Studia Psychologica, Vol. 63, No. 2, 2021, 204-220
https://doi.org/10.31577/sp.2021.02.821

# Teachers' Beliefs about Memory: A Vignette Study of Trainee and In-Service Teachers

Jonathan Firth
School of Education, University of Strathclyde, Glasgow, United Kingdom

Previous research has indicated that laypeople, students and legal professionals often hold flawed beliefs about memory, and the present study sought to extend this area of research to the teaching profession. Are teachers' beliefs about learning in line with the scientific consensus? A set of vignettes with contrasting options for classroom practice were presented to trainee ($n$ = 77) and in-service ($n$ = 44) teachers, and in each case a 7-point Likert scale prompted them to predict which would be the best course of action in terms of student outcomes. As hypothesized, responses were often out of line with research on 'desirable difficulties' in memory and learning such as retrieval practice, spacing, and interleaving, with choices indicating a lack of awareness of these evidence-based approaches, although they were more accurate than previous studies of students. Surprisingly, accuracy of response did not correlate with the duration of a teacher's classroom experience; trainee teachers outscored in-service teachers in certain areas, suggesting that recent familiarity with technical literature on learning could be advantageous.

*Key words*: beliefs, teaching practice, professional learning, desirable difficulties, metacognition

## Introduction

Memory is complex and its workings are not intuitively obvious, leading to what Pan and Bjork (2020) refer to as a faulty mental model of ourselves as learners. Perhaps because of this fact, misconceptions about memory appear to be widespread. Surveys by Simons and Chabris (2011, 2012) found numerous deviations between the views of the American public and the scientific consensus. For example, 82.7% of their sample strongly agreed or mostly agreed with the statement "*people suffering from amnesia typically cannot recall their own name or identity*" (Simons & Chabris, 2011, p. 3), but 0% of memory researchers questioned did so.

Flawed understandings of memory can be found among professionals too, perhaps

Correspondence concerning this article should be addressed to Jonathan Firth, School of Education, The University of Strathclyde, 141 St James Road, Glasgow, Scotland, United Kingdom.
E-mail: jonathan.firth@strath.ac.uk

Received August 24, 2020

most notably in legal settings. In a study that focused on the recollections of eyewitnesses to a crime, Melinder and Magnussen (2015) found that psychologists in Norway serving as expert witnesses in court often endorse memory myths, following up on an earlier study which found that judges make similar mistakes (Magnussen, Wise, Raja, et al., 2008). Neither group of professionals were any better than the general public overall in terms of their understanding of memory.

What is less clear is whether similar flawed beliefs about memory are held by the teaching profession. Teachers make daily decisions that have an impact on learning and memory, and their job performance is often judged on the basis of student grades (Kennedy, 2008). They are members of a highly educated profession, who engage regularly in learning-related ongoing professional development, and who also benefit from practical experience, which may help to illuminate how memory operates in the classroom. It is worth considering, then, what role education and experience are likely to have on the accuracy of their beliefs. If their beliefs are indeed as flawed as those held by other groups, then teaching decisions that relate to memory and learning could be flawed as well (Kornell & Bjork, 2007).

Recent research suggests that prior education has rather limited benefits in this area. Furnham (2018) presented survey participants with a set of myths about the mind and brain, and found that participants' education level did not correlate with accurate identification of myths, and that background study, specifically of psychology, was also unrelated. However, Furnham acknowledged that his work did not delve especially deeply into the level and manner in which participants had previously studied psychology; the questions also did not focus just on memory.

A teacher's experience level may be important for a number of reasons, but improvement in performance over time cannot be assumed (Wiliam, 2010). Indeed, it is perfectly possible – as found in other professions – for performance to deteriorate rather than improve post-training (Ericsson, 2017), and experience could also lead teachers to become overconfident in their own abilities. American data on student grades suggest that teachers do tend to become more effective for a few years immediately after the preparation/training stage and then plateau (Hood, 2016). However, early improvements could be due to a number of factors such as better classroom management or course knowledge – they need not imply a better understanding of learning and memory. There is also evidence that teachers' endorsement of myths such as 'learning styles' does not reduce in line with years of experience (Morehead, Rhodes, & DeLozier, 2016).

Given the doubts over whether either knowledge or experience will provide teachers with a full understanding of memory, and the misconceptions found among other samples, the present study will investigate the accuracy of teachers' beliefs about memory as judged by responses to vignette-based classroom examples. In particular, I will focus on three well-established strategies relating to the manner in which new learning takes place, all underpinned by research into human long-term memory. These techniques include retrieval practice (active recall activities are more effective than re-reading), the spacing effect (incorporating a delay before restudying leads to more durable learning compared to immediate restudy), and interleaving (mixing different types of practice items is beneficial). Dunlosky and Rawson (2015) provide a useful and accessible summary of these issues as they apply to education in their 'teacher-ready review', noting that the evidence for retrieval practice and spacing is 'strong', while the evidence for interleaved practice is 'moderate' (and the evi-

dence base for the latter has since expanded; see for example Brunmair & Richter, 2019).

These three techniques can all boost long-term learning, and they contrast with popular but ineffective learning strategies such as highlighting of notes and cramming (Dunlosky & Rawson, 2015). However, as *desirable difficulties*, they can all present significant challenges to shorter-term performance. As discussed in a review by Soderstrom and Bjork (2015), performance and learning are often negatively correlated (an immediate practice session, for example, would be quite easy for a student, but is not the best way to revise or consolidate). A practical implication is that the benefits of such techniques may not be obvious to learners during study or for some time afterwards.

For this reason, we may predict that students will make some very flawed study choices (as judged by long-term learning outcomes). And indeed, both field and laboratory research has revealed that participants tend to avoid and misunderstand effective learning techniques even when given a chance to try them (Kornell & Bjork, 2007; Kornell & Son, 2009; McCabe, 2011; Piza, Kesselheim, Perzhinsky, Drowos, Gillis, et al., 2019; Yan, Bjork, & Bjork, 2016; Zechmeister & Shaughnessy, 1980), with an impact on their grades (Hartwig & Dunlosky, 2011). It would therefore appear to fall to teachers to recommend these strategies. However, for this to be achieved, teachers themselves need an accurate understanding of desirable difficulties. How likely is it that teachers do indeed have this understanding?

There is some evidence that educators also favor flawed and passive learning strategies (Hunter & Lloyd, 2018), while student teachers' knowledge of learning strategies has been described as "knowledge in pieces" (Glogger-Frey, Deutscher, & Renkl, 2018, p. 228) – fragmentary, and with different elements rarely compared or formulated into a coherent mental model. In their review of performance and learning, Soderstrom and Bjork state several times that teachers are likely to misunderstand memory, but they do not report empirical evidence of this, for example suggesting that: "*fleeting gains during acquisition are likely to fool instructors…into thinking that permanent learning has taken place, creating powerful illusions of competence*" (Soderstrom & Bjork, 2015, p. 193). While this statement makes theoretical sense, it is important that any lack of metacognitive understanding of memory processes is established empirically, and in a sample of school teachers specifically.

A study by Firth (2018) found evidence that misconceptions about memory may be at a lower level among teachers than has been found in the general population in previous research. However, it revealed poor performance (in terms of being out of line with the research consensus) in response to items asking about spacing and retrieval practice, albeit via brief questions about memory which were presented out of context. The current research will improve on the methodology via the use of vignettes, as used by McCabe (2011, 2018). In the 2011 study, McCabe presented vignettes relating to desirable difficulties to students and found very poor accuracy in responses, with (for example), fewer than 10% endorsing spacing over massing. In the 2018 study, McCabe found evidence that academic study centers showed some support for evidence-based strategies, but neglected the benefits of spacing and interleaving.

The Firth (2018) study also provided evidence that beliefs do not become more accurate in line with years spent teaching, and this accords with the work of Morehead et al. (2016). Further, when Čavojová and Jurkovič (2017) studied the cognitive processes of teachers in Slovakia, experienced professionals

performed no better than trainees with the exception of an intertemporal choice task about patient outcomes. The task in their study was in some ways analogous to the application of desirable difficulties such as the spacing effect in that success involved suppressing impulsive and intuitive short-term gains in favor of correct ones which involve demanding long-term thinking. A study by Halamish (2018) tested this matter more directly; in her study of in-service and trainee teachers, both groups failed to endorse evidence-based strategies; again, there was no indication of an improvement with years of the experience, and indeed the more experienced teachers were less accurate in their judgements.

Together, these findings suggest that teacher competence and experienced-based improvements in their understanding of memory cannot be assumed. But it is important to follow up on these findings using teachers in a different setting. To do so, I will aim to study the memory beliefs of UK-based trainee teachers as a baseline, and to compare their performance with that of their more experienced peers as done by Halamish (2018). To distinguish book learning from insights gained from experience it will be helpful to gauge the level of prior knowledge that participants have relating to desirable difficulties.

It will also be useful to include a simple metacognitive measure asking in-service teachers to declare their confidence in their own responses after completing all of the classroom vignettes, as confidence is a key aspect of skilled professional practice (Nola & Molla, 2017), but confidence in metacognitive errors could pave the way for flawed classroom decisions. Lower confidence may also suggest a likelihood of participants autonomously engaging with future professional learning.

Four specific questions arise from the points above:

*1) How accurate are the participants in responding to vignettes about spacing, interleaving and retrieval practice?* Scores on the individual desirable difficulties will be compared to the neutral response on the Likert scale using *t*-tests.

*2) Are in-service teachers more accurate than trainee teachers?* The overall accuracy of both groups across all vignettes will be compared via a one-way ANOVA.

*3) Is there a relationship between years of classroom experience and response accuracy?* Experience, operationalized as number of years in service, will be compared with overall accuracy across vignettes for the in-service teachers only.

*4) Does a teacher's declared level of confidence in their own answers correlate to their accuracy in responding?* Metacognitive judgements will be compared to reality by asking participants to judge the accuracy of their own responses.

## Method

### Participants

Two groups were sampled. First, ethical permission was obtained to sample postgraduate trainee teachers at the Scottish university where the author works, the largest of the six institutions which are responsible for training all of the new teachers in the Scottish education system. Sampling was done by convenience; tutors in five secondary teaching subjects (Biology, Business, Psychology, Computing, and English) distributed the study to their student teachers, who then had the option of taking part if they chose to do so. 77 took part, with a mean age of 31.2 years (*SD* = 9.25).

The second group of participants consisted of experienced school teachers (*n* = 44). This part of the sample was obtained by approach-

ing local authorities for ethical approval, and then emailing the survey to headteachers with a request to distribute it. Two authorities participated (out of 32 in Scotland). In addition, to broaden the representativeness of the sample, one independent school was also invited to participate via a personal contact in the school management (independent schools comprise around 5% of Scotland's school system), as were a small number of retired teachers, again through personal contacts.

**Materials and Design**

Using a survey-based task, trainee and in-service teachers were compared in terms of their responses to three types of desirable difficulties (a 2 x 3 design), together with a correlational analysis of in-service teachers' experience level versus their accuracy on the task. The survey was prepared by the author and distributed via an online protocol on the PsyToolkit website (www.psytoolkit.org; see Stoet, 2017). It featured 3 demographic questions followed by nine scenarios/vignettes relating to memory in a classroom context.

Multiple subject areas were included across the vignettes; each vignette presented a professional choice relating to spacing, interleaving, or retrieval practice (3 of each), and were 104 words long on average. These are shown in Appendix 1. I drew two scenarios from McCabe (2011, 2018) with minor modifications to the wording to make the scenarios more recognizable to UK-based participants. The remaining seven are novel, developed to follow a similar style and length.

There was then a question asking participants to state how confident they felt in their own answers (practicing teachers only; for trainees, confidence was seen to be less relevant as it could not derive from reflections on classroom practice), and then three questions which asked them to report how well they felt they understood the three concepts at hand. The stages of the survey are summarized in Figure 1.

**Procedure**

Trainee teacher data were collected over weeks 3–4 of an initial teacher education course, in September 2019. This allowed the participants to be questioned at the very start of their teaching career, before they had spent any supervised teaching time in the classroom. The other participating teachers were surveyed between October and February of the same year.

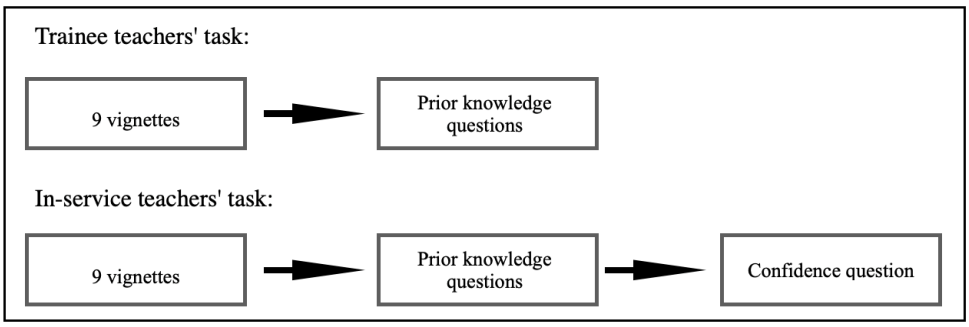


*Figure 1* Summary of question stages.

Scenarios were presented in the following order: *interleaving 1; spacing 1; retrieval 1; interleaving 2; retrieval 2; spacing 2; spacing 3; retrieval 3; interleaving 3*. Items were accompanied by a 7-point Likert scale which asked participants to judge which of the two options in the vignette would be more effective for learning. The terms 'interleaved', 'spaced' and 'retrieval' were avoided in favor of 'mixed' (vs. grouped), 'spread out' (vs. intensive) and 'a test' (vs. restudy), respectively. This decision aimed to ensure that participants thought about the scenario at hand, rather than relying on memory for terms which they may have seen recommended during professional reading (McCabe, 2011). In every case, endorsement of the desirable difficulty was viewed as the notional 'correct' answer.

All participants received a (software generated) unique participant code, allowing them to withdraw from the study in the period between completion of the task and the beginning of data analysis if they chose to do so.

**Measurements**

For all aspects of the analysis I reverse-scored four of the nine vignette items such that '7' always represented the optimal alternative for every question.

Comparison of each individual vignette drew on all responses for both groups of teachers combined, based on mean scores on the Likert scale; data from incomplete questionnaires were retained for this process.

However, for analysis of overall/consistent endorsement of the concepts, it was only meaningful to analyze complete questionnaires ($n$ = 100; 33 in-service teachers and 67 trainee teachers). The same data were used to compare each group's analysis of the three concepts overall to the midpoint (see Table 1).

For a more sensitive comparison between the two groups of participants it was desirable to use all data where possible, and so multiple imputation (MI) following analysis of the missing data[1] was used to replace the missing values for this analysis. The same imputed data were used to compare in-service teachers' confidence in their answers, for which they responded on a four-point scale: *not at all confident; slightly confident; moderately confident; very confident.*

Finally, when asked how well they believed they understood the three concepts, all participants were asked to select one of four responses: *not at all; very little; quite well, or very well*.

**Results**

The overall responses to the vignettes can be seen in Table 1, organized by concept. As can be seen, in all but one case the responses ranged from the minimum (1) to the maximum (7). The three interleaving vignettes averaged below the midpoint, and the others above.

Rates of consistent endorsement of a strategy across all three examples of that strategy are shown in Figure 2. This was operationalized as a score of 5 or higher on all three of the relevant vignettes (in line with the procedure followed by McCabe, 2011).

As can be seen from Figure 2, all three strategies were endorsed consistently by a minority of participants overall, although the overall total was close to the halfway point for spacing; slightly more than half (53.7%) of trainee teachers endorsed the spacing effect on every occasion. To investigate this pattern further, for both groups of participants, each scenario type (e.g., interleaving) was compared to the

[1] 25% of cases in the original data set had missing data with 8.3% of values missing overall. The main cause of missing data was participants dropping out mid-way through the task and failing to complete the final questions. No other pattern of bias was apparent.

Table 1 *Overall responses to each vignette*

| Vignette | *n* | *M* | *SD* | Range |
|---|---|---|---|---|
| Retrieval 1 | 109 | 4.29 | 2.04 | 6 |
| Retrieval 2 | 103 | 5.49 | 1.64 | 6 |
| Retrieval 3 | 101 | 4.45 | 1.81 | 6 |
| Spacing 1 | 110 | 4.95 | 1.89 | 6 |
| Spacing 2 | 101 | 5.22 | 1.64 | 6 |
| Spacing 3 | 101 | 4.75 | 1.79 | 6 |
| Interleaving 1 | 119 | 2.84 | 1.31 | 6 |
| Interleaving 2 | 104 | 3.49 | 1.68 | 5 |
| Interleaving 3 | 100 | 3.06 | 1.57 | 6 |

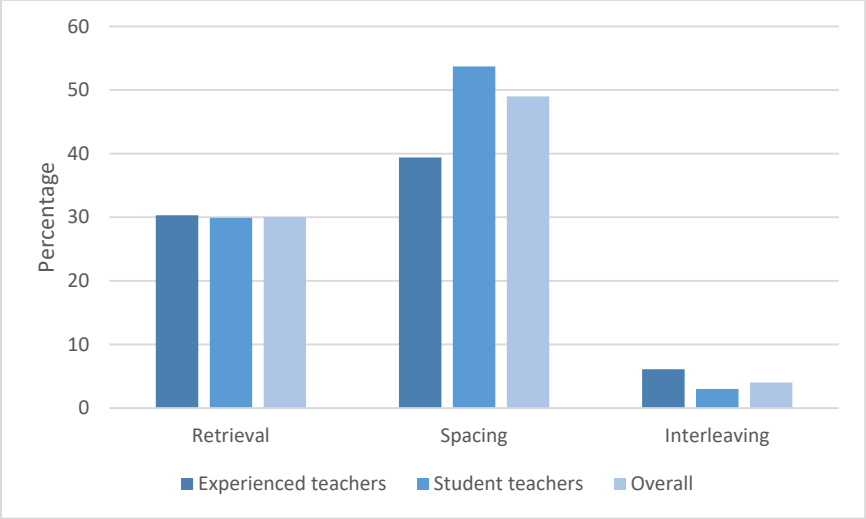


*Figure 2* Percentage of participants who consistently endorsed the desirable difficulties presented in the vignettes.

neutral score of 4 using a one-sample *t*-test. The findings are shown in Table 2.

Means for both retrieval and spacing were significantly above the mid-point for student teachers, indicating accuracy higher than that reported by Halamish (2018), who reported means below 4 for every category. For interleaving the responses were significantly below the midpoint. The effect sizes were large, in particular the result for spacing with in-service teachers (Cohen's $d$ = 1.78), and the two interleaving findings. Curiously, in contrast to the finding about consistent endorsement described earlier, responses for in-service teachers were higher than the mid-point for retrieval, while this was not true of spacing.

Table 2 *Responses to each concept by trainee and in-service teachers*

| | Trainee teachers | | In-service teachers | |
|---|---|---|---|---|
| *Scenario focus* | *Mean (SD)* | Mean score vs. 4 (the neutral score) | *Mean (SD)* | Mean score vs. 4 (the neutral score) |
| Retrieval | 4.82(1.28) | $t(66) = 5.27$, $p < .001$, Cohen's $d = 0.91$ | 4.70 (1.35) | $t(32) = 2.97$, $p = 0.006$, Cohen's $d = 0.73$ |
| Spacing | 5.27 (1.01) | $t(66) = 10.29$, $p < .001$, Cohen's $d = 1.78$ | 4.45 (1.56) | $t(32) = 1.67$, $p = 0.104$, Cohen's $d = 0.41$ |
| Interleaving | 3.18 (0.98) | $t(66) = -6.83$, $p < .001$, Cohen's $d = 1.18$ | 2.93 (1.22) | $t(32) = -5.05$, $p < .001$, Cohen's $d = 1.24$ |

A comparison between trainees and in-service teachers was run next using the pooled imputed values; equality of variance was rejected because Levene's test was significant ($p = 0.031$) for the spacing data, though it was not significant for the other two concepts. A one-way ANOVA revealed a significant effect of teaching level on accuracy of beliefs about spacing [$F(1, 110) = 8.50$, $p = .004$], but no such effect for retrieval practice [$F(1, 109) = 0.40$, $p = 0.53$] or for interleaving [$F(1, 118) = 1.46$, $p = 0.22$].[2]

The imputed data from the three concepts were also combined to form an overall average accuracy score for each in-service teacher with a minimum possible score of 3 and a maximum of 21, and this will henceforth be referred to as a participant's *total score*. A 2-tailed Pearson's product moment correlation revealed that there was no significant positive relationship between experience and accuracy by these measures, with the trend in the opposite direction ($n = 44$; $r = -0.113$; $p = 0.475$).[3]

[2] For the original, non-imputed scores, too, the spacing scores were significant [$F(1, 110) = 8.50$, $p = .004$] and other two non-significant.

[3] This compares to -0.232, $p = 0.155$, with the original data sets ($n = 33$).

Next, confidence scores were analyzed. There were insufficient data to complete MI for the confidence scores, and therefore only in-service teacher participants who had completed this question were included here. In effect, this amounted to listwise deletion of participants who did not finish the survey. Confidence was not normally distributed according to a Shapiro-Wilk test, and therefore bootstrapping was applied (see Efron & Tibshirani, 1993; Field, 2018) using the bias corrected and accelerated method, and using 1000 values. The analysis revealed that these two variables (accuracy and confidence) were not significantly correlated, with a weakly positive trend ($r = 0.11$; $p[33] = 0.540$). It should be noted, however, that according to a sensitivity analysis carried out by G-Power, the current sample size was too small to reliably detect correlations smaller than 0.44, and this finding should therefore be treated with caution.

Finally, after responding to the vignettes, all participants (trainee and in-service teachers) were asked to respond to characterize their background knowledge. 62.6% of trainees responded with 'quite well' or 'very well' to the three concepts overall compared to 35.3% of in-service teachers. Taking into account the

ordinal nature of the data, an independent samples Mann-Whitney U test was used to compare levels of prior knowledge with each of the response options scored from 1–4, and this confirmed an overall difference between the two categories of participants ($p$ = 0.01, $df$ = 98; $z$ = -3.39).

Interestingly, despite 51 out of 67 of the trainee teachers (76.1%) declaring that they understood the concept of interleaving at least 'quite well', only 3% were classified as consistent endorsers in the analysis described earlier. This may fit with an interpretation that trainees had encountered interleaving and other evidence-based strategies in their academic reading but were inexperienced at recognizing them in classroom scenarios.

## Discussion

This study has shown areas of mis-match between teacher beliefs and recommended evidence-based teaching practices, and provided evidence that contrary to what might be expected, in-service teachers neither perform better than trainee teachers overall, nor does their alignment with the evidence appear to improve in line with experience. Such findings have implications for the judgements teachers make in the classroom.

In terms of research question 1 (the accuracy of responses to vignettes about spacing, interleaving and retrieval practice), the way that both trainee teachers and practicing teachers interpret memory-relevant scenarios is not always well in line with the research consensus. However, there is some nuance to this. The results are more accurate than those found in the survey of university students by McCabe (2011) and in the work by Halamish (2018); in addition, just under 50% of the current sample of teachers endorsed spacing consistently (on all three scenarios that presented a spacing vs. massing choice), and the spacing and retrieval scenarios were in general better answered than those for interleaving. With exactly 30% endorsing retrieval practice consistently, and just 4% for interleaving, it seems possible that the idea of using 'tests' (retrieval) and mixing (interleaving) examples are seen as poor practice, perhaps due to negative associations with both terms. However, it should be noted that the in-service teachers' responses to the retrieval scenarios were significantly above the mean while those for spacing were not, perhaps suggesting that endorsement of spacing is often half-hearted, and that a sub-set of practicing teachers greatly value the use of quizzes.

Subsequent to the design and data gathering used in this study, I became aware of another survey which had also used vignettes about evidence-based learning strategies as well as neuromyths. Findings from around 200 American educators are included in a report from this work published by Boser (2019) on the website of *The Learning Agency*, and shows a rate of correct response at 31% for retrieval practice, approximately 60% for spacing, and 20% for interleaving. This supports three main conclusions that I have expressed so far: that the teachers' views were not strongly in line with the research evidence, that the teachers nevertheless scored higher on these matters than has been found in surveys of students and in certain previous findings, and that endorsement appears to be lowest in the case of interleaving.

A reasonable question could be raised over whether the scenarios presented are sufficiently in line with the research evidence such that the optimal answer is actually correct. It should therefore be noted that although they were largely novel, the vignettes did draw heavily on research evidence. Scenario 9, for example, was based around Birnbaum, Kornell, Bjork, and Bjork's (2013) study of interleaved butterfly images, and scenario 4 was

based on Zulkiply, McLean, Burt, and Bath's (2012) interleaved psychological case studies. While the only way to be certain whether the desirable difficulties would actually lead to superior outcomes would be to run empirical studies testing each one on school pupils, future studies could engage a team of experts to review and comment on the materials, as was done by Simons and Chabris (2011, 2012). In future studies, it would also be preferable to randomize the presentation order of scenarios.

Regarding the difference between trainee teachers and in-service teachers – research question 2 – these groups differed in how they responded. However, rather than their providing a baseline, it was trainees who performed better in terms of accuracy (at least with respect to items on spacing). Trainees also reported more awareness of the techniques under investigation.

This is surprising as the trainees were at a very early stage of their careers; most[4] had no formal classroom experience, and moreover, they had had little time to take on board the academic lessons of their course. In addition, prior research by Surma, Vanhoyweghen, Camp, and Kirschner (2018) has suggested that retrieval practice and spacing – the two concepts on which trainees answered more accurately – tend to be absent from teaching textbooks (and it can be assumed that interleaving is, too). It is entirely possible, however, that the trainee teachers had read about these concepts online or heard about them from tutors – any replication should seek to sample participants from more than one teacher education institution in order to analyze this possibility. Prior knowledge did not always relate to superior performance on the vignettes; declared familiarity with interleaving did not correspond to consistent endorsement of this technique, however it appears that trainee teachers' prior learning of spacing and retrieval practice was thorough enough for them to be able to apply their knowledge to vignettes despite their limited classroom experience.

This finding leads to two considerations for the future. One is that a replication should be attempted with a sample of individuals who have no experience of or interest in teaching at all, for example adults in other professions, matched according to past educational level. This would provide a baseline which is unbiased by academic reading or preparation prior to beginning as a trainee. Secondly, if classroom practice is to be optimal in our schools, more needs to be done to ensure that practicing teachers understand evidence-based techniques and are able to make use of them.

In terms of research question 3, it appears that accuracy in memory-related judgements does not improve in line with years of experience in the classroom. There was no evidence of a positive relationship between years of classroom experience and the accuracy of answers, with a non-significant trend in the other direction. The trainee teacher data also lend some support to this; it might be assumed that on the basis of a combination of both training and experience, practicing teachers would have a great advantage in the task, but trainee teachers at times outperformed their more experienced counterparts as discussed above. It may be interesting to note that the two (tied) highest overall scores (16.67) came from a practicing teacher with 3 years of experience and from a trainee teacher. The lowest (5.00) was from a teacher with 23 years of experience.

However, as with any correlational findings, the results have to be interpreted with caution, and particularly given the low sam-

[4] It is possible that some had shadowing experience prior to the PGDE course, or had perhaps worked in classrooms where a formal teaching diploma is not required, e.g. as a language assistant abroad.

ple size for in-service teachers they must be treated as indicative only. The more experienced teachers completed their education and/or training at an earlier point in time, and as with any inter-generational comparison, age effects have to be considered. For example, older teachers may have been using more conservative heuristics to judge the scenarios.

In terms of confidence in one's own choices (research question 4), past metacognitive research suggests that confidence is a poor indicator of accuracy, both in terms of judgements of learning and more broadly (e.g., in legal contexts; Loftus, 2019). In line with this, the correlation between performance on the task and declared confidence was low and non-significant. While again some caution is needed, the findings here seem to fit with the broader literature which suggests that when it comes to desirable difficulties, people have flawed metacognition – they do not know what they do not know. Teachers may be making flawed classroom choices and yet feel confident in those choices.

## Conclusion

The findings of this study suggest that teachers' judgements are often flawed with respect to certain memory processes. While teachers' responses to scenarios that relate to desirable difficulties were more accurate than those in previous studies, they were nevertheless out of line with the research consensus on desirable difficulties, in particular when it comes to interleaving. This finding extends research in a little-studied but practically important domain, fits with the broader metacognitive evidence that memory is counterintuitive, and revealed new evidence that years of classroom experience did not correlate with the accuracy of teachers' responses to desirable difficulties scenarios.

Although there is as yet a lack of research investigating the specific effects (if any) of such beliefs on classroom practice and thereby their potential impact on student attainment (and the effect of such beliefs is therefore just theoretical at present), there is evidence from research in students' independent study that connects flawed beliefs with poor outcomes (e.g., Hartwig & Dunlosky, 2012). The present research will help pave the way for such research to be conducted in classroom teaching contexts as well.

Overall, the current study has provided a clear message when it comes to beliefs about memory – teachers' beliefs are not always accurate, and the solution to this problem does not lie in experience alone.

**Acknowledgement**

The author would like to acknowledge the valuable comments of colleagues Ian Rivers and James Boyle on an earlier draft of this article.

**Author's ORCID**
Jonathan Firth
https://orcid.org/0000-0003-1213-0219

## References

Birnbaum, M. S., Kornell, N., Bjork, E. L., & Bjork, R. A. (2013). Why interleaving enhances inductive learning: The roles of discrimination and retrieval. *Memory & Cognition, 41*(3), 392–402. https://doi.org/10.3758/s13421-012-0272-7

Boser, U. (2019). *What do teachers know about the science of learning? A survey of educators on how students learn.* The Learning Agency. https://www.the-learning-agency.com/insights/what-do-teachers-know-about-the-science-of-learning

Brunmair, M., & Richter, T. (2019). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin, 145*, 1029–1052. https://doi.org/10.1037/bul0000209

Čavojová, V., & Jurkovič, M. (2017). Comparison of experienced vs. novice teachers in cognitive reflection and rationality. *Studia Psychologica*, *59*(2), 100–112. https://doi.org/10.21909/sp.2017.02.733

Dunlosky, J., & Rawson, K. A. (2015). Practice tests, spaced practice, and successive relearning: Tips for classroom use and for guiding students' learning. *Scholarship of Teaching and Learning in Psychology, 1*(1), 72–78. https://doi.org/10.1037/stl0000024

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

Ericsson, K. A. (2017). Expertise and individual differences: The search for the structure and acquisition of experts' superior performance. *WIREs Cognitive Science, 8*(1–2), e1382. https://doi.org/10.1002/wcs.1382

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

Firth, J. (2018). Teachers' beliefs about memory: What are the implications for in-service teacher education? *Psychology of Education Review, 42*(2), 15–22.

Furnham, A. (2018). Myths and misconceptions in developmental and neuro-psychology. *Psychology, 9*(02), 249–259. doi: 10.4236/psych.2018.92016

Glogger-Frey, I., Deutscher, M., & Renkl, A. (2018). Student teachers' prior knowledge as prerequisite to learn how to assess pupils' learning strategies. *Teaching and Teacher Education*, *76*, 227–241. https://doi.org/10.1016/j.tate.2018.01.012

Halamish, V. (2018). Pre-service and in-service teachers' metacognitive knowledge of learning strategies. *Frontiers in Psychology*, *9*, 2152. https://doi.org/10.3389/fpsyg.2018.02152

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement?. *Psychonomic Bulletin & Review, 19*(1), 126–134. https://doi.org/10.3758/s13423-011-0181-y

Hood, M. (2016). *Beyond the plateau: The case for an institute for advanced teaching.* Institute for Public Policy Research. https://www.ippr.org/files/publications/pdf/beyond-the-plateau_July2016.pdf

Hunter, A. S., & Lloyd, M. E. (2018). Faculty discuss study strategies, but not the best ones: A survey of suggested exam preparation techniques for difficult courses across disciplines. *Scholarship of Teaching and Learning in Psychology, 4*(2), 105–114. https://doi.org/10.1037/stl0000107

Kennedy, M. M. (2008). Sorting out teacher quality. *Phi Delta Kappan*, *90*(1), 59–63. https://doi.org/10.1177/003172170809000115

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, *14*(2), 219–224. https://doi.org/10.3758/BF03194055

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory*, *17*(5), 493–501. https://doi.org/10.1080/09658210902832915

Loftus, E. F. (2019). Eyewitness testimony. *Applied Cognitive Psychology, 33*(4), 498–503. https://doi.org/10.1002/acp.3542

McCabe, J. (2011). Metacognitive awareness of learning strategies in undergraduates. *Memory & Cognition, 39*, 462–476. https://doi.org/10.3758/s13421-010-0035-2

McCabe, J. A. (2018). What learning strategies do academic support centers recommend to undergraduates? *Journal of Applied Research in Memory and Cognition, 7*, 143–153. https://doi.org/10.1016/j.jarmac.2017.10.002

Magnussen, S., Wise, R. A., Raja, A. Q., Safer, M. A., Pawlenko, N., & Stridbeck, U. (2008). What judges know about eyewitness testimony: A comparison of Norwegian and US judges. *Psychology, Crime & Law, 14*, 177–188. https://doi.org/10.1080/10683160701580099

Melinder, A., & Magnussen, S. (2015). Psychologists and psychiatrists serving as expert witnesses in court: What do they know about eyewitness memory? *Psychology, Crime & Law, 21*(1), 53–61. https://doi.org/10.1080/1068316X.2014.915324

Morehead, K., Rhodes, M. G., & DeLozier, S. (2016). Instructor and student knowledge of study strategies. *Memory, 24*(2), 257–271. https://doi.org/10.1080/09658211.2014.1001992

Nolan, A., & Molla, T. (2017). Teacher confidence and professional capital. *Teaching and Teacher Education, 62*, 10–18. https://doi.org/10.1016/j.tate.2016.11.004

Pan, S. C., & Bjork, R. A. (2020). Acquiring an accurate mental model of learning: Towards an owner's manual. In A. Wagner & M. Kahana (Eds.), *Oxford handbook of learning & memory: Foun-*

*dations and applications*. Oxford: Oxford University Press.

Piza, F., Kesselheim, J. C., Perzhinsky, J., Drowos, J., Gillis, R., Moscovici, K., ... & Gooding, H. (2019). Awareness and usage of evidence-based learning strategies among health professions students and faculty. *Medical Teacher*, *41*(12), 1411–1418. https://doi.org/10.1080/0142159X.2019.1645950

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*, 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Simons, D. J., & Chabris, C. F. (2011). What people believe about how memory works: A representative survey of the US population. *PloS ONE 6*(8), e22757. https://doi.org/10.1371/journal.pone.0022757

Simons, D. J., & Chabris, C. F. (2012). Common (mis) beliefs about memory: A replication and comparison of telephone and Mechanical Turk survey methods. *PloS ONE, 7*(12), e51876. https://doi.org/10.1371/journal.pone.0051876

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science, 10*(2), 176–199. https://doi.org/10.1177/1745691615569000

Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology, 44*(1), 24–31. https://doi.org/10.1177/0098628316677643

Surma, T., Vanhoyweghen, K., Camp, G., & Kirschner, P. A. (2018). The coverage of distributed practice and retrieval practice in Flemish and Dutch teacher education textbooks. *Teaching and Teacher Education, 74*, 229–237. https://doi.org/10.1016/j.tate.2018.05.007

Wiliam, D. (2010, March). *Teacher quality: Why it matters, and how to get more of it.* The Spectator. http://www.vcsta.org/wp-content/uploads/2013/03/Spectator-talk.pdf

Yan, V. X., Bjork, E. L., & Bjork, R. A. (2016). On the difficulty of mending metacognitive illusions: A priori theories, fluency effects, and misattributions of the interleaving benefit. *Journal of Experimental Psychology: General*, 145(7), 918–933. https://doi.org/10.1037/xge0000177

Zechmeister, E. B., & Shaughnessy, J. J. (1980). When you know that you know and when you think that you know but you don't. *Bulletin of the Psychonomic Society, 15*, 41–44. https://doi.org/10.3758/BF03329756

Zulkiply, N., McLean, J., Burt, J. S., & Bath, D. (2012). Spacing and induction: Application to exemplars presented as auditory and visual text. *Learning and Instruction, 22*(3), 215–221. https://doi.org/10.1016/j.learninstruc.2011.11.002

## Appendix 1: Vignettes used

*Source: items 2 & 3 based on McCabe (2011, 2018); items 1 and 4-9 are novel. Note that vignettes are presented here in the order that they appeared to participants; labels (e.g., 'interleaving 1') can be used to match each vignette to the relevant data in Chapter 6, but these were not seen by participants.*

Instruction: For each of the scenarios, please decide for yourself which answer you think is most likely, and then circle or highlight your response on the sheet. There are 9 scenarios.

Scenario 1 ('interleaving 1')

A senior Geography class is learning about lakes. Their teacher divides them into two teams. Team A looks at pictures of different types of lakes all mixed together (MIXED), each with a picture of the lake, its name, and a label saying what type of lake it is. Team B see the same information categorized by type (that is, GROUPED), so that they see all examples of one type of lake together, then the next type, and so on. Finally, the two teams are given the same test.

Please provide a prediction on the 7-point scale in reference to typical pupils.

i. Team A (MIXED) would gain much higher test scores.
ii. Team A (MIXED) would gain moderately higher test scores.
iii. Team A (MIXED) would gain slightly higher test scores.
iv. Test scores for teams A and B would be about EQUAL.
v. Team B (GROUPED) would gain slightly higher test scores.
vi. Team B (GROUPED) would gain moderately higher test scores.
vii. Team B (GROUPED) would gain much higher test scores.

Scenario 2 ('spacing 1')

Two educational psychologists are working with a group of P5 pupils to help them with math. They have identified 10 target skills, and they have ten weekly half-hour sessions in which the pupils can practice these. In scenario A, the pupils spend an entire session focusing on just one skill, and then move on to the next skill the following week, and so on (that is, study is INTENSIVE). In scenario B, pupils look at a larger number of skills more briefly during each study session, and then return to these for further practice over the next few weeks (that is, study is SPREAD OUT). Pupils in both classes spend the same overall amount of time studying the skills. After the sessions are over, a math test on the skills studied is given to pupils from both classes.

Provide a prediction on the following 7-point scale in reference to typical pupils.

i. Scenario A (INTENSIVE) will result in much higher test scores
ii. Scenario A (INTENSIVE) will result in moderately higher test scores
iii. Scenario A (INTENSIVE) will result in slightly higher test scores
iv. Test scores for Scenarios A and B will be about EQUAL
v. Scenario B (SPREAD OUT) will result in slightly higher test scores
vi. Scenario B (SPREAD OUT) will result in moderately higher test scores
vii. Scenario B (SPREAD OUT) will result in much higher test scores

Scenario 3 ('retrieval 1')

In two different secondary classes, a 275-word prose passage about a specific topic is presented. In Lesson A, students first study the passage for 5 minutes, and then are asked to write down from memory as much of the material from the passage as they can for a further 5-minute period (they take a TEST). In Lesson B, learners first study the passage for 5 minutes, and then are asked to study the passage again for another 5 minutes (they RESTUDY). After 1 week, all students are asked to recall as much of the passage as they can remember.

Provide a prediction on the following 7-point scale in reference to typical secondary pupils:

i. Lesson A (TEST) will result in much higher test scores
ii. Lesson A (TEST) will result in moderately higher test scores

iii. Lesson A (TEST) will result in slightly higher test scores
iv. Test scores for Lessons A and B will be about EQUAL
v. Lesson B (RESTUDY) will result in slightly higher test scores
vi. Lesson B (RESTUDY) will result in moderately higher test scores
vii. Lesson B (RESTUDY) will result in much higher test scores

Scenario 4 ('interleaving 2')

Two P7 classes are learning about mental health. Their teacher has prepared examples of teenagers who suffer from three key types of mental health problems. The school pupils are presented with these examples, together with suggested solutions. In class A, pupils look at examples of the same type of mental health problem consecutively (i.e., GROUPED). In class B, pupils see the examples of the three types in an intermingled fashion (that is, MIXED), such that an example of one type is followed by an example of a different type, until all examples have been presented. After viewing all the examples, the learners are given a test with a selection of novel (previously unseen) case studies of individuals with mental health problems, and they are asked to identify suitable solutions.

Provide a prediction on the 7-point scale in reference to typical pupils.

i. Class A (GROUPED) will do much better on the test
ii. Class A (GROUPED) will do moderately better on the test
iii. Class A (GROUPED) will do slightly better on the test
iv. Test performance for Classes A and B will be about EQUAL
v. Class B (MIXED) will do slightly better on the test
vi. Class B (MIXED) will do moderately better on the test
vii. Class B (MIXED) will do much better on the test

Scenario 5 ('retrieval 2')

Two schools are running revision group ahead of end-of-year exams. In School X, learners spend their study periods reading over lesson notes, and looking at lesson slides (they RESTUDY). In School Y, learners spend each of their study periods testing themselves using flashcards (they take a TEST). A few weeks later, all pupils from both schools sit an identical exam during which they have to remember and apply the information, and they gain a percentage mark.

Please provide a prediction on the following 7-point scale in reference to typical secondary pupils:

i. School X (RESTUDY) will obtain much higher percentage exam results.
ii. School X (RESTUDY) will obtain moderately higher percentage exam results.
iii. School X (RESTUDY) will obtain slightly higher percentage exam results.
iv. Exam results for Schools X and Y will be about EQUAL
v. School Y (TEST) will obtain slightly higher percentage exam results.

vi. School Y (TEST) will obtain moderately higher percentage exam results.
vii. School Y (TEST) will obtain much higher percentage exam results.

Scenario 6 ('spacing 2')

Two secondary school deputy headteachers are planning the S3-S4 curriculum. The deputy in one school, Alpha High, plans the topics such that they are distributed across the year, with topics being partially covered and then returned to at a later date (learning is SPREAD OUT). The deputy in another school Beta High, plans the topics such that each topic is covered in full within a couple of weeks, and pupils then move on to a different topic (learning is INTENSIVE). The same overall amount of lesson time is spent on the topics in both schools. Pupils are then given an end-of-year test which covers all of the topics.

Provide a prediction on the following 7-point scale in reference to typical pupils, assuming that the pupils are generally similar in every other respect.

i. Pupils at Alpha High (SPREAD OUT) will gain much higher end-of year test scores.
ii. Pupils at Alpha High (SPREAD OUT) will gain moderately higher end-of year test scores.
iii. Pupils at Alpha High (SPREAD OUT) will gain slightly higher end-of year test scores.
iv. End-of-year test scores for Alpha High and Beta High A will be about EQUAL
v. Pupils at Beta High (INTENSIVE) will gain slightly higher end-of year test scores.
vi. Pupils at Beta High (INTENSIVE) will gain moderately higher end-of year test scores.
vii. Pupils at Beta High (INTENSIVE) will gain much higher end-of year test scores.

Scenario 7 ('spacing 3')

Two computer science classes are learning coding skills. In one class, Class A, the teacher presents new coding processes, and these are then practiced several times within the same lesson (that is, the learning is INTENSIVE). In the other class, Class B, the same coding processes are practiced across several lessons, only once per lesson (that is, the learning is SPREAD OUT). The same overall time is spent on the terms by both classes. At the end of the topic, both classes are given the same test.

Please provide a prediction on the following 7-point scale in reference to typical pupils:

i. Class A (INTENSIVE) will gain much higher test scores.
ii. Class A (INTENSIVE) will gain moderately higher test scores.
iii. Class A (INTENSIVE) will gain slightly higher test scores.
iv. Test scores for classes A and B would be about EQUAL.
v. Class B (SPREAD OUT) will gain slightly higher test scores.
vi. Class B (SPREAD OUT) will gain moderately higher test scores.
vii. Class B (SPREAD OUT) will gain much higher test scores.

Scenario 8 (‘retrieval 3’)

Two similar classes of pupils are learning terminology for their latest topic. In one class, Mrs. Smith shows the pupils the terminology one term per slide on a PowerPoint, and then shows the same PowerPoint two more times in follow-up lessons (that is, they RESTUDY). Mrs. Jones shows the pupils the terminology one term per slide on a PowerPoint, and then tests them on the items two times in follow-up lessons (that is, they take a TEST). A couple of weeks later, both classes are given a multiple-choice quiz on the terminology.

Please provide a prediction on the following 7-point scale in reference to typical secondary pupils:

i. Mrs Jones’s class (TEST) will gain much better scores on the quiz.
ii. Mrs Jones’s class (TEST) will gain moderately better scores on the quiz.
iii. Mrs Jones’s class (TEST) will gain slightly better scores on the quiz.
iv. Test scores for both classes will be about EQUAL
v. Mrs Smith’s class (RESTUDY) will gain slightly better scores on the quiz.
vi. Mrs Smith’s class (RESTUDY) will gain moderately better scores on the quiz.
vii. Mrs. Smith’s class (RESTUDY) will gain much better scores on the quiz.

Scenario 9 (‘interleaving 3’)

A visiting biologist presents pictures of butterflies to 11-year-old pupils in two schools. She shows the children four examples each of 16 species of butterfly. In School A, she shows all four examples of a single species consecutively and then moves on to examples of the next species, and so on, until all pictures have been presented (the images are GROUPED by species). In School B she presents the various species in an intermingled fashion, such that an example of one species is followed by an example of a different species, until all pictures have been presented (the images are MIXED). After viewing all the pictures, children are given a test that requires them to correctly identify previously presented pictures of the butterflies.

Please provide a prediction on the following 7-point scale in reference to typical pupils.

i. Pupils at School A (GROUPED) will get much higher test scores.
ii. Pupils at School A (GROUPED) will get moderately higher test scores.
iii. Pupils at School A (GROUPED) will get slightly higher test scores.
iv. Test scores for Schools A and B will be about EQUAL
v. Pupils at School B (MIXED) will get slightly higher test scores
vi. Pupils at School B (MIXED) will get moderately higher test scores
vii. Pupils at School B (MIXED) will get much higher test scores