academic ability within science. Those pupils who are more academically able within science undertake the Triple Award and those who find Science more challenging undertake the Combined Award. This analysis is limited by the fact that a few of the pupils are advised to undertake the Combined Science Award but elect to do the Triple Award. The ethical constraints did not allow for this to be factored into the analysis as no internal school data was used at all.

For this analysis, a numerical value was placed against the responses, with Strongly Agree being 5 and Strongly Disagree being 1. Considering the second-round data, taken after 8 weeks from the start of the intervention, we can see that from the 64 responses there were 43 pupils taking the Triple Award and 21 undertaking the Combined Award at GCSE in Key Stage 4. I undertook a statistical testing using SPSS, licenced to the University of Oxford, (Figure 11). It can be seen that the means are similar and that using a T Test, there appears no significance that either Combined Science or Triple Science students prefer retrieval, or that retrieval is significantly perceived as being helpful by either group.

Group Statistics

	Assessmenttype	N	Mean	Std. Deviation	Std. Error Mean
Preparedscore	Combined	21	3.9048	1.09109	.23810
	Triple	43	3.9535	.78539	.11977

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means							
						Significance				95% Confidence Interval of the Difference	
		F	Sig.	t	df	One-Sided p	Two-Sided p	Mean Difference	Std. Error Difference	Lower	Upper
Preparedscore	Equal variances assumed	4.162	.046	-.204	62	.419	.839	-.04873	.23840	-.52528	.42782
	Equal variances not assumed			-.183	30.473	.428	.856	-.04873	.26652	-.59268	.49523

Figure 11

Several questions in the survey were asking similar questions, to try and gain an understanding for this perception. For example, “retrieval testing at the start of the lesson is a waste of time” and “the revision at the start of the lesson is a good use of time” can be both used to correlate against the final question relating to retrieval practice being effective in preparing the pupils for their exams.

	Revision at the start of the lesson is a waste of time	Revision at the start of the lesson is good	I think testing is preparing me well for my exams

Strongly Agree	3	15	19
Agree	9	17	25
Neutral	24	21	18
Disagree	15	10	1
Strongly Disagree	13	1	1
Mean	2.593	3.547	3.938
Standard Deviation	1.109	1.068	0.889

Figure 12

Figure 12, is of some interest, the general correlation between the perception of retrieval revision at the start of the lesson being a “good use of time” is in line with the perception that retrieval is preparing the pupil for their examination. Yet, there is a more marked difference when asking “is retrieval at the start of the lesson a waste of time?” Here, the mean of 2.593 places the average score into “Disagree”. This would use useful to investigate in a much larger study.

In order to continue investigating the importance of the fidelity of implementation, a question was asked whether the retrieval activity should be set for homework instead of taking up lesson time.

This question, as with others, was asked at the start of the intervention and towards the end of it. In the second survey round (n=64) the mean was 2.73 compared to 2.42, meaning that there was a slight increase in the tendency towards preferring the retrieval to be set as a homework task. That said, the latter mean of 2.73 is still disagreeing overall with the question. Analysing this in SPSS, the P value of 0.072 highlights there is no statistically significant change. (Figure 13)

Group Statistics

	round	N	Mean	Std. Deviation	Std. Error Mean
HomeworkScore	One	63	2.4286	1.10299	.13896
	Two	64	2.7344	1.23754	.15469

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means							
						Significance				95% Confidence Interval of the Difference	
		F	Sig.	t	df	One-Sided p	Two-Sided p	Mean Difference	Std. Error Difference	Lower	Upper
HomeworkScore	Equal variances assumed	1.488	.225	-1.469	125	.072	.144	-.30580	.20813	-.71772	.10612
	Equal variances not assumed			-1.471	123.791	.072	.144	-.30580	.20794	-.71739	.10578

Figure 13

Summary

The first section of results has tackled the first research question: Do pupils perceive retrieval testing to be helpful in preparing them for their exams? The literature review indicated that there is a clear link between the introduction of retrieval and enhanced outcomes. Although this paper examined the pupils and teachers' perceptions of this, the results suggest some perceived benefits, although much depends upon the quality of the intervention and therefore the subsequent level of pupil engagement. We will now move onto address the second and third research questions: Do pupils and staff perceive hints and cues to be supportive? and How important do pupils and staff perceive feedback following retrieval to be?

6.2 Research Question 2: Do pupils and staff perceive hints and cues to be supportive? Research Question 3: How important do pupils and staff perceive feedback following retrieval to be?

6.2.1 Feedback

Rosenshine subscribes to the use of effective feedback to the learner once a piece of work has been completed. The DfE NPQH programme promotes this in its course literature for aspiring Headteachers and OFSTED as well as ISI, in the independent sector, consider it to be an important aspect of pedagogy. The stark results of the survey suggest that having completed the intervention for a number of weeks, the majority of the pupils acknowledged the importance of feedback following the retrieval task. Only 2 pupils from the sample of 64 disagreed with it being “more helpful” and 42 of the 64 agreed or strongly agreed. Considering these results, especially when Central Tendency Basis is factored in, it would appear that the general feeling amongst the pupil cohort is that feedback from teachers is supportive to learning.

Although the differences before and after the intervention are not significant, the reduction of those strongly disagreeing with the benefits of feedback, from 3 to absolute zero, is worthy of note and support the answering of the second research question.

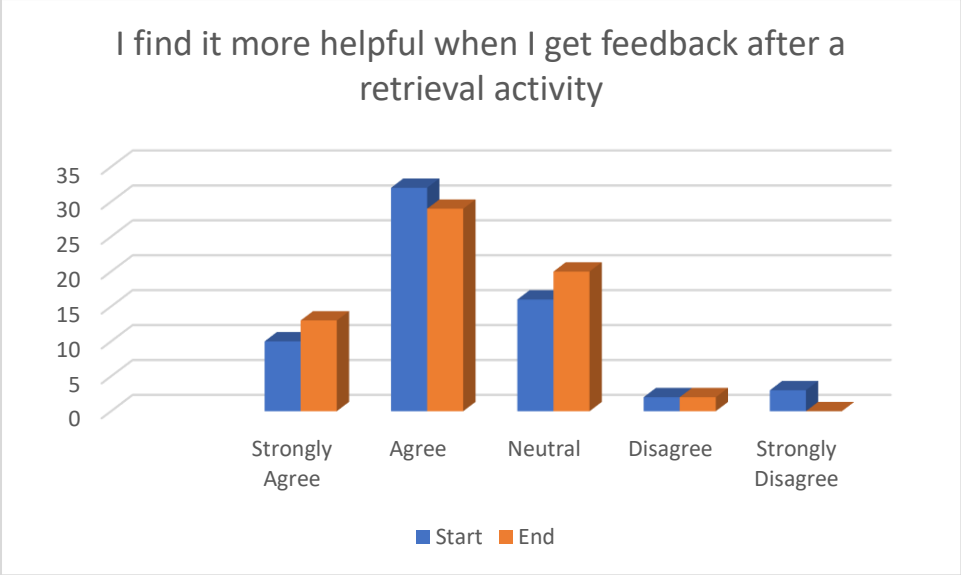

Figure 14

The results from this question tend to be in line with the general findings that feedback is supportive for retrieval and recall (Leggett, 2018; Roediger & Butler, 2010; Kang et al, 2007; Storm et al, 2010). During the teacher interviews, Teacher #1 discussed how retrieval is beneficial to the pupils, increasing their confidence and allowing them to continuously review their work. However, they also indicated that as there was a "broad spectrum of ability in the classroom, it is important that feedback is targeted to allow all pupils to make progress". Here the feedback is being used as part of an adaptive teaching process, which will lead to potential progress across all key groups within the class.

6.2.2 Cues

As part of the intervention, the teachers provided pupils with cues or hints at times during the retrieval exercises. Pupil perception of the usefulness of cues sees more mixed results than the perception relating to feedback, with the mean score being 3.578 compared to 3.828 (see table xxx). This was further analysed by academic ability within

science, with interesting results that cues are less well received by those taking the Triple Award compared to the Combined GCSE Assessment. Here the difference in the mean scores for both are more marked with 3.441 for Triple and 3.857 for the Combined, with standard deviations of 0.1007 and 0.857 accordingly. This low deviation, essentially showing a lack of spread in the data, points to an overall average "neutral" position, which tends to suggest that there is no evidence of significant helpfulness or unhelpfulness of hints, although further testing of this significance would be supportive. A large limiting factor is the relatively small sample size of 64 pupils, although this reflects the size of the school and the genuine ability for pupils to deselect themselves from the research. However, that said, there are some comments that could be carefully made; first there may appear to be the initial glimpses that those pupils in the more able class have a more subdued response to the provision of cues and hints, and certainly more subdued than their perception of feedback being positive. The definitive zero that disagree or strongly disagree with feedback is greater in the combined group, although not significantly at all. This would need to be investigated further, expanding the sample size, perhaps into Key Stage 3, or setting up a collaborative sample group in another school or a selection. I have begun discussing this latter opportunity with a fellow colleague on the MLT course for next academic year. Another point of note relates to the mean for the Triple group being 0.443 higher for feedback compared to cues, whereas the Combined group saw a mean 0.143 higher for cues compared to feedback.

	I find Cues helpful	I find feedback helpful	Feedback helpful (Triple) (43)	Cues helpful (Triple) (43)	Feedback helpful (Combined) (21)	Cues helpful (Combined) (21)
Strongly Agree	10	13	8	5	5	5
Agree	26	29	22	17	7	9
Neutral	22	20	13	16	7	6
Disagree	3	2	0	2	2	1
Strongly Disagree	3	0	0	3	0	0
Mean Score	3.578	3.828	3.884	3.441	3.714	3.857
Standard Deviation	0.973	0.788	0.697	1.007	0.956	0.854

Figure 15

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means							
						Significance				95% Confidence Interval of the Difference	
		F	Sig.	t	df	One-Sided p	Two-Sided p	Mean Difference	Std. Error Difference	Lower	Upper
FeedbackScore	Equal variances assumed	4.886	.031	.806	62	.212	.424	.16944	.21034	-.25104	.58991
	Equal variances not assumed			.724	30.749	.237	.475	.16944	.23419	-.30835	.64722
CuesScore	Equal variances assumed	.818	.369	-1.624	62	.055	.109	-.41528	.25566	-.92633	.09577
	Equal variances not assumed			-1.720	46.261	.046	.092	-.41528	.24142	-.90117	.07060

Figure 16

Further statistical testing, using SPSS, considers the significance with the comparison of perceived helpfulness of cues between those within the Combined Science class and those in the Single Science Trilogy class. Assuming equal variances, we note a one-sided p insignificance of 0.055; with only results below 0.05 indicating statistical significance.

However, as some pupils with true lower science ability are placed into the trilogy award due to parental expectation, it limits the scope of this investigation to some extent. An internal test, using school baseline data would be more supportive.

It is important to devote more time in the future to understanding the rationale for these results and the potential to undertake an inter-school research project would be highly supportive here and should be explored further.

This was further investigated in both the pupil and staff interviews in some depth. Here there were a range of comments made, such as Pupil #1 who felt that hints were “good, they stimulate your brain”, they also alluded to the fact that the use of clues reduce mental health concerns and potentially prevent pupils from worrying too much in class. This positive view of hints was also shared by Pupil #4 who felt that “they unlock the brain when you are having a blank” and also can support, therefore, the process of long-term retrieval. Both of these pupils were undertaking the combined science specification, both finding science challenging at times. On the other hand, Pupil #4 commented “if you get stuck on a question then it is helpful. But, often I am working on it and then I get stopped by the teacher for a hint to the class and it is disruptive and counterproductive” They went onto say that the hint can also reduce the focus in the classroom, leading to increases in noise levels too. Pupil #2, a pupil taking Triple Award, was more negative about hints during retrieval, believing that they were not helpful and reduced pupil independence, stressing that by making it easier and more accessible, it may be limiting the level of examination preparation.

The perception of staff would appear to triangulate with these views too. Teacher #2 believes that cues are helpful, particularly with lower ability children and that by

providing them it “allows them to believe that they can do the retrieval practice”. Yet, there is also a reluctance to provide too much help too. Teacher #4 was reticent in their response “ummm, I think it is good, but I let them have a go and will only guide them if they are struggling”, a point echoed by Teacher #3 who would rather they have a go and then be supported if they cannot access the material. Interestingly, comparing the comments of Teacher #5 and Teacher #1 one can draw some partial conclusions relating to the fidelity of implementation. On one hand Teacher #5 talks of being happy to provide cues when asked but also simply asks the pupils to reread their notes. Whereas Teacher #1 felt that strongly that retrieval should provide scaffolding to ensure that teaching is adapted for each pupil. This was elaborated on by discussing the need to have different retrieval activities with varying levels of challenge in them.

Finally, I tested to see if there was any significance between the scores for males and females. Here, I considered those at the end of the intervention who identified as male (n=24) and female (n=36). Four pupils who did not identified as male or female were discounted. Interestingly, only one pupil did not identify as male or female during the first round of data at the start of the intervention. Using SPSS produced an overall mean of 3.60 for the overall male and female cohort. This equates to 3.625 for the male mean and

3.583 for the female mean. Further testing for statistical significance confirms no significance with one-sided p results less than 0.05. (see Figure 17)
Therefore, it might appear that the level of academic ability has slightly more impact on the perception of helpfulness of hints and cues, but gender does not.

Group Statistics

	Gender	N	Mean	Std. Deviation	Std. Error Mean
CueScore	Male	24	3.6250	.87539	.17869
	Female	36	3.5833	.96732	.16122

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means							
						Significance				95% Confidence Interval of the Difference	
		F	Sig.	t	df	One-Sided p	Two-Sided p	Mean Difference	Std. Error Difference	Lower	Upper
CueScore	Equal variances assumed	.610	.438	.170	58	.433	.866	.04167	.24559	-.44994	.53327
	Equal variances not assumed			.173	52.727	.432	.863	.04167	.24067	-.44111	.52445

Figure 17

Summary

Although no results suggested any significance there are elements within these results that are strongly supportive of feedback, across the key groups of both gender and ability. The definitive zero response of those who strongly disagree with the benefits of feedback following the intervention is particularly supportive in answering research question three.

The case of the implementation of cues is less strong, with those pupils who have slightly higher academic ability finding them, at times, disruptive. Yet again, a large factor in determining these results is the quality of the delivery of the intervention. Where it is well conceived and delivered by the teacher, it leads to more positive perceptions. Cues provided as part of adaptive teaching are far more beneficial than those delivered universally to the class cohort. Again, these results provide a rationale for the delivery of well-designed and targeted professional development linked to pedagogy. We now move to the fourth and final research question within the paper.

6.3 Research Question IV: Testing the premise that pupils do not test at home and rely on re-reading

It was the general premise within the literature review that pupils tend to neglect the concept of self-testing within their revision strategies and prefer to focus on re-reading their notes. This, therefore, has the implication of making retrieval practice within lessons more important.

The survey questionnaire investigated the prominence of this assumption. First, a statement "Testing myself is a large part of my revising strategy" was asked, which appear to counter the findings from previous research. Although, there is some disparity between round 1 and round 2 figures, the general tendency is similar, with approximately 60% of the sample cohort agreeing with the statement.

	Round 1 (%) (beginning of intervention)	Round 2 (%) (end of intervention)
Strongly Agree	30.16	29.69
Agree	28.57	32.81
Neutral	20.63	21.88
Disagree	19.05	14.06
Strongly Disagree	1.59	1.56

Figure 18: Testing is a large part of my revision strategy

This was triangulated with the statement "I test myself by asking myself questions about the topic". Figure 19 highlighting the results of this, once again suggest that testing plays an important part of a pupil's revision strategy, although it is noted that the changes from round 1 to round 2 are more marked in terms of disagreeing with the statement. This might be due to the use of testing each lesson which develops the perception of it not being so needed in self-regulated study.

	Round 1 (%) (beginning of intervention)	Round 2 (%) (end of intervention)
Strongly Agree	20.63	34.38
Agree	42.86	28.13
Neutral	22.22	15.63
Disagree	12.70	18.75
Strongly Disagree	1.59	3.13

Figure 19: I test myself by asking myself questions about the topic.

	Round 1 (%) (beginning of intervention)	Round 2 (%) (end of intervention)
Strongly Agree	14.29	26.56
Agree	28.57	37.50
Neutral	20.63	20.31
Disagree	28.57	6.25
Strongly Disagree	7.94	9.38

Figure 20: I often re-read my notes during revision.

The third statement in this triangulation was “I often re-read my notes during revision” (see Figure 20) This presents conflicting results, which potentially may be an unintended consequence of the intervention. In March 2024, 43% of the sample indicated that they re-read their notes as part of their revision strategy, yet this had increased to 64% of the sample in May 2024 at the end of the intervention. This potential conflict in the data was investigated further in using qualitative data from the pupil interviews. Pupils talked of how they use testing in their self-regulated revision, with comments such as “exams stress me loads and therefore I need to do exam style questions in my revision at home” (Pupil #1). In response to a question asking about the use of testing at home “well we have to do it, don’t we? Although it is one of my main forms of revision, testing is not the best way for me” (Pupil #4) This pupil then stated “I like rereading my notes, it is really helpful. If I get stuck I can go through them and jog my memory, it helps a lot”

Here we see that the testing of material is not recognised as overly supportive, but is being completed anyway, whereas others were honest about not using testing, such as Pupil #2 who stated “to be honest, I don’t do much” and “I like to reread stuff because it then makes me get it” and “I reread my books the night before an exam”. Overall, there appears to be a disconnect between the comments from the interviews and the results of both the Round 1 and Round 2 surveys, with the interviews supporting some of the results from previous studies. Perhaps one informed comment came from Pupil #3 who said, when asked about re-reading their notes during revision “I do this quite a lot. It is really helpful to me, but perhaps not so helpful as testing myself, ummm, (pause) but I reread the most, maybe rereading is not the best though” This interesting self-reflection is much in line with previous research in this area (Rowley & McCrudden, 2019; Callender & McDaniel, 2009; Weissgerber & Rummer, 2023).

Summary

The results from the survey tended to support the evidence that pupils rely heavily on re-reading of notes and textbooks during self-supported study, yet, at the same time the results also highlighted a larger weighting of pupils who also used self-testing in their self-supported study.

In the second round of the survey, following the retrieval intervention, the percentage of, pupils engaging in self-testing marginally fell. On one hand, this might be seen as a disappointing outcome of the intervention, but it could also be seen as positive as the lower order assessments skills are being regularly covered during retrieval testing in class and therefore the pupils are then focusing on the more higher order skills of

analysis and evaluation during self-supported revision. Further testing of this assumption is required.

7. Discussion and Limitations

7.1 Discussion

The aims of this paper were to consider the widely reported benefits of retrieval and determine whether the curriculum-wide implementation of retrieval might enhance learning and teaching at school. The literature review also considered the importance of self-regulated study and the potential over-reliance by pupils for re-reading rather than for self-testing.

In terms of this latter point, the findings of the study broadly support the reliance on re-reading and highlight how pupils find this a secure way of revising. Although there is some clear benefit in this approach to revision (Rawson et al., 2000), it is somewhat limited, particularly if the re-reading takes place immediately after the initial study (Callender & McDaniel, 2009). However, others are more subdued in terms of their appraisal of re-reading, such as Bjork et al. (2013) who state that it is “not an efficient way to learn” (Bjork et al., 2009, p420).

The results of the questionnaire, which was then triangulated with the pupil and staff interviews, provided a mixed picture here. Pupils acknowledged that they recognised that testing was beneficial to their memory retention, a result in line with Weissgerber & Rummer (2023), but also were keen to support their revision with re-reading of notes and books. The survey suggested that this cohort sample, both across Year 10 and Year 11, tested themselves to a greater level than some of the previous studies suggested, but the interviews highlighted that this did not sit naturally with them.

In terms of the results relating to the other research questions, the outcomes, although not significant, did highlight some nuances which other papers had not reported. This is due to the focus on pupil and staff perceptions rather than academic outcomes. The overwhelming agreement across a number of papers is that feedback is beneficial to future retrieval outcomes (Carpenter, 2023; Grimaldi & Karpicke, 2012; Hostetter et al., 2017; Kornell et al., 2011; Leggett et al, 2018; Roediger III & Butler, 2011; Rodrigues de Lima, 2020 & Roediger III & Karpicke, 2006) This general positive opinion is only diminished by Agarwal et al. (2016) who, to their surprise, reported that feedback had little impact on attainment during the course of their experiments. A more moderate view comes from Smith et al. (2016). Their paper concluded that feedback does not always help, particularly if it is not understood. Once again, although my results are not significant, there are elements of the results that concur with Smith et al. (2016). The higher ability groups within both Year 10 and Year 11 found feedback more helpful, potentially indicating that those will lower scientific ability found the feedback less accessible. It would have been more helpful to examine this further during the course of the interviews.

The research related to the use of cues also presented mixed, although non-significant results. Previous research results tend to suggest that there is a benefit to future attainment if cues are used (Karpicke & Blunt, 2011; Klier & Buratto, 2022; Rodrigues de Lima, 2020; & Vaughn et al., 2017). Although this paper found that the overall perception by pupils and teachers agrees with this, it is once again worth noting the subtle nuance in the data which relates to scientific academic ability. The incidence of agreement that cues are beneficial to retrieval is greater with those pupils in the combined science classes, which tend to have the lower academic ability pupils in them.

Those pupils in the Triple Science classes found them less helpful, referring to the prevention of creative and independent thinking as well as a distraction in the lesson when they are used by the teachers.

This correlation between feedback and scientific ability and the use of cues and academic ability is worthy of note and needs to be examined further as retrieval practice is adopted across other areas of the curriculum. It also supports the need for careful CPD training and the refinement of adaptive teaching for teachers.

Overall, the results from this paper are in line with my reading of the former literature on retrieval. Although this research focuses on perception, the findings suggest that both pupils and teachers find retrieval supportive of their preparations for examinations, which is in line with the general consensus that it supports memory retention and academic outcomes.

7.2 Limitations

Throughout the study, I have been extremely conscious that my position as the Head of the school might influence the outcome. I have been relentless in ensuring that all stakeholders understood that they were not being judged in any form and that their responses should be honest and open. Although I am relatively satisfied that the data is honest and truthful, I also acknowledge that both pupils and colleagues might have been careful in what they said.

The relatively small sample size was always going to be somewhat of a limitation, but that was governed by the small size of the school. A larger data set might have produced some results displaying greater significance, but there is enough data to make some general conclusions.

I have already acknowledged the limitations of using Likert Scales including the difficulty of applying a numeric scale to them. Although this is once again acknowledged at a limitation of the study it was not factored into the findings when applying statistical tools to them. Central tendency bias, as well as Pseudoneglect are also limitations in terms of potentially skewing the data, and although they are acknowledged they have not been factored into the results and findings. I might also have considered a wider scale, potentially seven points, to determine where there might have been more significance in the results. Finally, a future study might address whether the academic outcomes improved due to the intervention. This would be fraught with potential problems in terms of not being able to work within a ceteris paribus model. Schools are alive with interventions, revision sessions, study skills workshops as well as alterations to the curriculum. It would therefore be challenging to apportion any improvement linked to retrieval without artificially limiting other changes and innovations.

8. Organisational implications, collaboration and personal reflections on my practice and leadership style

The concept of teacher collaboration is seen as a group of colleagues, either in a department, a school or across a group of schools, who come together to work on a shared project or task (de Jong, 2022). It has become increasingly popular in the education sector; indeed, many senior leaders are placing pressure on their staff to collaborate with each other as well as engage in inter-school collaborations too (Vangrieken et al. 2015). Although collaboration can lead to many positive outcomes, enhancing the delivery of learning and teaching, development pastoral care and potentially leading to a decline in individual workloads through the sharing of good practice and resources, it can also be damaging too. If the collaboration is forced upon colleagues, restricting individual creativity by creating a uniformity in teaching styles and delivery in the classroom, it can be seen as increasing workload and lowering teacher motivations (Vangrieken et al, 2015).

The development of collaborative teamwork ought to not only reduce individual workloads but also diminish the sense of isolation for teachers. Yet, it can also create competitiveness within the collaboration team, which in the short term may increase standards of teaching but can lead to resentment and decline in morale in the longer term. That said, teachers are “the most important agents in shaping education for students and in bringing change and innovation in educational practices” (Bakkenes et al. 2010, p.533). It is therefore important that effective collaboration occurs to ensure that these innovations and pedagogical improvements are shared across the profession.

This research, a small collaborative investigation with a group of science teachers, has led to many of these positive aspects. It was not forced, although, as the Head, it is difficult to be completely confident that the staff did not unconsciously simply agree, and the team are keen to continue adopting the intervention even though the timeframe for it has ended. As with the collaboration conducted by Cowan (Cowan & Farrell, 2023), the collaboration in this research project involved both experienced and newly qualified practitioners. As with the Cowan paper, the sharing of good practice and willingness to learn from each other has both enhanced teaching and learning and further fostered a sense of teamwork within the department.

I was particularly keen to avoid the criticisms made of school leaders by Bakkenes et al. where teachers are simply expected to adapt their classroom practice to accommodate the latest Learning and Teaching policy being implemented by the Senior Leadership Team. Although it is challenging to prove, and therefore a limitation in this study, I believe that this was achieved. There is now a willingness to adapt and extend this research project, into different phases within the science curriculum, and through exploring other aspects of Rosenshine's Principles.

During the course of the Summer Term the discussions relating to retrieval practice have widened to involve teaching colleagues outside of the department. The Head of Science spoke at the Heads of Department meeting in May about the intervention, explaining to others how they implemented it. The appetite to discuss teaching and learning and aspects of pedagogy has grown. In my own teaching, I have experimented with different types of retrieval, conscious of the initial results relating to students' perception of its implementation. An embryonic teaching and learning group has been

established, which, being the Head, I did not feel that I should join. It meets on a regular basis and is open to all staff. Writing this at the end of the summer term, its members have agreed to keep meeting with the view of sharing good practice with aspects of pedagogy, including retrieval. This is very much a grassroots led group, with members of the Science team being part of it. As part of this, two books have been purchased for all Heads of Department. Both from the same publisher, the first being Tom Sherrington's "Rosenshine's Principles in Action (2019) and the second Claire Grimes' "Rosenshine's Principles in Action – The Workbook" (2020).

Heads of Department agreed at their last meeting in early July that they would place teaching and learning on their agendas moving forwards. What this research has not done is to implement retrieval across the curriculum, but it has encouraged, through teacher collaboration, more discussion of it and other aspects of pedagogy. It is clear from the results of my research that the fidelity of retrieval's implementation plays a significant part in its success. Therefore, the forced implementation, through school-wide policy would, according to the results, be counter-productive. This is a different outcome to the one that I had perhaps intended at the start of the process, when I had envisaged retrieval to be adopted across the curriculum. However, the process of conducting this research has shown two things to be very clear; teachers need to buy into policy and agree to it, indeed, own it. Secondly, the need to share good practice and engage in effective continued professional development to allow this to happen is vital. This has led to one significant change in the next academic year, the implementation of one additional in-service training day completed developed to teaching and learning and the sharing of good practice. It is very much hoped that this culture of discussing pedagogy will continue.

In terms of my own practice, next academic year I am teaching groups of pupils taking A Level Economics. My intention is to experiment with aspects of pedagogy, including retrieval, challenge, feedback and aspects of gamification (my first-year project), ensuring that at all times I gather pupil feedback as to their perceptions. I intend to experiment and introduce different types of retrieval into my lessons and wish to explore retrieval opportunities through verbalisation as well as testing. Examples of this might include encouraging pupils to teach each other as well as self-explanation to the class (Desy et al., 2017). The literature also has encouraged me to experiment with introducing some higher order skill testing into my own retrieval. Fact based retrieval tends to focus on the lower order assessment skills on Bloom's taxonomy and designing quizzes which incorporate analysis and evaluation will benefit the pupils' overall understanding (Agarwal, 2019).

Keen to collaborate, I am also intending to share the outcomes of this research with a fellow student on the MLT course, who is also a Headteacher. We have currently agreed to expand each other's research into our schools to gain a greater sample. The project that they are working on is the effective use of feedback to pupils which is another of Rosenshein's Principles of Instruction. My reading of the literature has increased my understanding of pedagogy and has directly led to not simply adopting widely discussed ideas without testing them within the school first of all. It has also deepened my belief that pupils should, at some level, be part of their own educational journey. I therefore plan to survey pupils at the end of each term in my own classes as to what pedagogical approaches they prefer. The same is true with the Head of Science and we have decided, at their request, to continue collaborating on this.

I hope that in the next academic year, other Heads of Department will see the benefits of pupil engagement and feedback and implement such feedback forms in their classes too. I also hope that during the coming academic year the thirst for discussion of pedagogy continues across middle leaders.

References

Adam, E., Heissel, J., Hittner, E., Doleac, J., Meer, J., & Figlio, D. (2017) Adolescent cortisol responses to high-stakes testing in school-based settings. Psychoneurodocrinology, 83, pp1-89.

Adeoye- Olatunde, O., & Olenbik, N. (2021) Research and scholarly methods: Semi-structured interviews. Journal of the American College of Clinical Pharmacy, 4, pp.1358-1367.

Adesope, O., Trevisan, D., & Sundararajan, N. (2017) Rethinking the Use of Tests: A Meta-Analysis of Practice Testing. Review of Educational Research, 87 (3), pp. 659-701.

Agarwal, P. (2019). Retrieval Practice & Bloom's Taxonomy: Do Students Need Fact Knowledge Before Higher Order Learning? American Psychological Association, 111 (2), pp. 189-209

Agarwal, P. (2023). Personal Reflections on Science Communication and Sharing Retrieval Practice Research with Teachers. Educational Psychology Review, 36 (3), pp. 1-12

Agarwal, P., Bain, P., & Chamberlain, R. (2012). The Value of Applied Research: Retrieval Practice Improves Classroom Learning and Recommendations from a Teacher, a Principal, and a Scientist. Educational Psychology Review, 24, pp. 437-448

Anderson, D., Harris, I., Allen, S., Satran, L., Bland, C., Davis-Feickert, J., Poland, G., & Miller, W. (1991) Comparing students' feedback about clinical instruction with their performances. Academic Medicine, 66, pp. 29-34.

Awang, Z., Afthanorhan, A., Mamat, M., Bucharest, N. (2016) The Likert scale analysis using parametric based Structural Equation Modelling (SEM). Computational Methods in Social Sciences, 4 (1), pp. 13-21.

Baars, M., Wijnia, L., Bruin, A., & Pass, F. (2020). The Relationship Between Students' Effort and Monitoring Judgements During Learning: A Meta-analysis. Educational Psychological Review, 32, pp. 979-1002.

Bakkenes, I., Vermunt, J., & Wubbels, T. (2010) Teacher learning in the context of educational innovation: Learning activities and learning outcomes of experienced teachers. Learning and Instruction, 20, pp. 533-548.

Bangert-Drowns, R., Kulik, J., & Kulik, C. (1991). Effects of Frequent Classroom Testing. The Journal of Educational Research, 85 (2), pp. 89-99.

Bartoszewki, B., & Gurung, R. (2015) Comparing the Relationship of Learning Techniques and Exam Score. Scholarship of Teaching and Learning in Psychology, 1 (3), pp. 219-228.

Beck, K., Rohr, H., Reme, B., & Flato, M. (2024). Distressing testing: A propensity score analysis of high-stakes exam failure and mental health. Child Development, 95 (1), pp. 242-260.

Bjork, R. & Bjork, E. (1992) A new theory of disuse and an old theory of stimulus fluctuation. Learning processes to cognitive processes: Essays in honor of William K. Estes, 2, pp. 35-67.

Bjork, R., Dunlosky, J., & Kornell, N. (2013) Self-Regulated Learning: Beliefs, Techniques, and Illusions. Annual Review of Psychology, 64, pp. 417-444.

Blasiman, R., Dunlosky, J., & Rawson, K. (2016). The what, how much, and when of study strategies: comparing intended verses actual study behaviour. Memory, 25 (6), pp. 784-792.

Brown, A. & Danaher, P. (2017). CHE Principles: facilitating authentic and dialogical semi-structured interviews in educational research. Internal Journal of Research & Method in Education, 42 (1), pp. 76-90.

Callender, A., & McDaniel, M. (2009). The limited benefits of rereading educational texts. Contempory Educational Psychology, 34, pp. 30-41.

Carifio, J., & Perla, R. (2008). Resolving the 50-year debate around using and misusing Likert scales. Medical Education, 42, pp. 1151-1152.

Carvalho, P., Rachatasumrit, N., & Koedinger, K. (2022). Learning depends on knowledge: The benefits of retrieval practice. Proceedings of the Annual Meeting of the Cognitive Science Society, 44.

Carpenter, S. (2023) Encouraging Students to Use Retrieval Practice: a review of Emerging Research from Five Types of interventions. Educational Psychology Review, 35(96), pp. 1-17

Carpenter, S., Pashler, H., & Cepeda, N. (2008). Using Tests to Enhance 8th Grade Students' Retention of U.S. History Facts. Applied Cognitive Psychology, 23, pp. 760-771.

Cowan, P., & Farrell, R. (2023). Using Virtual Reality to Support Retrieval Practice in Blended Learning: An Interdisciplinary Professional Development Collaboration between Novice and Expert Teachers. Digital, 2, pp. 251-272.

Crawford, R. & Tan, H. (2019). Responding to evidence-based practice: An examination of mixed methods research in teacher education. The Australian Educational Researcher, 46, pp. 775-797.

Cuddy, L., & Jacoby, L. (1982) When Forgetting Helps Memory: An Analysis of Repetition Effects. Journal of Verbal Learning and Verbal Behavior, 21, pp. 451-467.

De Jong, L., Meirink, J., & Admiraal, W. (2022) School-based collaboration as a learning context for teachers: A systematic review. International Journal of Educational Research, 112, pp. 1-15.

Department for Education. (2022). 'How pupils learn' exemplar materials. Accessed 24th June 2024. https://assets.publishing.service.gov.uk/media/637b7702e90e0728475ed534/How_Pupils_Learn_Exemplars.pdf

Department for Education. (2020). National Professional Qualification Headship Framework. https://assets.publishing.service.gov.uk/media/63a1dc9d8fa8f539169b1d74/NPQ_Headship_FINAL_Ref.pdf Accessed 13th June 2024

Desy, J., Busche, K., Cusano, R., Veale, P., Coderre, S., & McLaughlin, K. (2018) How teachers can help learners build storage and retrieval strength. Medical Teacher, 40 (4), pp. 407-413.

Dirkx, K., Camp, G., Kester, L., & Kirschner, P. (2019). Do secondary school students make use of effective study strategies when they study on their own? Applied Cognitive Psychology, 33, pp. 952-957.

Dolnicar, S. (2021). 5/7-point Likert scales aren't always the best option. Their validity is undermined by lack of reliability, response style bias, long completion times and limitations to permissible statistical procedures. Annals of Tourism Research, 91, pp. 1-3.

Douven, I. (2018). A Bayesian perspective on Likert scales and central tendency. Psychonomic Bulletin Review, 25, pp. 1203-1211.

Dutro, E., & Selland, M. (2015) "I like to read, but I know I'm not good at it" Children's perspectives on high-stakes testing in high-poverty school. Curriculum Inquiry, 42 (3), pp. 340-367.

Edwards, R., & Holland, J. (2013). How have qualitative interviews developed? What is Qualitative Interviewing? Bloomsbury Academic, February, pp. 11-28.

Emerson, R. (2017). Likert Scales. Journal of Visual Impairment & Blindness, September, pp. 488-489.

Endres, T., & Renkl, A. (2015). Mechanisms behind the testing effect: an empirical investigation of retrieval practice in meaningful learning. Frontiers in Psychology, 6 (1054), p. 1-6.

Ethical Guidelines for Educational Research, fourth edition (2018). British Educational Research Association. https://www.bera.ac.uk/publication/ethical-guidelines-for-educational-research-2018-online (accessed 15th February 2024)

Gates, A. (1917) Recitation as a factor in memorizing. Archives of Psychology, 40, pp. 1-104.

Gaviao, L., Sant'Anna, A., Lima, G., & de Almada Garcia, P. (2023). Composition of Probabilistic Preferences in Multicriteria Problems with variables Measured in Likert Scales and Fitted by Empirical Distributions. Standards, 3, pp. 268-282.

Geller, J., Toftness, A., Armstrong, P., Carpenter, S., Manz, C., Coffman, C., & Lamm, H. (2018) Study strategies and beliefs about learning as a function of academic achievement and achievement goals. Memory, 26 (5), pp. 683-690.

Glover, J., Krug, D., Hannon, S., and Shine, A. (1990) The Testing Effect and Restricted Retrieval Rehearsal. The Psychological Record, 40, pp. 215-226.

Gunderman, R., & Chan, S. (2013). The 13-Point Likert Scale. Academic Radiology, 20, pp. 1466-1467.

Hartwig, M., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? Psychonomic Bulletin and Review, 19, pp. 126-134.

Higham, P., Fastrich, G., Potts, R., Murayama, K., Pickering, J., & Hadwin, J. (2023). Spaced Retrieval Practice: Can Restudying Trump Retrieval? Educational Psychology Review, 35 (98), pp. 1-41.

Hinze, S., & Rapp, D. (2014). Retrieval (Sometimes) Enhances Learning: Performance Pressure Reduces the Benefits of Retrieval Practice. Applied Cognitive Psychology, 28, pp. 597-606.

Hostetter, A., Penix, A., Norman, M., Batsell Jr, W., & Carr, T. (2019). The role of retrieval practice and analogical problem-solving. Quarterly Journal of Experimental Psychology, 72 (4), pp. 858-871.

Howard, E. (2020). A review of the literature concerning anxiety for educational assessments. Ofqual research report. Retrieved 29th July 2024. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/865832/A_review_of_the_literature_concerning_anxiety_for_educational_assessment.pdf

Hui, L., Bruin, A., Donkers, J., & Merrienboer, J. (2022). Why students do (or do not) choose retrieval practice: Their perceptions of mental effort during task performance matter. Applied Cognitive Psychology, 36, pp. 433-444.

Jacob, S., & Furgerson, S. (2012). Writing Interview Protocols and Conducting Interviews: Tips for Students New to the Field of Qualitative Research. The Qualitative Report, 17 (42), pp. 1-10.

Jagerskog, A., Jonsson, F., Selander, S., & Jonsson, B. (2019). Multimedia learning trumps retrieval practice in psychological teaching. Scandinanian Journal of Psychology, 60, pp. 222-230.

Jerrim, J. (2022) The power of positive emotions? The link between young people's positive and negative affect and performance in high-stakes examinations. Assessment in Education: Principles, Policy and Practice, 29 (3), pp. 310-331.

Johnson, R., & Onwuegbuzie, A. (2004). Mixed Methods Research: A Research Paradigm Whole Time Has Come. Educational Researcher, 33 (7), pp. 14-26.

Jones, K. (2019) Retrieval Practice: Resources and research for every classroom. John Catt Educational Ltd, Suffolk.

Kang, S., McDermott, K., & Roediger, H, III. (2007) Test format and corrective feedback modify the effect of testing on long-term retention. European Journal of Cognitive Psychology, 19, pp. 528-558.

Karlsen, A. & Ohna, S. (2021) Pupil voices in teachers' collaborative professional learning in Lesson Study. International Journal of Educational Research, 100, pp. 1-13.

Karpicke, J., & Blunt, J. (2011). Retrieval Practice Produces More Learning than Elaborative Studying with Concept Mapping. Science Express, January 2011, pp. 1-5.

Karpicke, J., & Roediger, H. (2008) The critical importance of retrieval for learning. Science, 319 (5865) pp. 966-968.

Keeley, J., Smith, D., & Buskist, W. (2006). The Teacher Behaviors Checklist: Factor Analysis of Its Utility for Evaluating Teaching. Teaching of Psychology, 33 (2), pp. 84- 91.

Klier, C., & Buratto, L. (2023). The benefit of retrieval practice on cued recall under stress depends on item difficulty. Neuroscience Letters, 797, pp. 1-9

Knose, L., Rawson, K., & Dunlosky, J. (2020). How much do college students with ADHD benefit from retrieval practice when learning key-term definitions? Learning and Instruction, 68, pp. 1-11.

Knapp, T. (1990). Treating ordinal scales as interval scales: An attempt to resolve the controversy. Nursing Research, 39, pp. 121-123.

Kornell, N., & Bjork, A. (2007). The promise and perils of self-regulated study. Psychonomic Bulletin and Review, 14 (2), pp. 219-224.

Kornell, N., Bjork, R., & Garcia, M. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. Journal of Memory and Language, 65, pp. 85-97.

Kuhbandner, C., & Emmerdinger, K. (2019). Do students really prefer repeated rereading over testing when studying textbooks? A re-examination. Memory, 27 (7), pp. 952-961.

Kutrovatz, K. (2017). Conducting Qualitative Interviews with Children – Methodological and Ethical Challenges. Corvinus Journal of Sociology and Social Policy, 8, pp. 65-88.

Kvale, S. (2006). Dominance Through Interviews and Dialogues, Qualitative Inquiry, 12 (3), pp. 480-500.

Lakshmi, Y. (2019). Mixed Methods Research in Education. International Journal of Higher Education and Research, 9 (1), pp. 281-290.

Leggett, J., Burt, J., & Carroll, A. (2018). Retrieval practice can improve classroom review despite low practice test performance. Applied Cognitive Psychology, 33, pp. 757-770.

Letterman, D. (2013). Students' Perception of Homework Assignments and What Influences Their Ideas. Journal of College Teaching and Learning, 10 (2), pp. 113-121.

MacDaniel, M., Agarwal, P., Huelser, B., McDermott, K., & Roediger III, H. (2011). Test-enhanced Learning in a Middle School Science Classroom: The Effects of Quiz Frequency and Placement. Journal of Educational Psychology, 103 (2), pp. 399-414.

Mason, J. (2006). Mixing methods in a qualitatively driven way. Qualitative Research, 6 (1), pp. 9-25.

Marsh, H. (2012). Relationships for learning: using pupil voice to define teacher – pupil relationships that enhance pupil engagement. Management in education, 26 (3), pp. 161-163.

McDaniel, M., Kowitz, M., & Dunay, P. (1989). Altering memory through recall. The effects of cue-guided retrieval processing. Memory and Cognition, 11, p. 423-434.

McDaniel, M., & Masson, M. (1985). Altering memory representations through retrieval. Journal of Experimental Psychology: Learning, Memory and Cognition, 11, p.371-385.

Minear, M., Coane, H., Boland, S., Cooney, L., & Albat, M. (2018). The Benefits of Retrieval Practice Depend on Item Difficulty and Intelligence. Journal of Experimental Psychology, 44 (9), pp. 1474-1486.

Morano, S. (2019). Retrieval Practice for Retention and Transfer. Teaching Exceptional Children, 51 (6), pp. 436-444.

Morehead, K., Rhodes, M., & DeLozier, S. (2016). Instructor and student knowledge of study strategies. Memory, 24 (2), pp. 257-271.

Nadler, J., Weston, R., Voyles, E. (2015). Stuck in the Middle: The Use and Interpretation of Mid-Points in Items on Questionnaires. The Journal of General Psychology, 142 (2), pp. 71-89.

Nelson, T., & Dunlosky, J. (1994) Norms of paired-associate recall during multitrial learning of Swahili-English translation equivalents. Memory, 2, pp. 325-335.

Nicholls, M., Orr, C., Okubo, M., & Loftus, A. (2006). Satisfaction Guaranteed. The Effect of Spatial Biases on Responses to Likert Scales. Psychological Science, 17 (12), pp. 1027-1028.

OFSTED. (2019) Education Inspection Framework. Overview of research. https://assets.publishing.service.gov.uk/media/6034be17d3bf7f265dbbe2ef/Research_for_EIF_framework_updated_references_22_Feb_2021.pdf Accessed 13th June 2024.

Pashler, H., Rohrer, D., Cepeda, N., & Carpenter, S. (2007). Enhancing learning and retarding forgetting: Choices and consequences. Psychonomic Bulletin and Review, 14 (2), pp. 187-193.

Pearce, J. (2021) Retrieval practice – A common good or just commonplace? Educational Endowment Foundation blog. Retrieved 5th July 2024. https://educationendowmentfoundation.org.uk/news/guest-blog-retrieval-practice-a-common-good-or-just-commonplace

Putwain, D. (2008) Test anxiety and GCSE performance: The effect of gender and socio-economic background. Educational Psychology in Practice, 24 (4), pp. 319-334.

Pyc, M., & Rawson, K. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? Journal of Memory and Language, 60, pp. 437-447.

Rawson, K., Dunlosky, J., & Thiede, K. (2000). The rereading effect: Metacomprehension accuracy improves across reading trials. Memory and Cognition, 28 (6) pp. 1004-1010.

Rawson, K., & Kintsch, W. (2005). Rereading Effects Depend on Time of Test. Journal of Educational Psychology, 97 (1), pp. 70-80.

Roediger III, H., & Butler, A. (2011). The critical role of retrieval practice in long-term retention. Trends in Cognitive Sciences, 15 (1) pp.20-26.

Roediger III, H., & Karpicke, D. (2006). Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. Psychological Science, 17 (3), pp. 249-255.

Rowland, C. (2014). The Effect of Testing Verses Restudy on Retention: A Meta-Analytic Review of the Testing Effect. American Psychological Association, Psychological Bulletin, 140 (6), pp. 1432-1463.

Rowland, C., & DeLosh, E. (2014). Mnemonic benefits of retrieval practice at short retention intervals. Memory, 23 (3), pp. 403-419.

Rodrigues de Lima, M., Cavendish, B., Silva de Deus, J., & Buratto, L. (2020). Retrieval Practice in Memory – and Language – Impaired Populations: A Systematic Review. Archives of Clinical Neuropsychology, 35, pp. 1078-1093.

Rosenshine, B. (2010) Principles of Instruction. Educational Practices Series, 21. International Academy of Education.

Rosenshine, B. (2012) Principles of Instruction – research-based strategies that all teachers should know. American Educator, 36 (1), pp. 12-39.

Rowley, T., & McCrudden, M. (2020)Retrieval practice and retention of course content in a middle school science classroom. Applied Cognitive Psychology, 34, pp. 1510-1515.

Russo, J., & Hopkins, S., (2023). Not so Simple Addition: Comparing Student Performance and Teacher Perceptions of Retrieval. International Journal of Science and Mathematical Education, 21, pp. 2279-2301.

Sana, F., & Yan, V. (2022). Interleaving Retrieval Practice Promotes Science Learning. Psychological Science, 35 (7), pp. 782-788.

Schuetze, B., Eglington, L., & Kang, S. (2019). Retrieval practice benefits memory precision. Memory, 27 (8), pp. 1091-1098.

Sherrington, T. (2019) Rosensine's principles in action. John Catt Education, Suffolk.

Smith, M., Blunt, J., Whiffen, J., & Karpicke, J. (2016). Does Providing Prompts During Retrieval Practice Improve Learning? Applied Cognitive Psychology, 30, pp. 544-553.

Solorzano, R. (2019) High Stakes Testing and Educational Ineqlauty in K-12 Schools. Oxford Research Encyclopaedia of Education, December 2019.

Stine-Morrow, E., Gagne, D., Morrow, D., & DeWall, B. (2004). Age differences in rereading. Memory and Cognition, 32 (5), pp. 696-710.

Storm, B., Bjork, R., & Storm, J. (2010). Optimizing retrieval as a learning event: When and why expanding retrieval practice enhances long-term retention. Memory and Cognition, 38 (2), pp. 244-253.

Toppino, T., LaVan, M., & Iaconelli, R. (2018). Metacognitive control in self-regulated learning: Conditions affecting the choice of restudying verses retrieval practice. Memory and Cognition, 46, pp. 1164-1177.

Tse, C. & Pu, X. (2012) The effectiveness of test-enhanced learning depends on trait test anxiety and working memory capacity. Journal of Experimental Psychology: Applied, 18, pp. 253-264.

Tullis, J., Finley, J., & Benjamin, A. (2013). Metacognition of the testing effect: Guiding learners to predict the benefits of retrieval. Memory and Cognition, 41, pp. 429-442.

Vandenbussche, J., Griffiths, W., & Scherrrer, C. (2014). Students' Perceptions of Homework Policies in Lower- and Intermediate- level Mathematical Courses, Mathematics and Computer Education, 48 (2), pp. 149- 163.

Vangrieken, K., Dochy, F., Raes, E., & Kyndt, E. (2015) Teacher collaboration: A systematic review. Educational Research Review, 15, pp. 17-40.

Vaughn, K., Hausman, H., & Kornell, N. (2017). Retrieval attempts enhance learning regardless of time spent trying to retrieve. Memory, 25 (3), pp. 298-316.

Voskamp, A., Kuiper, E., & Volman, M. (2020). Teaching practices for self-directed and self-regulated learning: case studies in Dutch innovative secondary schools. Educational Studies, 48 (6), pp. 772-789.

Weissgerber, S., & Rummer, R. (2023). More accurate than assumed: Learners' metacognitive beliefs about the effectiveness of retrieval practice. Learning and Instruction, 83, pp. 1-10.

Wiklund-Hornqvist, C., Stillesjo, S., Andersson, M., Jonsson, B., & Nyberg, L. (2020) Retrieval practice facilitates learning by strengthening processing in both the anterior and posterior hippocampus. Brain and Behaviour, 11, pp. 1-9.

Wu, H., & Leung, S. (2017). Can Likert Scales be Treated as Interval Scales? – A Simulation Study. Journal of Social Service Research, 43 (4), pp. 527-532.

Ysseldyke, J., & Nelson, J. (2004) What we know and need to know about the consequences of High-Stakes testing for students with disabilities. Exceptional Children, 71 (1), pp. 7-120.

Appendices

Appendix 1: CUREC FORM

SOCIAL SCIENCES & HUMANITIES

INTERDIVISIONAL RESEARCH ETHICS COMMITTEE

DEPARTMENTAL RESEARCH ETHICS COMMITTEE

Department of Education

15 Norham Gardens, Oxford OX2 6PY

student.curec@education.ox.ac.uk; staff.curec@education.ox.ac.uk

[redacted]

Department of Education, Social Sciences Division

University of Oxford

8 March 2024

Dear [redacted],

Research Ethics Approval Research Title: Do staff and pupils believe that structured retrieval practice in KS4 Science prepares the pupils for their end of year and GCSE examinations?

Research Ethics Reference: EDUC_C1A_24_037

The above application has been considered on behalf of the Education Departmental Research Ethics Committee (DREC) in accordance with the University's procedures for ethical approval of all research involving human participants.

I am pleased to confirm that, on the basis of the information provided to the DREC, ethics approval has now been granted for this study. Please note the following:

- **Personal data**: It is the responsibility of the PI to ensure that all personal data collected during the project is managed in accordance with the University's guidance and legal requirements.
- **In-person activities**: Any data collection involving in-person interactions with participants must have an up-to-date fieldwork risk assessment in place; further guidance is available from the Safety Office's website.
- **Amendments**: Please notify the committee if you intend to make any amendments to the information in your ethics application as submitted at date of this approval, as all changes must receive ethical approval prior to implementation. The amendment form is available on the SSH IDREC webpage.

We welcome feedback on your experience of the ethical review process and suggestions for improvement. Please email any comments you might have to staff.curec@education.ox.ac.uk / student.curec@education.ox.ac.uk or ethics@socsci.ox.ac.uk.

Yours sincerely,

[redacted]

Departmental Lecturer in Applied Linguistics

DREC Member

[redacted]

Appendix 2: Microsoft Form Template for the survey

Survey of Year 10 and Year 11 on revision and testing techniques in class

Thank you for taking part in this research project. This is being conducted as part of some educational research linked to a MSc research degree at the University of Oxford. The focus of the research is consider the learning styles of pupils in Year 10 and Year 11 who are taking GCSE examinations in Science.

There is no compulsion in completing this research and you do not have to complete this questionnaire. However, it will only take 6 minutes and the research will be used by the School when shaping teaching and learning in future years.

I thank you very much for completing this short questionnaire. A copy of the results will be available at the end of the academic year to those who are interested.

Thank you

* Required

1. Year Group *

- ◯ Year 10
- ◯ Year 11

2. Gender *

- ◯ Male
- ◯ Female
- ◯ Non-binary

3. Science qualification being studied *

- ◯ Triple Award
- ◯ Combined Science (Higher)
- ◯ Combined Science (Foundation)

4. Please add your initials (these will not be used in any written findings) *

5 The next few que tions relate to how you revise at home and out of class. Please now move onto Question 6 *

6. I have devised a revision timetable for my exams *

Strongly agree	Agree	Neutral	Disagree	Disagree	Strongly disagree
○	○	○	○	○	○

7. I often re-read my notes during revision *

Strongly Agree	Agree	Neutral	Disagree	Strongly disagree
○	○	○	○	○

8. I find revising at home easy *

Strongly Agree	Agree	Neutral	Disagree	Strongly disagree
○	○	○	○	○

9. I highlight my notes during revision *

Strongly Agree	Agree	Neutral	Disagree	Strongly disagree
○	○	○	○	○

10. I use aids like Post It Notes *

Strongly Agree	Agree	Neutral	Disagree	Strongly disagree
○	○	○	○	○

11. I write down the main points into a short summary *

Strongly agree	Agree	Neutral	Disagree	Strongly disagree
○	○	○	○	○

12 I test myself by a king myself questions about the topic *

Strongly agree	Agree	Neutral	Disagree	Strongly disagree
○	○	○	○	○

13. I make lists of key points *

Strongly agree	Agree	Neutral	Disagree	Strongly disagree
○	○	○	○	○

14. I struggle with my organisation when it comes to revising on my own *

Strongly agree	Agree	Neutral	Disagree	Strongly disagree
○	○	○	○	○

15. The harder you revise the better the outcome *

Strongly agree	Agree	Neutral	Disagree	Strongly disagree
○	○	○	○	○

16. I make lists of key vocabulary and points *

Strongly agree	Agree	Neutral	Disagree	Strongly disagree
○	○	○	○	○

17. I tend to revise the easier parts of the specification first *

Strongly agree	Agree	Neutral	Disagree	Strongly disagree
○	○	○	○	○

18. Once I feel that I understand a concept I tend not to spend time revising it again *

Strongly agree	Agree	Neutral	Disagree	Strongly disagree
○	○	○	○	○

19 Thank you for an wering these questions. The final section of this survey relates to your Science lessons. At the start of many lessons you are undertaking retrieval activities, which test your understanding of prior topics. These questions relate to this and other aspects of your Science lessons too. Please move immediately to Question 20. *

20. I like my teacher testing us on previous work at the start of the lesson *

Strongly agree	Agree	Neutral	Disagree	Strongly disagree
○	○	○	○	○

21. The revision at the start of the lesson is a good use of time *

Strongly agree	Agree	Neutral	Disagree	Strongly disagree
○	○	○	○	○

22. I would prefer that the revision activities were set as homework rather than be at the start of the lesson *

Strongly agree	Agree	Neutral	Disagree	Strongly disagree
○	○	○	○	○

23. I prefer learning new things to revising in class *

Strongly agree	Agree	Neutral	Disagree	Strongly disagree
○	○	○	○	○

24. I would prefer that the teacher provides us with information about previous topics rather than testing us at the start of the lesson *

Strongly agree	Agree	Neutral	Disagree	Strongly disagree
○	○	○	○	○

25. Retrieval testing at the start of the lesson is a waste of time *

Strongly agree	Agree	Neutral	Disagree	Strongly disagree
○	○	○	○	○

26 I find being testing on topics we covered a while ago more beneficial than being tested on recent topics *

Strongly agree	Agree	Neutral	Disagree	Strongly disagree
○	○	○	○	○

27. I find it more helpful when we get feedback after the retrieval activity *

Strongly agree	Agree	Neutral	Disagree	Strongly disagree
○	○	○	○	○

28. I find it helpful when we are given cues (or clues) about the answers needed *

Strongly agree	Agree	Neutral	Disagree	Strongly disagree
○	○	○	○	○

29. I much prefer being tested on very recent topics *

Strongly agree	Agree	Neutral	Disagree	Strongly disagree
◯	◯	◯	◯	◯

Appendix 3: Pupil Questions

1. How do you find revising at home for exams? (I)
2. What about testing yourself? (IV)
3. What about re-reading your notes. Do you do this in revision? (IV)
4. And what do you think about the retrieval exercises at the start of the lessons in Science? (I)
5. What are your thoughts on the frequency of the retrieval activities? (I)
6. Finally, what do you think about hints and cues? (II and III)

Figure 2: Core Pupil Interview Questions with research question in brackets

Appendix 4: Staff Questions

1. What have you been doing in the Department?
2. And do you undertake RP at the start of a lesson? (I)
3. So not just a quiz? (I)
4. Does it benefit the pupils? (I)
5. Are you saying that failing RP can be detrimental? (I)
6. What is your view on providing cues? (II)
7. What are your views on providing feedback (III)
8. And what about your view of the length of duration? (I)
9. What do you think the pupils' perception of RP is? (I)
10. And is this common place? (I)
11. So, does RP lead to an increase in workload? (I)
12. And do your Do It Now's all have RP in them? (I)
13. Do you think it is therefore worthwhile? (I)

Figure 3: Core Interview Questions with teaching staff with research question in brackets