Final Report

"Impact assessment of the publication of questions of theoretical examinations for Part 66 and Part FCL "
for the EUROPEAN AVIATION SAFETY AGENCY

Research Contract EASA.2008.C52

Submitted by:	Keiko Moebus Project Manager
Date:	7 August 2009 Version 2.1
Company:	MOEBUS Aviation Consulting
Address:	P.O. Box 8058 Zurich Airport SWITZERLAND
Phone: Mobile: Email:	+41 44 586 90 57 +41 79 552 50 97 keiko@moebus-aviation.ch

Document Identification Sheet

DOCUMENT DESCRIPTION	
Document Title End of Phase 4 "Project Wrap-up" Report "Impact assessment of the publication of questions of theoretical examinations for Part 66 and Part FCL "	
Project Reference Number	EASA.2008.C52
Edition Date	7 August 2009

Abstract

The cognitive processes involved in human learning and memory are discussed with particular emphasis on the differences between rote-learning and meaningful learning. An experiment was conducted to assess the risk of affecting the results of the flight crew theoretical examinations by the publication of the CQB. The results of this study indicate that it is difficult to pass a multiple choice examination using rote-learning if the number of questions in the database is high and if the examination period is short enough. The risks of publishing the questions and answers of the theoretical multiple choice examinations as well as proposals on how to mitigate such risks are discussed.

Keywords

CQB	JAR Part-FCL	EASA Part 66
Rote-learning	Deep-learning	Memory
Examination	Learning Strategies	Assessment
MCQ	Risk	Mitigation

Authors

Keiko Moebus	MOEBUS Aviation Consulting	keiko@moebus-aviation.ch
Philipp Moebus	MOEBUS Aviation Consulting	philipp@moebus-aviation.ch

Prof. Dr. Adrian Schwaninger	Center for Adaptive Security Research and Applications (CASRA) APSS, University of Applied Sciences Northwestern Switzerland and University of Zurich	adrian.schwaninger@fhnw.ch
Dr. Saskia Koller	Center for Adaptive Security Research and Applications (CASRA) APSS, University of Applied Sciences Northwestern Switzerland and University of Zurich	saskia.koller@fhnw.ch
Sandra Iglesias	Center for Adaptive Security Research and Applications (CASRA) APSS, University of Applied Sciences Northwestern Switzerland and University of Zurich	sandra.iglesias@casra.ch
Alan Wales	Center for Adaptive Security Research and Applications (CASRA) APSS, University of Applied Sciences Northwestern Switzerland and University of Zurich	alan.wales@casra.ch

Document Control

Edition	Date	Status
0.1	25 June 2009	First Draft
		Internal review
1.0	29 June 2009	Version 1.0
		Received EASA comments and updated
2.0	08 July 2009	Version 2.0
		Received EASA comments and updated
2.1	7 August 2009	Version 2.1

MOEBUS Aviation Consulting

P.O. Box

8058 Zurich Airport

Phone: +41 44 586 90 57
Email: info@moebus-aviation.ch
MOEBUS AVIATION Web Page: www.moebus-aviation.ch

TABLE OF CONTENTS

1. Executive Summary

The team of MOEBUS Aviation Consulting (Moebus) and the scientists from the Center for Adaptive Security Research and Applications (CASRA) were given a project; 1) to assess the risk of affecting the results of the flight crew theoretical examinations by using the European Aviation Safety Agency's (EASA) rulemaking process, in particular the publication of the Central Question Bank (CQB), and 2) to make proposals so as to mitigate this risk. The Agency asked to mainly assess the risk of students having good results out of mere rote-learning and such assessment should take into account the number of questions available, the examination procedures (e.g. number of attempts, number of sittings allowed etc.) and the learning processes. At the end of the project, possible mitigation measures should then be proposed.

The 5-month project was broken into four phases with specific goals: 1) Identify the cognitive processes involved in rote-learning and assess how it would apply to a student pilot in theoretical training, 2) Calculate the probability for a student pilot to learn by heart all the questions available for each basic examination involved in the study, together with the correct answer, 3) Calculate the probability for such a student to pass the corresponding examination; this includes assessing how the applicable examination procedures may influence the result and 4) Deduct the risk of affecting the results of the examinations and make proposals to mitigate this risk.

The literature shows that the cognitive processes involved in rote learning are not conducive to good retention of memory. Rote learning also decreases the ability of students to apply what they have learned to similar problems or even problems that are related but outside the scope of their studies. Students in theoretical training may be tempted to learn questions by rote given the huge amount of information to be retained across the 13 learning categories. Research by the FAA shows that students will go as far as to learn mnemonic devices or 'flashcard' training to create simplified methods of remembering. Several studies suggest that the relationship between the number of questions and performance is non-linear, and likely to be decaying power function. Thus, the number of questions to be published as well as whether they are tested all at once or in several exams with several months in between are key factors which will determine whether students will apply a rote learning strategy or not.

To explore the important questions empirically, we have conducted an experiment in which a sample of twenty university students attempted to rote-learn a battery of questions provided by a flight training school. A custom-designed learning management system was used to facilitate in student learning and data collection and the subjects were then asked to learn 136 questions and answers from the 050 Meteorology module in one day. The following day the students undertook the computer-based examination which contains the 136 exact questions and 136 reworded questions presented in random order. A week later, the subjects then learned a new set of 136 questions and undertook the final 544-question exam on the following day.
Overall, the results showed that typical university students were capable of exceeding the pass-mark of 75% in each of the conditions in our experiment. One day of concentrated study was enough to achieve these high scores, where the students memorized a total of 272 questions by rote. The results were marginally worse for the reworded questions, but the average still exceeded 75%. Rather impressively, seven subjects managed an overall score above 90% for the full 544 question battery off just two days of study.
This experiment gives an insight into how well the CQB can be rote-learned if the questions and answers are made available prior to testing, enough time is allocated and the amount of items to remember is reasonable. However, there will be a difference in the probabilities to rote memorize questions depending on the examination format and the time between examinations.
The size of the dataset to be memorized is a determining factor as to whether an individual would either attempt memorization, or whether it is even within the realms of possibility. As the data set size increases, the potential payoff for memorization arguably decreases.

To estimate the probability that students can in fact rote-learn a battery of up to 10,000 CQB questions and answers, projections have been made from the sample of students used in this experiment. Previous literature has suggested that the relationship between number of questions and performance is non-linear, and likely to be a decaying power function. The likelihood of memorization is different depending on how close together the examinations are spaced and whether there are test resits or not. If the full CQB is to be memorized because the examinations are close together, then the probability of passing the test is very low. However, we have estimated that if learning is spread across a sufficiently large amount of time then it may be possible to pass the examinations by rote by learning around two modules at a time.

Another interesting result was found by asking the students about their hypothetical learning strategy and if they could use text books as well as the questions and answers for which they would be tested. Students reported that they would mainly work with the multiple choice questions to orient their learning and use the textbooks just to consult some misunderstandings, relying basically on a rote-learning strategy to possess a higher probability to pass the exam. When rote learning is combined with textbook study or classroom study then the student is engaging with the material to understand it fully, and will result in better-retained and understanding of the syllabus than rote learning alone. The students used in the experiment would have liked to have read a book as well as rote learning to understand the material better. This qualitative finding showed an insight into the typical student mindset when faced with a MCQ examination rather than a written, essay-style examination. The learning strategies subtly change depending on what the demands for the examination are.

In terms of regulations and knowledge of procedures and essential flight statistics, both the literature review as well as the experimental study support the conclusion that some meaningful learning does occur with rote-learning. That the learning applied to reworded questions supports the idea that material is learned flexibly and lasts in memory up to a week, even when competing knowledge is introduced during that week. Rote learning of the CQB will result in memorization of vital flight information when presented in MCQ format.

Based on the above findings, we have listed 8 risk evaluation options to be assessed whether there is a potential risk of employing rote-learning as a strategy for flight students to score enough correct answers to pass the examination. Applied was the condition of the full CQB published, no changes in actual examinations implemented and non-standardised examination procedures as it is currently the case among the EASA member states.

Under the actual current examination practices, any straight recommendation either way is difficult to make. However, when applying mitigating measures such as standardisation of testing procedures as well as a limited time frame in which a student is to complete all subjects, the risk of regurgitation of information only is greatly reduced. Likewise, the CQB could be increased to such an extent at which memorization of the data available becomes futile and students are discouraged from rote-learning only.

To this extent, we recommend the agency to develop such EASA-wide standardised examination procedures such as, enrolment only through a certified flight training organization, and a very limited time frame in which the examination is to take place regardless of modular or integrated training. Further the CQB should be increased to at least double the volume and should be centralized so as to ensure that each EASA member state is sourcing the same test questions. These and other proposed standards should be released for consultation by the member states which is to be eventually implemented throughout all EASA member states prior to publishing the CQB. This may require some further evaluation of current national supervisory agencies procedures and their certification process in order to derive a harmonisation of EASA-wide standards fulfilling the mitigating requirements to publish the CQB.

2. Introduction

2.1 Background Information about the project

The EASA Basic Regulation 216/2008 requires applicants for a pilot license to demonstrate their level of theoretical knowledge. For this purpose, many EASA Member States organise theoretical examinations using forced-choice questions of a Central Question Bank (CQB) operated by the JAA. In order not to lose this heritage, EASA envisages keeping the question bank in its regulatory environment. Furthermore, it is envisaged to adopt a similar system for the examinations related to the issuance of the aircraft maintenance license (EASA Part 66).

The questions of the CQB are confidential. As the questions put in the database shall be used in the certification process, they have the status of certification specifications. The agreed rulemaking procedures that EASA is committed to follow require an open and public consultation on the content of a future rule, as well as its publication when finally adopted.

The Agency considers it necessary to assess whether the present rulemaking procedure could affect the current examination system by comparing examinations using confidential questions with examinations where the questions are known.

2.2 Scope of the project

The scope of the project given by EASA to MOEBUS Aviation and the scientists from CASRA was; 1) to assess the risk of affecting the results of the flight crew theoretical examinations by using the Agency's rulemaking process, in particular the publication of the CQB, and 2) to make proposals so as to mitigate this risk.

Furthermore, the Agency asked to mainly assess the risk of students having good results out of mere rote-learning and such assessment should take into account the number of questions available, the examination procedures (e.g. number of attempts, number of sittings allowed etc.) and the learning processes.

At the end of the project, possible mitigation measures should then be proposed.

The 5-month project was broken into four phases and each phase carried the specific goal to achieve as following:

Phase I: Identify the cognitive processes involved in rote-learning and assess how it would apply to a student pilot in theoretical training;

Phase II: Calculate the probability for a student pilot to learn by heart all the questions available for each basic examination involved in the study, together with the correct answer;

Phase III: Calculate the probability for such a student to pass the corresponding examination; this includes assessing how the applicable examination procedures may influence the result;

Phase IV: Deduct the risk of affecting the results of the examinations and make proposals to mitigate this risk.

2.3 Scope of the document

This report is organized in 8 chapters which were written by different teams. Chapters 1, 2, 7 and 8 were written by Moebus whereas chapters 3-6 were written by CASRA.

Chapter 1 contains the executive summary and Chapter 2 provides background information of the project.

Chapter 3 contains a literature review covering the diverse topics of cognitive processes involved in learning and memorization, and in particular, rote-learning processes. Based on that literature search, the CASRA team has then formed specific hypotheses to be asked in a scientific experiment. The data collected from the experiment was studied and statistically analysed, and later, we used those results and findings to generate a list of possible risks associated with publishing the Central Question Bank (CQB) to flight students in theoretical training and flight training institutions and eventual mitigation measures that the Agency may consider prior to the publication.

Chapter 4 describes the scientific experiment and data collection.

Chapter 5 summarises the results of the data collection and statistical analyses are presented.

The conclusions and summary of the scientific study findings are given in chapter 6.

Chapter 7 contains a summary written by Moebus on how they have prepared and reasoned to conduct a risk assessment and evaluation for the publication of CQB as well as the results of their risk analysis.

Chapter 8 concludes the risk analysis results and recommendations for future mitigation by Moebus.

2.4 Acronyms

ATPL	Airline Transport Pilot License
CASRA	Center for Adaptive Security Research and Applications
CPL	Commercial Pilot License
CQB	Central Question Bank
EASA	European Aviation Safety Agency
FCL	Flight Crew Licensing
JAA	Joint Aviation Authority
JAR	Joint Aviation Regulation
MCQ	Multiple Choice Question
PPL	Private Pilot License
NASA	National Aeronautics and Space Administration

3 Literature review

3.1 Introduction

The term "memory" describes the ability to recall or recognize information or events that have been previously learned or experienced (Ormrod, 2001). Memory refers not only to the process of retaining information for a period of time, but also to a mental representation of the new information and a mental location of where it is kept. Human memory is thought to consist of three central components: 1) storage, 2) encoding and 3) retrieval (Baddeley, 1999).

The first point, storage, refers to the process of putting new information into memory. When information is stored in memory, it is usually modified in some way; like simplifying the information, to bring it to a verbal form or associating it to previous knowledge. This act forms the second part of forming memories in a process known as encoding. This newly-formed information can now be used again by retrieving it from the memory store. Human memory can be divided into different memory systems, shown in Figure 1.

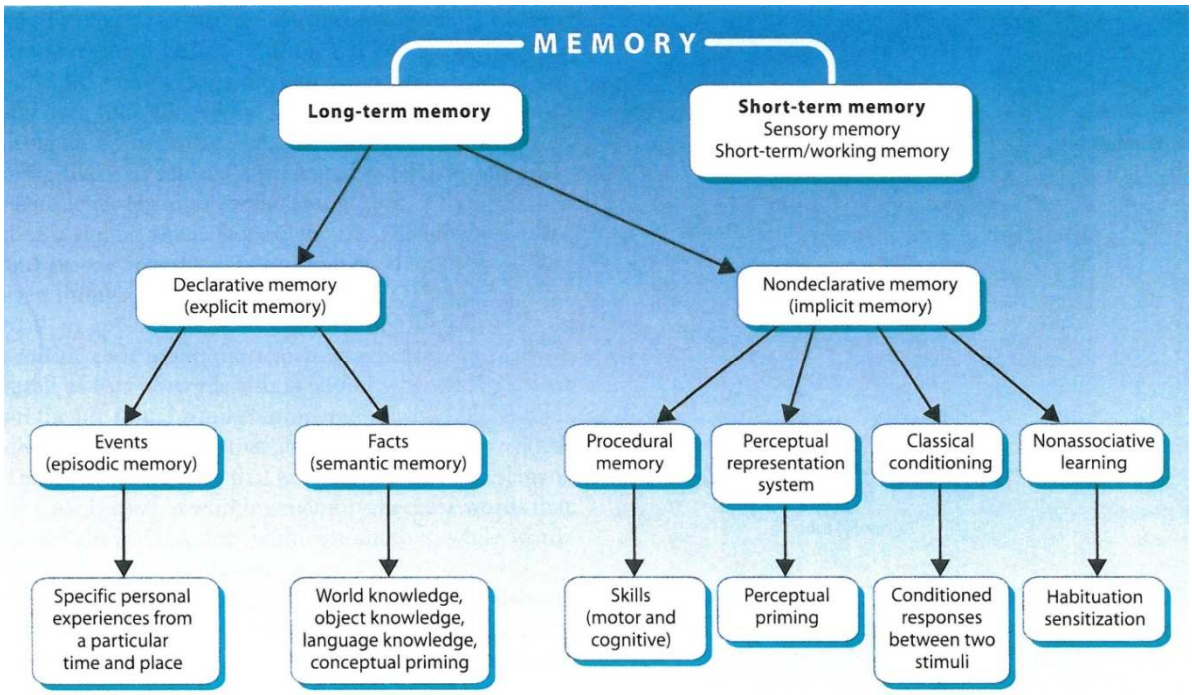

Figure 1: Structure of human memory.

The classical theory of memory (Baddeley, 1999; Ormrod, 2001; Gazzaniga, Ivry, & Mangun, 2002) makes the distinction between short-term memory (STM) and long-term memory (LTM) stores. This theory emphasizes the distinction between memory systems that retain information for different periods of time, and those that contain different kinds of memory codes or that have different limitations on the amount of information retained. STM or working memory, as formulated by Baddeley and Hitch (1974), represents a complex set of interactive subsystems, which have temporary capacities and allow for the holding and manipulation of limited amounts of information. LTM represents information that is stored for a durable period of time. Depending on the level of processing, this information persists in the memory store for several seconds up to a lifetime.

Studies have shown with aphasics and other memory-disorder patients that this distinction can be held as well on the neural level by demonstrating that, depending on the lesion in a certain brain area, an inability in either short or long-term memory is the consequence. Due to the nature of this upcoming study into the comparison of rote-learning vs. meaningful learning, we will mainly focus on levels of processing information related to long-term memory models.

3.2 Storing information in the long-term

Long-term memory is split into declarative and non-declarative memory, of which declarative memories are those facts and intellectually acquired knowledge that can be stated or shown as being accurate or inaccurate (Baddeley, 1999). In contrast, procedural memories are those memories which encompass movements or actions that are remembered and can be repeated at a later date. An example of declarative memories would be the knowledge of flight rules and regulations, or how to speak a foreign language. Procedural memory can be evidenced when operating a joystick within reasonable operating limits, or can be as simple as remembering how to ride a bicycle.

The differentiation between declarative and procedural memories depends on the task demands and previous exposures to the stimuli. Flight students may recall how to operate technical machinery from a workshop and physically interacting with the device, or they may have learned its operation by reading an operating manual. There is strong evidence that procedural memories are strongly stored in memory such that people never forget how to operate complex tasks like using a keyboard, despite the task being difficult in explanatory terms. However, declarative memories can be difficult to remember if a long period of time has expired between learning and recall (Baddeley, 1999).

Long-term memory storage is not simple to define in unambiguous terms; although some information may be stored easily, most information must be consciously and actively processed before it is stored. People store information in long-term memory most successfully when they have understood it, organized it, and integrated it with information they already have. Retrieval from long-term memory is closely tied to storage processes: the more completely information has been understood, the better it has been organized and the more closely it has been integrated with previously stored concepts.

This leads onto the application of memorization by learning, as it is clear that different learning strategies and encoding result in different performance capabilities. As we will show, learning by rote-memorization is not encoded in memory as strongly or is as useful to performance as other learning strategies.

3.3 Storing memories by learning

Learning is defined as a relatively permanent change in behaviour and in mental representations as a result of previous experiences (Ormrod, 2001). When retaining this information for a period of time, a memory is generated and this process is related to the ability to recall information that has previously been learned; but not everything that is learned is remembered permanently.

This is because long-term memories are not all formed equally, whereby the persistence of the memory and how easily recalled the memory is, tends to be a function of the second process of memorization – encoding. Craik and Lockhart (1972) are acknowledged as the first researchers to formulate a model of memory based on the level of cognitive processing, which was seen as a determinant of how that information would be encoded and stored in memory. They tested three conditions; the first condition was considered to be a superficial processing task as the students had to decide if a word was written in upper- or lowercase. The second condition was considered to be an intermediate level of processing as the students had to decide if two words rhymed with each other or not. The third condition involved semantic processing and consisted of making a judgment about the meaning of the word, which was considered a deep level of processing. In the test phase students showed better memory for words when they learned the stimuli more deeply than for those who had processed the stimuli more superficially in the learning phase. Many studies and experiments since then have confirmed that there are two basic forms of memory encoding: *surface learning (such as rote-learning) and deep learning ('meaningful' learning).*

Rote-learning is a learning technique that often involves little or no explicit understanding of the material that is learned (Ormrod, 2001). The material is memorized by repetition, which means that no transformation of the knowledge is required and the person simply tries to remember the material learned without context or a view to apply the knowledge to other situations. By contrast, **meaningful learning** refers to a learning technique in which the learning material is learned consciously and fully understood. Meaningful learning leads to knowledge acquisition that enables the application of this information to novel situations. An elaboration of the processes involved in meaningful learning is provided in the appendix.

3.4 Cognitive processes of rote-learning

In Western countries, memorization and rote-learning are typically used as synonyms commonly believed that they do not lead to deep understanding. Rote-learning is widely used in the mastery of foundational knowledge such as phonics in reading, the periodic table in chemistry, multiplication tables in mathematics, anatomy in medication, laws etc. Instead of viewing rote-learning as the opposite of understanding materials, it can be viewed as a complementary role. However, **rote-learning (surface learning) should be distinguished from memorization that involves understanding.**

- If students are repeating the material to be learned without understanding, then they are engaging in rote-learning, and this may not be followed by understanding.
- If students try to understand what they are repeating, memorization with understanding is taking place, such that the memorized information can be used as well in other circumstances. This is how Chinese traditional learning takes place; 'surface' learning is applied but for the purpose of a further 'deep' learning (Marton, Dall'Alba & Kun, 1996; Tweed & Lehman, 2002).

Therefore, it is important to differentiate between these two approaches of memorization; one that relies on regurgitation and does not give students' an application of knowledge and connectivity of information, and the other application of memorization when facts are rote-learned but connected to previously stored knowledge with the purpose of facilitating the recall and understanding of the facts in a long-term (Kember, 1996).

Memorization of information is critical for a flight student to master the wide variety of knowledge that is necessary to safely operate an airplane, but how that information is accrued is also important to the understanding involved when in the cockpit. Expertise can be accumulated by practice, but prior to practice the theory has to be demonstrated to be understood at a high level of competence. In rote-learning, on the other hand, new knowledge may be acquired simply by verbatim memorization and arbitrarily incorporated into a person's knowledge structure without interacting with what is already there, as previously existing knowledge might not be activated.

Under these circumstances if such information is stored in long-term memory at all, it is stored in relative isolation from other information. Information stored in this unconnected fashion becomes difficult to retrieve and to hold for a long period in memory. Because of this, much of the educational reform movement of the late 1950s and 1960s was an attempt to get away from rote-learning in schools by advancing instructional programs that encouraged discovery, or inquiry learning (Novak & Gowin, 1984).

Bruer (1994) argued that schools provide students with command of lower-level, rote skills, such as a computation in math, recalling facts in science, decoding works in reading, spelling, grammar, and punctuation in writing. He said that many students can remember facts to solve routine textbook problems, and apply formulas, but that "many if not most students have difficulty using what they know to interpret an experiment, comprehend a text, or persuade an audience. "They can't rise above the rote, factual level to think critically or creatively" (p. 147).

Students learn facts and propositions to pass exams, but the existing educational model does often not reward higher-order reasoning and learning skills. This issue is pervasive throughout the educational psychology literature, as the examinations students sit should reflect divergent cognitive abilities to succeed in applied settings, rather than regurgitate factoids from a system that, according to some educational psychologists, rewards 'cramming' of information.

There is no argument that many of the questions used for flight-school examinations require some level of factual knowledge, as a 2004 NASA technical document reported when analysing pilots' learning for the theoretical examination. "A new student pilot's understanding of flaps may be little more than a collection of memorized facts," (Casner *et al.*, 2004, p. 11) the researchers noted, which is the reality of learning for an examination that covers many diverse topics and requires hard-wired knowledge of certain factoids.

3.5 Retention and transfer of knowledge from learning

Retention is the ability to remember material at some later time in much the same way it was presented during the instruction. Sulman (1987) points out that many students at age 17 are memorizing basic math facts but have not been trained to apply them to problem-solving at school. This is not the same as understanding the basic principles of math and prevents transfer of knowledge into new problems that require flexibility of thought. Transfer, on the other hand, is the ability to use what was learned to solve new problems, answer new questions or facilitate learning new subject matters (Mayer & Wittrock, 1996).
It is easier to teach and test for retention of facts than to teach and assess objectives aimed at promoting transfer (Baxter, Elder, & Glaswer, 1996).

Transfer of knowledge is enhanced when abstract rules are coupled with examples, learners are confronted with similar problems, problem-solving is trained and when feedback on performance is available. By applying the appropriate instructions and assessment tasks educators can show how they can manipulate new information and in so doing promote retention and transfer of knowledge.

Rote-learning is sometimes considered to be 'cramming' because one who engages in rote-learning may give a misleading impression of having understood what they have written or said by memorizing an abundance of information into a short amount of time, thereby weakly committing the information to memory. This practice is strongly discouraged by many new curriculum standards but rote-learning is still widely practiced in schools in countries such as India, China, Japan, Russia, Turkey, Malta and Greece. This is typically because these nations tend to highly encourage students to earn high test scores in international comparisons with regards to other nations around the world. On the other hand, repetition of facts as a complement to meaningful learning is an important tool to store the new information in long-term memory.

3.6 The probability to rote-learn multiple choice questions

If students believe that the test will just require verbatim recall then they are more likely to engage in rote-learning. If students believe that they will have to connect and apply the material to be learned they are going to engage in meaningful learning, and as has been shown, rote-learning helps to store the information in long-term memory. This is the central differentiation between multiple-choice examinations and essay-style examination.

Brian *et al.* (1998) studied learning behavior of medical students at different time periods. In this study, 88% of the medical students answered previously learned multiple-choice questions (MCQ) correctly at the end of the learning condition, but only 35% answered the questions correctly based on nearly identical knowledge one day later when presented in a different format. Furthermore the learned MCQ presented five months later was more often answered correctly than the modified question presented one day after the problem-based learning. This study raises questions about the MCQ format when students can pass the examination, but are not able to apply this information to similar situations.

There is strong evidence in the literature that systematically "testing" oneself, as opposed to basic paired-associate learning, has benefits for the recall of questions and answers. In the domain of language learning, where native-foreign pairs are learned (i.e. known native word to matching foreign word), studies have shown that self-testing produces improvements in performance (Mozer, Howe & Pashler, 2004).

Figure 2 shows the difference between students who had used a self-test and study procedure, and those who had simply rote-memorized the question and answer pairs (Carpenter *et al*., 2007). It should be noted, that neither method improves general knowledge of an area, but the self-test procedure does improve memory encoding- even up to 40 days after the original set was learned.

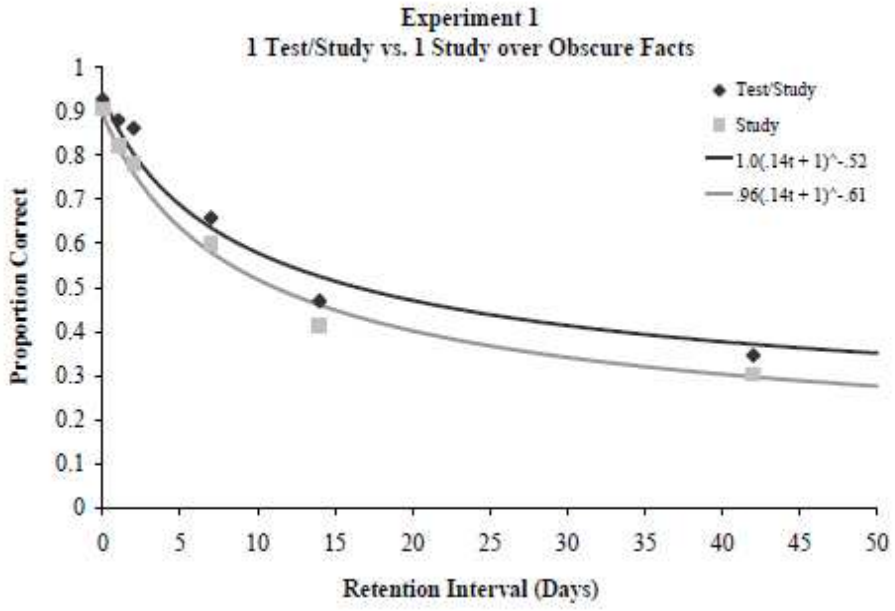

Figure 2: The difference between self-testing with study, and study alone

The more a student learns the subject material, the lower the probability of error becomes.

An example of this sort of relationship is shown in Figure 3, where blocks are sets of sixteen stimuli that have been repeatedly exposed, and the conditions (I-VI) represent the difficulty of the stimuli sets as determined by categorization difficulties with VI being the most complex. The steepness of the curve is determined by the set difficulty, although the decaying power relationship holds as a cumulative function of the amount of exposure to

the same set of stimuli. The more time a student spends learning paired-associate stimuli, the better they perform by rote-memorization.

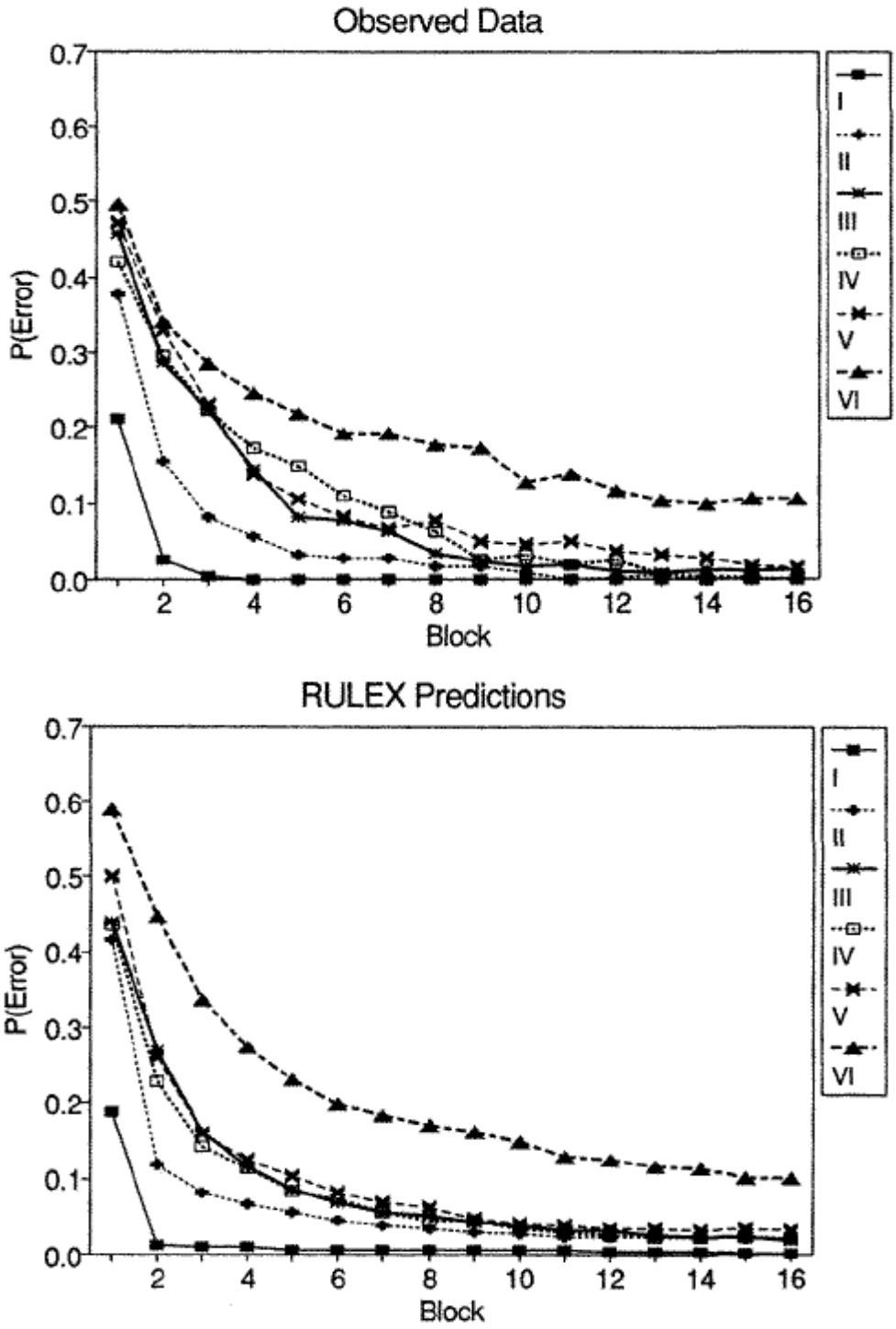

Figure 3: Predicted and actual results from a memory-retrieval paradigm.

When stimuli are given for a fixed amount of time (i.e. shown only once), retrieval of paired-associate learning is almost always a function of *forgetting*, which is accentuated when more questions have to be learned. As the training load increases, a sharp decline in capabilities are shown, producing a non-linear relationship between load and performance.

An example of this is the difference between trying to learn 100 or 1000 question and answer pairs. Performance will be lower for the larger question set if learning time is kept constant. The problem with rote-memorization, then, is that students may be able to

remember question-answer pairs and perform well on an examination, but recall very little of what was learned a year later. The rate of forgetting is typically very large in the first month, and then slower after that (Figure 4: Landauer, 1986).

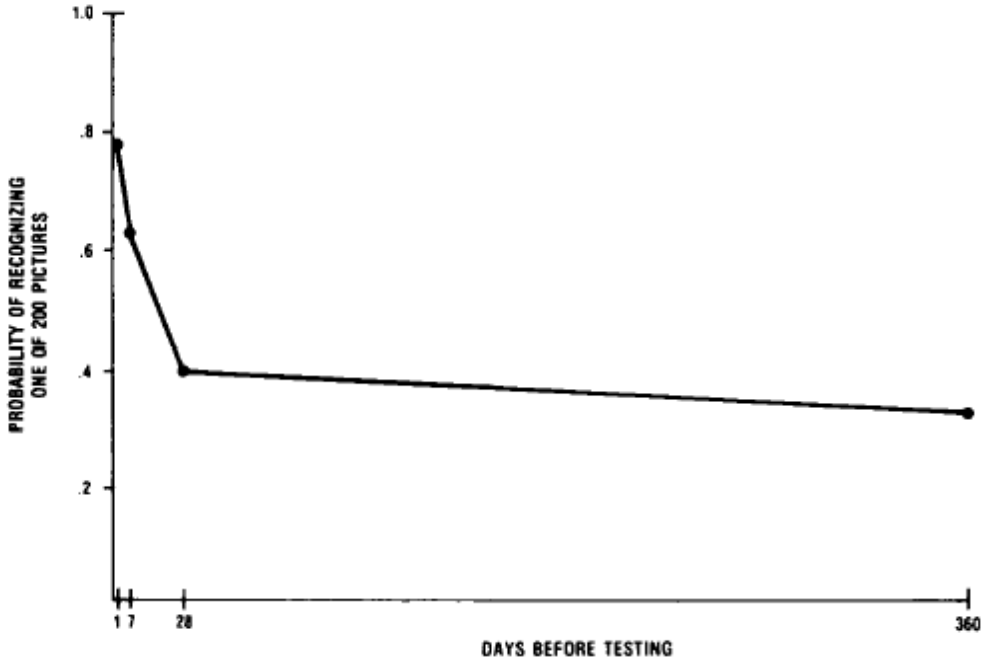

Figure 4: The rate of forgetting is largest for the first month

3.7 The NASA-rote-learning study

Since the mid 1980's into the early 21st century, the Federal Aviation Administration (FAA) of the United States made the question bank available to the public. Suspicions about the performances and implications of rote-learning became evident among flight examiners across the United States. The US National Aeronautics and Space Administration (NASA) commissioned a study (Casner, Jones, Puentes, Irani, 2003) to investigate the learning habits of the students who are admitted to FAA private pilot licence examination. Their protocol was to present flight school students (n = 48) with a pen-and-pencil knowledge test that contained a variety of questions that were either:

1) Unaltered skills questions
2) Different data skills (same wording, different values)
3) Unaltered knowledge questions
4) Shuffled knowledge questions
5) Reworded knowledge questions
6) Different knowledge questions matched for difficulty of actual items

Two tests were given: the first contained 50 questions addressing points 1-5 above, while the second test was only 20 questions long and paired points 3 and 6. The results are given in Figure 5. There was a significant difference between the different knowledge and control knowledge ($F_{(1,23)} = 31.2$, $p < 0.0001$) for the 20-item test, and for the different data skills compared to the control skills ($F_{(1,23)} = 15.4$, $p < 0.001$). However, the students still achieved 73.8% in the hardest condition for the first test, which is above the pass mark. There was no difference between average scores for unaltered, shuffled, and reworded knowledge questions. The authors conclude that "these results seem to rule out our worst fear: that participants relied solely on the crudest of memorization strategies in which learners used superficial cues available in the questions and answers." (p. 10). The score for the second, 20-item test averaged a fail mark for the different knowledge which indicates "fairly serious knowledge deficiencies" (p. 10).

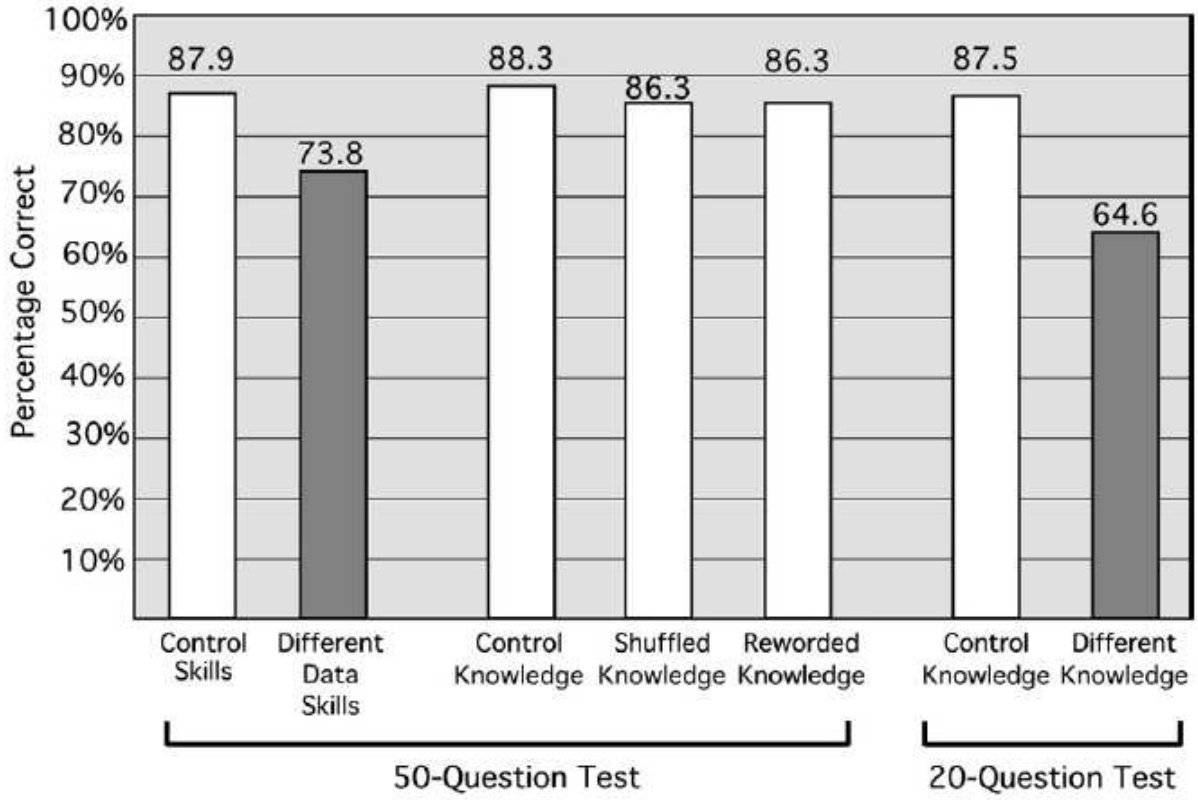

Figure 5: Casner *et al.'*s (2003) findings. For the 50-question test the average marks were in excess of the pass mark.

In this study the authors concluded that, with regards to test performances, "the FAA data clearly suggests that memorization is at work" (p. 3). The principal reason for this was an examination of test completion times, which were often completed "in far less time than would be required for the average human to even read the questions and answers on the test." The FAA (c.f. Casner *et al*., 2003) reported that some questions were answered, on average, in an astonishing half a second – and sometimes even quicker – for calculated answers. In light of such compelling evidence, there exists a real risk that students will attempt to memorize the tests when presented the opportunity.

However, it should be noted that the FAA examination, at the time, was vastly shorter than the European equivalent. The JAA required 500 hours of classroom study, whereas the FAA examination required only 35 hours of classroom study (Verheijen, 2002) and the final FAA examination was 75-questions long. So while there was strong evidence that students were memorizing questions and answers, this may be an artifact of the size of the question pool, which is much more manageable for the examination that was used for Casner and colleagues' study than for the proposed EASA-regulated examinations. **The size of the dataset to be memorized is a determining factor as to whether an individual would either attempt memorization, or whether it is even within the realms of possibility**. As the data set size increases, the potential payoff for memorization arguably decreases.

There are significant limitations to the findings found by the NASA study when examined in this present context and with a view to extending their findings to this present report. Firstly, the researchers used a sample of flight school students who had recently sat the actual FAA examination, and thus there was no control over learning techniques or the sample itself. The researchers found that the control skills questions (the same as those given for the FAA examination) were answered very well, but that the results for completely different data skills also exceeded the pass mark in the larger test and 60% in the smaller, 20-question test. It could be concluded, then, that students were aided somewhat by having access to the questions, but would have likely passed the examination even if it hadn't been published. The students could apply their knowledge to a completely unknown test showing a deep understanding and transfer of knowledge. Similarly worded questions and shuffled answers were also completed to a very high degree. Knowledge questions were questions relating to facts, whereas skills questions were more deductive and involved working out an answer.

Two years later, a follow-up study was released (Casner, Herladex and Jones, 2006). This study sought to determine how much of what qualified pilots had learned during their previous studies was retained in the form of a follow-up test. A short 10 question test was administered to 60 pilots, with the results showing that 62% of participants passed the pass-mark of 70%, 23% scored 70% and 23% failed the test. Only 38% scored higher than the national average despite having worked as pilots for several years, with the authors concluding that significant forgetting had taken place. They found that there is a trend for pilots to remember information that is relevant to their day-to-day operation of aircraft, and that irrelevant knowledge is preferentially forgotten. The authors claim that their study shows evidence that pilots devise and use simplifications of aeronautical knowledge to aid memorization, such that rote memorization of facts is better remembered than mnemonics. The researchers recommended regular study to reinforce the knowledge gained for the FAA pilot license exam, but that the areas of knowledge that are forgotten are too complex to make unambiguous claims.

3.8 The limits of memory

It has been well established since the 1950s that performance for recall of items stored in memory through rote-learning decreases as a power function with the amount of items that has been learned. Underwood (1957) compared several memory experiments from different laboratories where a main learning task had used subjects who had previously had to learn different lists as part of different experiments. These previous experiments were unrelated to the current experiment in which the students were participating; however the previous lists that had been stored had an influence on the current list to be learned through interference. These lists contained words that didn't resemble English, and sometimes contained geometric forms rather than words, and thus had to be rote-learned. Figure 6 shows the results from this meta-analysis, which demonstrates that as subjects have to rote-learn more items, their performance decreases as a mathematical power function.

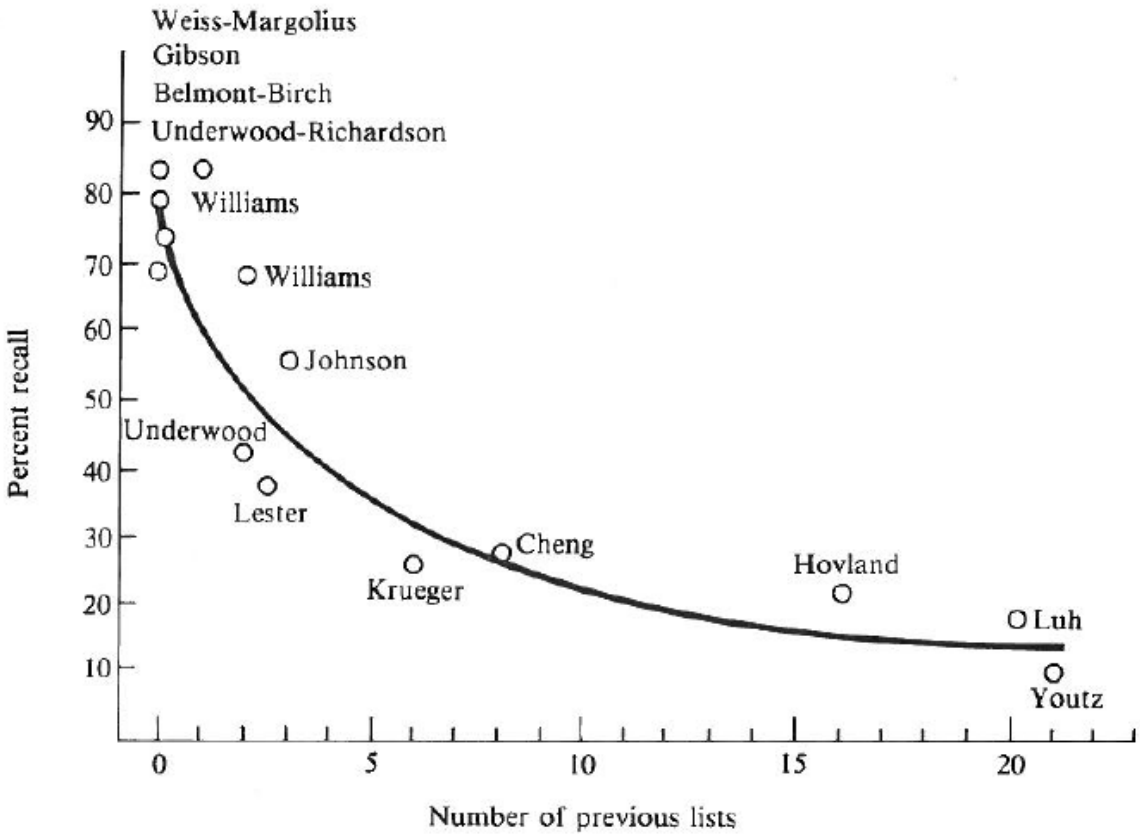

Figure 6: As more lists had to be learned by the students, the worse their performance for a new list became (Underwood, 1957).

These results are similar to short-term memory paradigms, which also show a decreasing ability to remember items in memory as more items are due to be stored. Gunter, Barry and Clifford (1981; c.f. Baddeley, 1999) showed that as successive news items were presented, the ability to correctly recall details of those new items diminished. However, performance could be improved by semantically changing the topic, whether recall was immediate or delayed. The mechanism is thought to be the same for both short and long-term memory, which is an interference effect. As the topic of study changes, it is less likely that there will be interference from similar items, which is an important consideration when evaluating whether the modular nature of the Part FCL examinations actually promotes or increases the probability of rote learning.

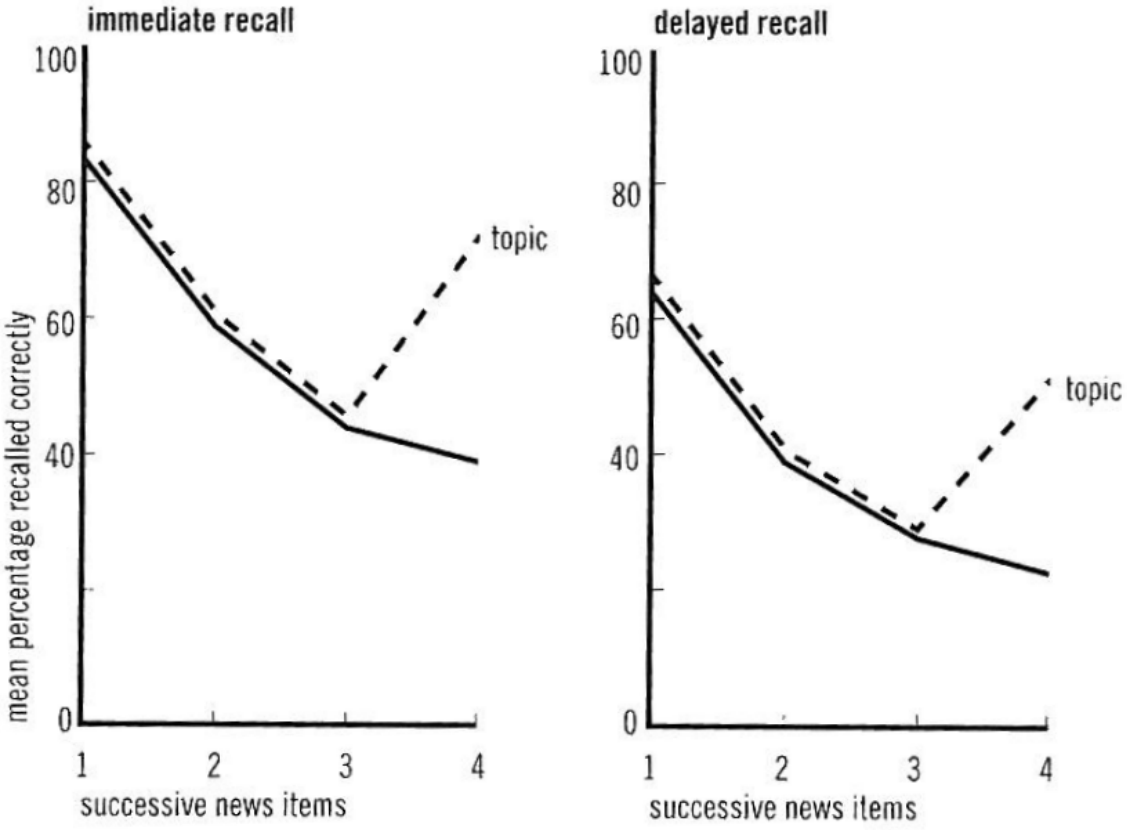

Figure 7: A short-term recall study (Gunter, Barry & Clifford, 1981) shows a similar curve to that found by Underwood (1957), who had used students that had learned several lists.

Previous research on memory storage capacity and cognitive load has concentrated on memorization of verbal lists that consist of a small number of nonsense words or meaningful words. The main reason for this research was to understand short-term memory capacity limits influenced by serial position effects, chunking, and interference. To our knowledge no other studies have investigated rote-learning capability as a function of up to 10,000 meaningful items and related long-term memory storage capacity. Brady et al. (2008) studied the information capacity of visual long-term memory by presenting several thousand familiar objects. Subjects were asked to remember all of the details of the items that have been presented for 3s each. Figure 8 shows the memory performance in the three test conditions. In the novel condition, the viewed object was paired with a new object from a distinct category. In the exemplar condition, the old object was paired with a new object of the same category, and in the state condition, the viewed object was paired with exactly the same object, but in a different state. Subjects were asked to identify the viewed object in all conditions. Their performance was best in the novel condition, where 92.5% (s.d. = 1.6%) of objects were remembered. However, the subjects' performance was as well high in the other two conditions, where detailed information was required. These results show empirically that human visual memory is capable of storing thousands of images and related detailed information in long-term memory with successful recall.

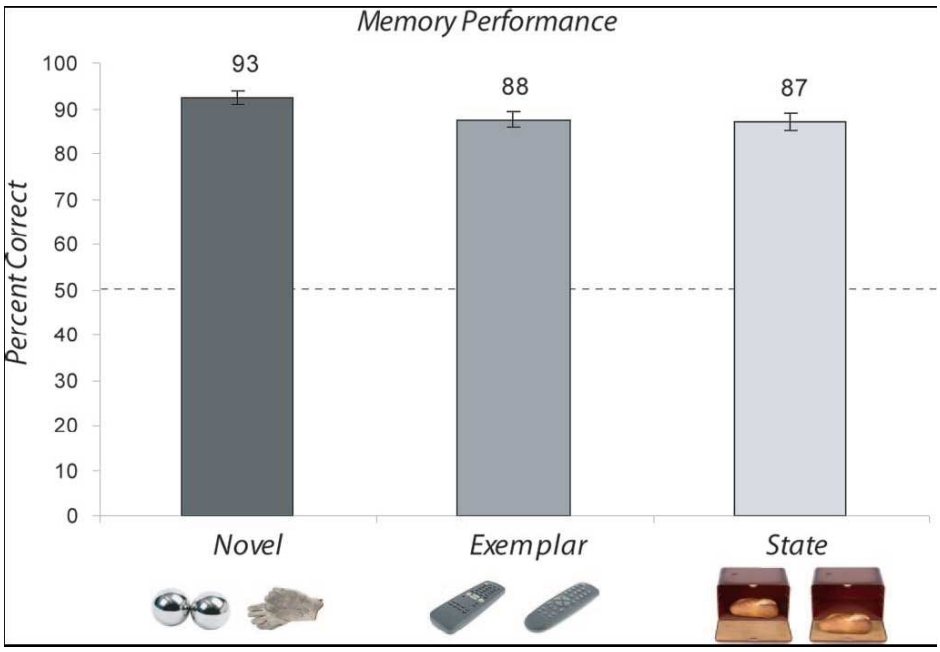

Figure 8: Memory performance for the three test conditions where the score for the discrimination between a novel object and a viewed object was highest.

There is a wide variance in human memory and capabilities, which makes concluding in absolutes impossible. In fact, there are some humans in the world who challenge the notion that human memory is imperfect; it may simply be a deficit in recall. Mnemonists have long shown the amazing range of memory found in human beings, with some individuals being able to exhibit almost perfect recall of hundreds of items even decades later (Luria, 1968). Although exceptional, these mnemonists form the upper-limit of what is possible by humans. The limits of memory for the average human being is quantifiably difficult to estimate without knowing the exact circumstances and probable demographics, which is why experimentation is important to making informed policies in specialised cases.

3.9 Examination Format

Research has found that students changed their study activities by employing surface rather than deep learning strategies when presented with an end-of-year multiple choice question (MCQ) examination (Scuoller & Prosser, 1994), but those students with a tendency towards deep learning approaches seem to continue relying on these approaches. In contrast, students are more likely to employ deep learning strategies in an open-ended examination method (Thomas & Bain, 1984). Scuoller (1998) investigated student's learning behaviour when preparing for two different methods of assessment: an assignment essay and an MCQ examination. Students who focused on surface learning approaches when preparing for a MCQ examination perceived their own levels of intellectual processing to be rather low. By contrast, students who prepared for their assignment essays by engaging in deep learning perceived themselves as having a higher level of intellectual understanding. Thus, the assessment method seems not only to influence students' learning approach but also their perception of the intellectual abilities that are requested to pass the examination.

The open question or assessment essay formats motivates students to gain a wider understanding of the topic, to relate concepts between various topics and to develop own ideas to make the material meaningful to them. MCQ assessments are perceived to examine students' recognition abilities of the learned material more than to examine understanding, and is seen to be easier, involve less effort and thinking. Furthermore, the publication of the CQB, or the fact of having at one's disposal the examination questions prior to the examination could influence students' learning behaviour even more towards the direction of a surface approach. So far little research has been done to study the effects of publishing examination questions on students' performance (Bandaranayake et al. 1990, Casner et al. 2004, Brian et al.1998). The main findings are that students performed significantly better on previously published examination questions than on unavailable examination questions. Surprisingly more capable students showed a significantly greater benefit from the publication than the less able students. Furthermore, the simple fact of preparing for an MCQ examination by practicing multiple-choice questions leads to practice effects on MCQ examination (Lundeberg & Fox, 1991, Sax & Collet, 1968), resulting in improved performance.

These findings have implications on the teaching approach in flight students' theoretical training. As the assessment method influences in such a strong way students' learning behaviour, it is essential that teachers motivate flight students in the classroom to engage in deep understanding of the material, to use their knowledge about facts and rules on a flexible way applying it to a variation of situations and circumstances and engaging in problem-solving tasks to promote transfer of knowledge rather than rote-memorization of information (Billing, 2007; Ormrod, 2004; Ward & Walker, 2008).

In summary, students adopt different learning strategies depending on the examination method and the intellectual skills being assessed. Therefore, assessment has been identified as possibly the single most potent influence on student learning, such that it informs students what they should learn, how to learn and how much they need to learn (Lundeberg & Fox, 1991; Ramsden 1992; Scouller 1994, 1996; Scouller & Prosser 1994). Thus the availability of the CQB could possibly affect pilot students' learning strategies, negatively motivating students to use a surface learning approach and to rote-learn the questions to which they had access.

3.10 Summary

The literature shows evidence that rote-learning is a poor memory strategy for the retention and transfer of knowledge when compared to other learning methods. The suspicions of the FAA that mass rote-learning by pilots was being used to pass their theoretical examinations were supported empirically by their internal response-time data. Often, the responses were given so fast that the average student wouldn't have even been able to read the questions and possible answers within many of the response-times given. For certain questions it was clear that mnemonics were being used for crude paired-associate learning. Yet, it is debatable how much of this study concerned rote memorization as the flight students already had a wide knowledge of their area coming into the examination. The findings from the results shown in Figure 5 have shaped the experiment that was conducted in our own study. In addition, several findings were taken into account, which were illustrated in Figures 2-4 and 6:

1) Figure **2** – Self testing as a memory strategy is more efficient than study alone. To replicate the way flight schools offer learning packages to students, our learning management system is based on self-testing.
2) Figure 3 – Greater exposure to questions will promote greater memorization, although not necessarily understanding. Our subjects will be encouraged to read the question and answer pairs multiple times to mimic a rote-learning strategy. Set size is manipulated between testing days, such that the second set is twice as large as the first set.

3) Figure 4 – A retest of the original questions is given a week apart to test for forgetting effects. Students will not be able to relearn the original questions so the rate of forgetting can be estimated using our study from Figure 4 as reference.
4) Figure 6 – As more information is to be retained in memory, the ability to learn new information that interacts with the previously-listed information is diminished as a power function.

In summary, repetition (thus rote learning) enhances remembering, but is consistent with the view of learning as knowledge acquisition, in which students seek to add new information to their memories. Yet it has been shown that learning the material in different ways enhances understanding. Furthermore, engaging in meaningful learning helps to establish a long-term memory, for example regarding the repetition of facts, so that the facts are not simply memorised but linked to previous stored information and as result a network of stored knowledge is built up.

Study guides are available with similar questions as those found in the actual theoretical examinations for pilot licenses. There exists a real risk that some students will form learning strategies that are designed to pass the test, but ultimately result in a lack of conceptual or flexible understanding of the subject matter. Some rote-learning is implicit in the examination design, but adopting a pure rote-learning strategy is a dangerous proposition, especially when taking into account that the principles that are being examined are life-long information that could be tested at an inopportune moment while flying a real aircraft. Furthermore, knowing which questions students will be examined on, students might exclusively engage in learning the topics covered by these questions, ignoring the remaining theory.

Previous research has shown that FAA pilot license students perform significantly worse in examination when data skills and knowledge is switched from the items they had learned to similar items, but functionally different. This is a result of poor transfer abilities to apply their knowledge to test questions that are different to those they had learned for the examination. Although, it should be stated that reworded questions were still very well answered, showing promise for the students' abilities to apply knowledge to similar items that they had already learned.

The NASA researchers used a 50-question sample for the 900+ questions used in the FAA question databank. By contrast, the EASA proposal consists of a much larger databank of questions. It is unsure how the students from their first study would have retained memory of these questions some point after the exam, but a follow-up study using a different sample found that significant forgetting had occurred (Casner, Heraldez, & Jones, 2006). The study conducted for this report used a within-subjects analysis rather than the between-subjects employed in the NASA studies, which should allow greater accuracy in modeling and prediction.

To summarise:

1) The cognitive processes involved in rote learning are not conducive to good retention of memory. Rote learning also decreases the ability of students to apply what they have learned to similar problems or even problems that are related but outside the scope of their studies.
2) Students in theoretical training may be tempted to learn questions by rote given the huge amount of information to be retained across the 13 learning categories. Research by the FAA shows that students will go as far as to learn mnemonic devices or 'flashcard' training to create simplified methods of remembering.
3) Several studies suggest that the relationship between the number of questions and performance is non-linear, and likely to be decaying power function. Thus, the number of questions to be published as well as whether they are tested all at once or in several exams with several months in between are key factors which will determine whether students will apply a rote learning strategy or not.

4 Experimental study

4.1 Introduction and hypotheses

The experimental protocol was designed to investigate whether flight school students could rote-learn a set of published questions to pass the Airline Transport Pilot License (ATPL) examination. Running an experiment is an important aspect of making informed decisions, as the results can provide an indication of what is likely to happen in the field when a databank such as the Central Question Bank (CQB) is made public. However, the experimental outcome can only be taken as an indication of what is a probable outcome, and the results should be interpreted with caution as there are a number of factors that can differ between the laboratory and the real-world. Nevertheless, in combination with a literature review of existing scientific knowledge, a controlled experiment is the most reliable method of simulating and quantifying how well students without a wider knowledge of aviation can learn the CQB.

The following hypotheses were investigated in our experiment:

1. If rote-learning is an effective strategy, then the students will be able to achieve the pass mark of 75% without having in-depth knowledge of the field.
2. If students only learn the exact question and answer pairs, then there will be marked difference between the original and reworded questions.
3. If rote-learned material is not stored effectively in memory, then there will be an effect of forgetting between two test periods spaced a week apart.
4. If rote-learned material is not stored effectively in memory, the decay in performance in the post-test will be larger for the reworded questions (hypothesis of interaction between type of questions and test phase on performance, see below).
5. If there is an effect of test size, then performance for the second test will be worse than the first despite proportional learning time given to both. The difference has to be larger than the difference in forgetting to be valid half of the questions had been learned a week prior to the final test.

4.2 Experimental design

The experimental design was a within-subjects repeated-measures design using a sample of twenty University students who would attempt to rote-learn a battery of questions provided by a leading European flight school. A custom-designed learning management system was written to facilitate learning and data collection. The protocol would take place over the space of a week and each subject would:

1. Learn 136 questions from the flight school preparation exam. These questions are actual questions that students use to prepare for the ATPL(A). Subjects were required to read each question at least five times to rote-learn the material.

2. The following day the student reported to the CASRA research centre to undertake the computer-based examination. This examination encompassed the 136 rote-learned questions and 136 reworded questions presented in random order. These reworded questions were similar in content to the learned questions, such that no new knowledge was required to answer them correctly.

3. A week later the subject then learned a new set of 136 questions, with the same requirement of reading the questions five times. The following day they undertook the final 544-question exam, which included the learned and reworded questions from the first test, as well as the second set of 136 questions and reworded questions from this set, presented in random order. This was designed so that the rate of memory retention could be measured.

4. The two sets of questions were counterbalanced so that there could not be any influence of set difficulty if such a difficulty did exist between the two sets of questions and reworded questions.

All responses during learning were stored in a database and monitored prior to the examination to ensure that the subjects had spent the requisite time learning the material. All of the subjects were found to have complied with instruction, resulting in 28,729 questions being read across the twenty subjects with an average reading time of 17.2s. For the tests themselves, 16,320 answers were stored, including reworded questions, with an average response time of 14.5s.

Different question conditions were used in the first and second test:

First Test	
Set 1 Test 1 Original	The first set of questions to be learned in the same wording as during learning.
Set 1 Test 1 Reworded	The first set of questions, but reworded and the answers shuffled.
Second Test	
Set 1 Test 2 Original	The questions from the first round of learning, repeated a week later.
Set 1 Test 2 Reworded	The reworded questions from the first test, repeated a week later.
Set 2 Test 2 Original	The second set of questions given in the second round of learning.
Set 2 Test 2 Reworded	The reworded version of the second round of questions, with answers shuffled.

Table 1: Classification of the question conditions

Two examples that typify the difficulty of the 272-question battery are given below for reference:

Q: FL 80, an OAT +6°C is measured. What will the temperature be at FL 130, if you consider the temperature gradient of the Standard Atmosphere?

a) -4°C Correct
b) +2°C Incorrect
c) 0°C Incorrect
d) -6°C Incorrect

Q: FL 110, an OAT -5°C is measured. What will the temperature be at FL 50, if you consider the temperature gradient of the Standard Atmosphere?

a) -3°C Incorrect
b) +3°C Incorrect
c) 0°C Incorrect
d) +7°C Correct

Our laboratory, learning management system and strict design allowed control over:

1. The number of questions read during learning for each of the subjects.
2. The time interval between learning and taking the test was constant for all participants.
3. The time interval between testing sessions was kept constant at one week.
4. Room temperature, ambience, noise, comfort, and interface were constant for all participants.
5. Participant expertise and education was kept constant during screening.

4.3 Subjects

All applicants for the study were screened for lifestyle, health and education requirements such that only healthy and reasonably-well educated subjects would form part of the examination. Participants had to have low weekly alcohol intake, abstain from recreational or performance-enhancing drugs and consume only a low to moderate amount of caffeine during the week. Finally, subjects were required to be fluent in English, which is the language which the ATPL(A) examination is administered in.

Subjects with previous exposure to flight schools, specialised aviation experience or flight learning materials were excluded from the study, such that only new-to-aviation subjects would be part of the experiment. This was designed so that the only possible strategy to pass the examination would be through rote-learning of the material the day before the examination. After completion of the experiment the participants were paid an honorarium of 290CHF (~ 186 euros), and the best three performances were awarded with a bonus of up to 100CHF extra as incentive to perform well.

Twenty subjects aged between 20 and 27 (mean = 24.15 ± 2.21) were recruited, of which half were males and half females. We sought to determine whether there existed a difference in learning strategy or outcome between the genders.

4.4 Statistical Analyses

Statistical analyses are used in psychological research to empirically evaluate the importance or size of the difference between dependent variables. When comparisons are made these are known as inferential statistics, or when information is given (such as the mean percent correct in a test) then these are known as descriptive statistics. Inferential statistics are typically accompanied by a measure of the experimental change compared to individual and group differences, and a portion of unexplained variance. There is also a measure of 'experimental significance', which is an estimate of whether the observed results are equal to or exceed results that would be expected in a chance scenario. If the results significantly appear to exceed chance results, then the p-value alpha statistic will be less than 0.05, otherwise we assume that the results are caused by random effects. Finally, an effect size measure (d or η^2) is used to determine whether the difference in means can be considered genuinely large, or if a significant p-value is not significant in magnitude. For the results given, a d of less than 0.2 is relatively small, between 0.2-0.5 is medium and above 0.8 is considered large (Cohen, 1988). η^2 (partial eta-squared), as it is calculated here (Bakeman, 2005) should be considered as less than 0.01 as small, between 0.01-0.06 as medium and higher than 0.14 as large.

Although statistical analysis is a cornerstone of good psychological research, caution must be exercised in their interpretation (Cohen, 1994; Frick, 1996). The actual values and change in values across conditions are often more insightful than the outcome of statistical significance testing (i.e. p-values). Therefore, we urge the reader to place emphasis on the interpretation of graphical data that is provided for every result than misinterpret 'statistical significance.' The standard deviations are given for each graph, which is a measure of inter-individual differences.

4.5 The learning management system

In order to achieve an automated learning environment similar to existing systems used to train flight students, a custom-written learning and test suite was created. The suite was web-based so that students could learn the questions and answers at home and then undertake the examination at the CASRA research centre using a similar interface and the same login credentials. Once students had registered and set up an account, they were able to proceed to the learning area which showed them a question and all of the possible answers. They would then click the "next" button where they were shown the actual answer to learn (Figure 9).

If students were clicking too fast they were given a warning to slow down (reading time < 10s) to ensure that the material was fully learned. The program kept track of how many times the questions were read and how fast the students were going through learning. If a student logged out, the system would resume from the point where she or he had left.

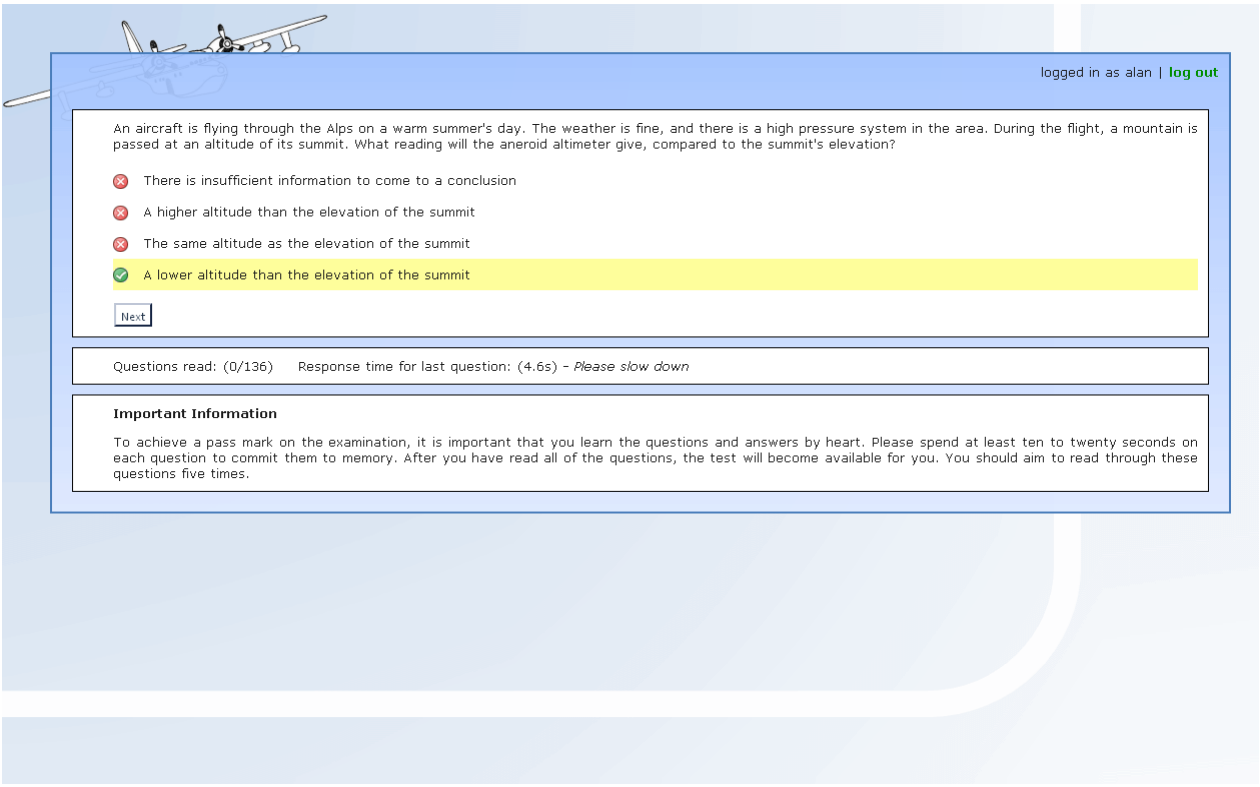

Figure 9: Screenshot of the learning management system. The question at first does not show the answer, and on the "next" click the correct answer is highlighted as shown in the screenshot.

In regards to testing, the system changed to give a random question from the appropriate set with a radio-button list of the possible answers. The students had to pick an answer from these and were not able to select multiple answers or fail to answer the question. The system gave an indication of how many questions remained to be answered, and then gave a confirmation message of completion when the test was finished.

5 Results of the experiment

5.1 Results of parametric statistics (t-tests and ANOVAs)

Twenty students (10 males, 10 females) with a mean age of 24.15 (s.d. = 2.21), all from the University of Zurich, completed the study. One additional male was recruited but abandoned the experiment during the first phase of study. For the first set of questions the students read the questions an average of 5.22 times (range 5.0 – 6.0), and for the second set the students had read the questions an average of 5.13 times (range 5.0 – 6.0).

The overall results for each question condition are given in Figure 10 below. The worst result occurred in the second test for the reworded questions of the new set of questions (set 2), while the best performance occurred in the first test for the original questions. However, all of the results exceeded the pass mark of 75% as an average. For the overall average of the second test (544 questions) and the first test (272 questions), there were 7 subjects who exceeded 90% correct, 78 who exceeded 80% correct, 1 who exceeded 75% correct and 5 subjects who failed to break the pass mark of 75% (75% of subjects achieved marks over 75%, while 85% achieved marks above 70%).

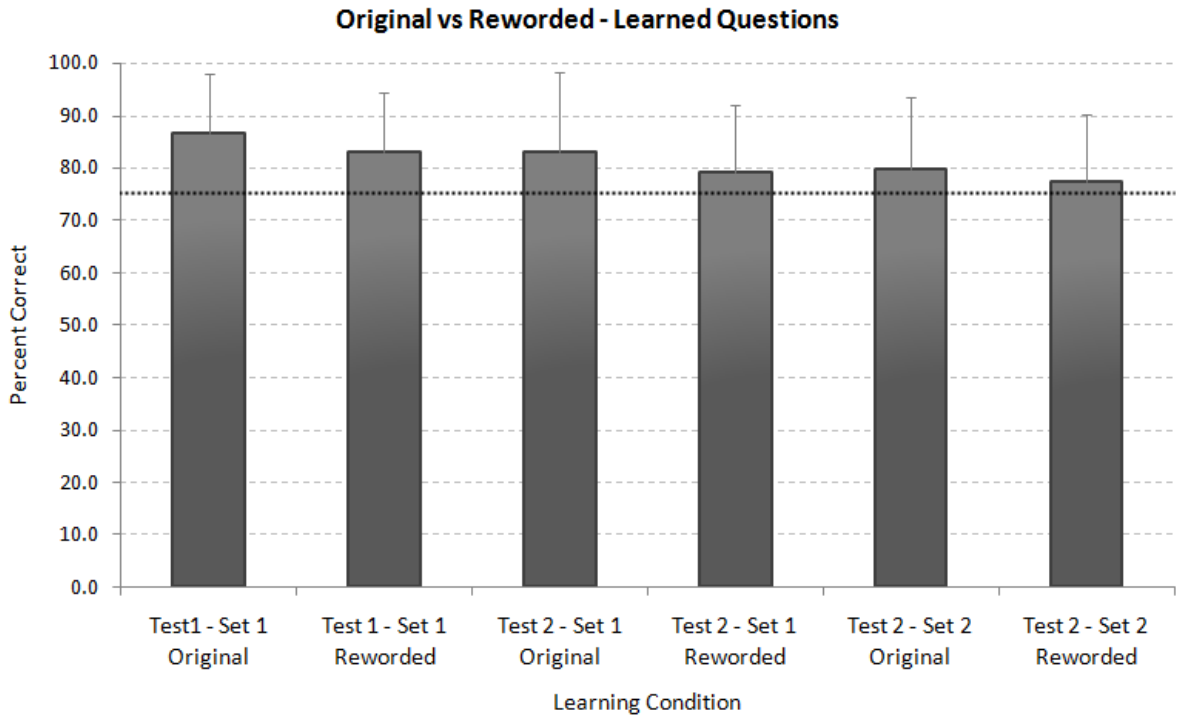

Figure 10: Breakdown of the overall results in each condition. The error bars represent the standard deviation, while the dotted line represents the pass mark (75%).

The overall results from Figure 10 can be compared and broken down into other meaningful comparisons. Two-way paired t-tests were used for each comparison as each measure used the same subjects for each question condition. The rate of forgetting is a comparison of performance for the first set of learned questions between the two test periods, where only one day of learning was permitted (Figure 11). There was a statistically significant decrement in performance between the two tests for both the original questions ($t_{(20)} = 2.390$, $p < 0.05$, $d = 0.53$) and the reworded questions ($t_{(20)} = 2.824$, $p < 0.05$, $d = 0.63$). The magnitude of difference (d values) is considered medium to large (based on Cohen, 1988); although subjects still managed to exceed 80% correct a week after learning.

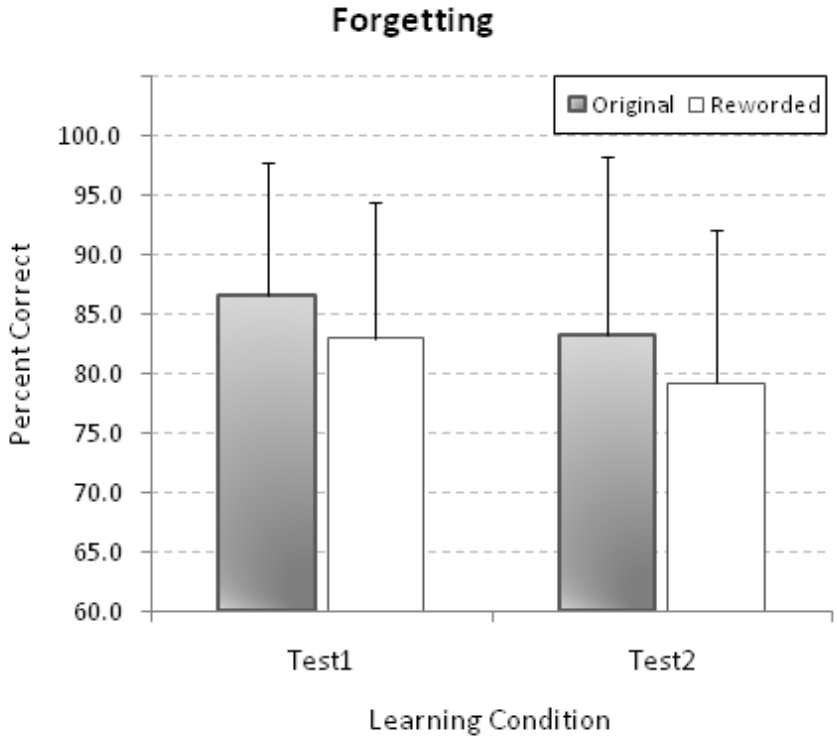

Figure 11: Comparison of performance between the two test periods for the first set of learned questions.

The effect of rewording the questions is compared by analysing the results from the first test original and reworded questions, and the second set from the second test (Figure **12**). There was a highly statistically significant main effect of question rewording for the first set of questions ($t_{(20)} = 5.302$, $p < 0.01$, $d = 1.19$) and for the second set of questions ($t_{(20)} = 2.969$, $p < 0.01$, $d = 0.66$). However, in absolute terms, this represents a difference in percent correct of only 3.67% for the first test and 2.54% for the second test. Although the statistical significance appears high, the actual values are not what would typically be considered as large.

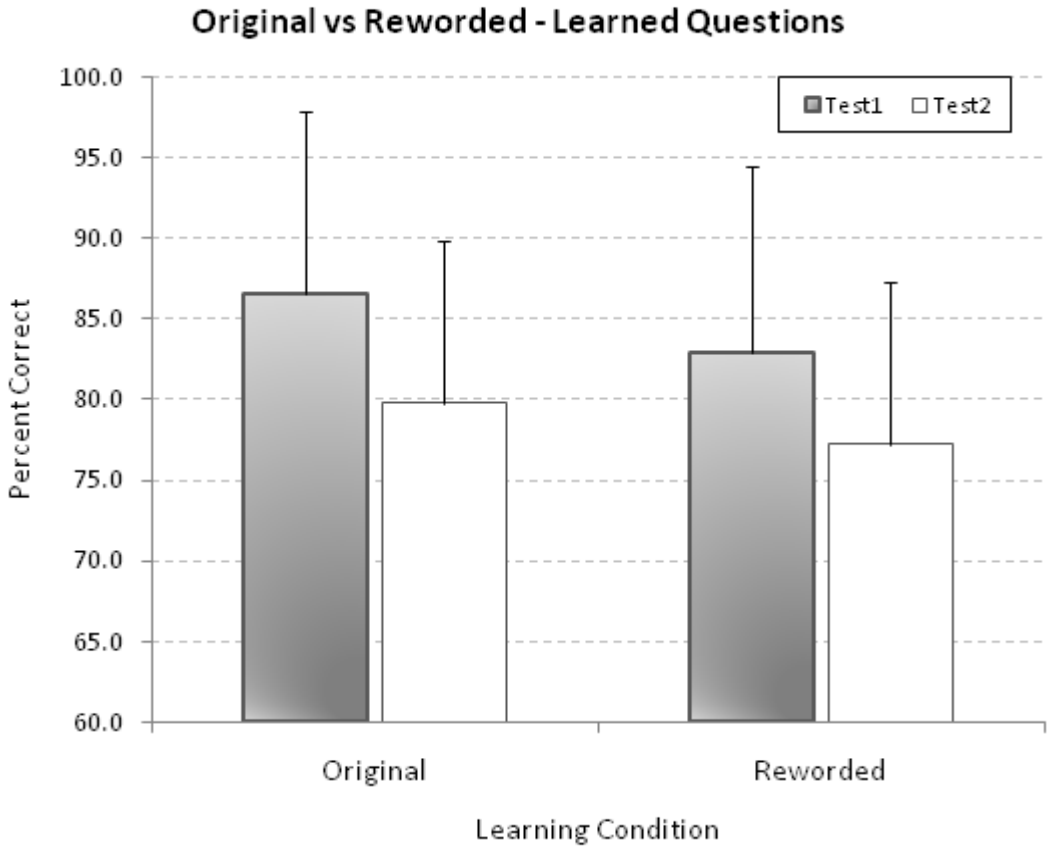

Figure 12: Comparison of the original questions and the reworded questions. Set 1 is compared for the first test and set 2 for the second test.

The overall results from the first, smaller test and the second test is shown in Figure 13. Here was a statistically large decrease in performance between the two tests ($t_{(20)} = 4.998$, $p < 0.001$, $d = 1.12$), with an absolute difference of 4.91%. There was no statistically significant difference between the genders or as a split-half analysis of age. The rate of forgetting was shown as a difference of 3.42% for the original questions and 3.75% for reworded, so the test size statistic needs to take this into account as half of the questions were learned a week prior to the other half.

Figure 13: Comparison of the first test (272 questions) against the second test (544 questions).

Subjects rated how well they thought they had performed for each of the tests after the second test on a scale of 0-100, with 100 meaning they had managed to achieve a perfect score, and 0 representing all questions were answered incorrectly. For the first test there was a large correlation between the subjects' perceived performance and their actual performance, see Figure **14** ($r = 0.81$). The same relationship was true for the second test (Figure **15**), where the correlation values and actual scores were correlated marginally higher ($r = 0.83$).

A two-way repeated-measures analysis of variance (ANOVA) was used to test for interaction effects between the original questions and the reworded questions, with the second level of time (first test vs. the second test a week later). The output from this analysis is provided in the appendix. There was a statistically significant finding of question presentation ($F_{(1,19)} = 34.278$, $p < 0.05$, $\eta^2 = 0.64$) as well as forgetting ($F_{(1,19)} = 28.53$, $p < 0.05$, $\eta^2 = 0.60$). There was no interaction between the question presentation and the time that had elapsed between the tests ($F_{(1,19)} = 1.22$, $p = 0.28$, $\eta^2 = 0.06$).

Using the results from the second test only, another two-way repeated-measure ANOVA was used to investigate the effects of the first set of questions compared to the second set, and the associated reworded questions between the two sets of questions. There was a statistically significant difference between the original questions ($F_{(1,19)} = 19.01$, $p < 0.05$, $\eta^2 = 0.50$) but not for the reworded questions ($F_{(1,19)} = 4.34$, $p = 0.051$, $\eta^2 = 0.19$). There was no interaction effect ($F_{(1,19)} = 2.93$, $p = 0.103$, $\eta^2 = 0.13$).

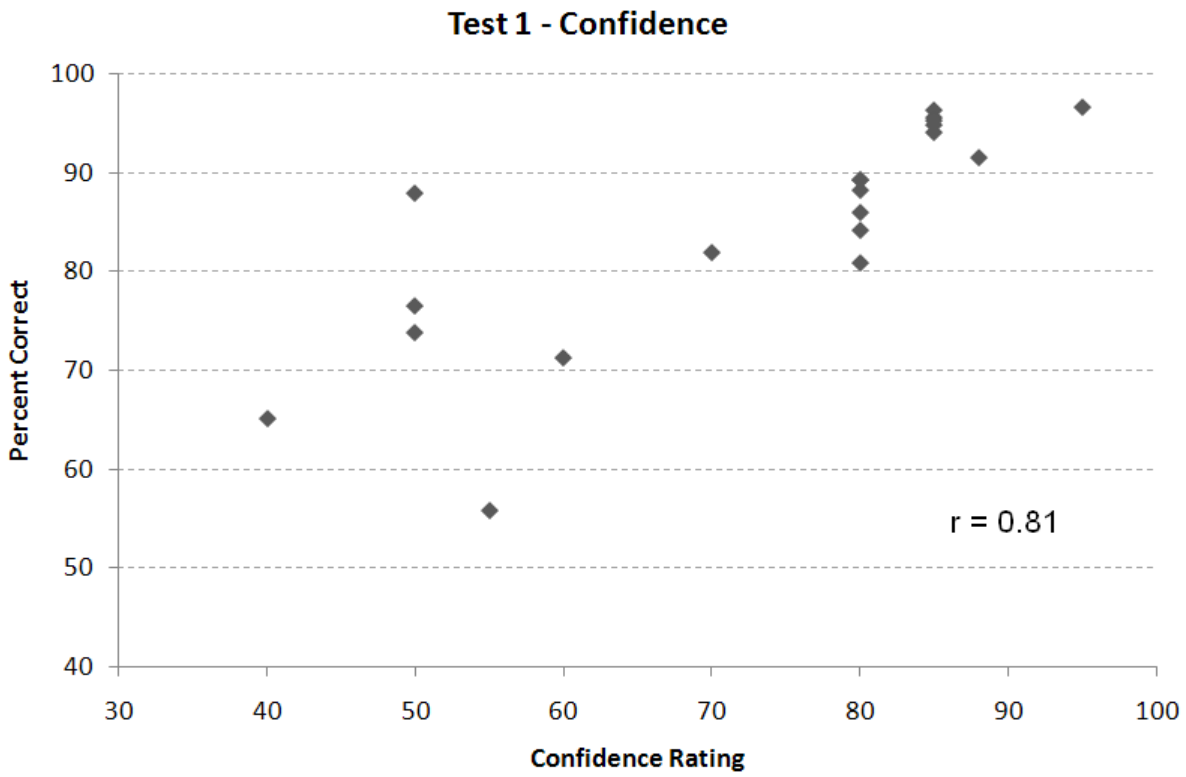

Figure 14: Correlation between the confidence rating and the actual score for the first test.

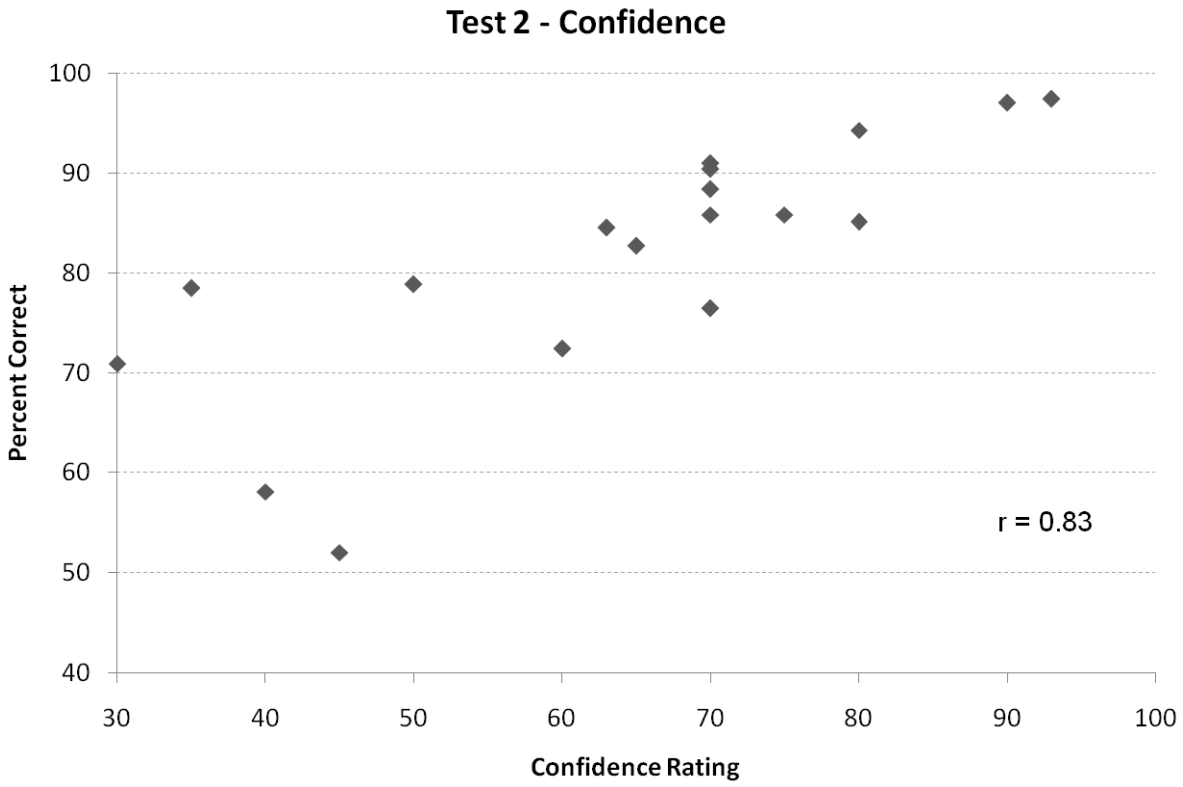

Figure 15: Correlation between the confidence rating and the actual score for the second test.

5.2 Statistical probability to memorize the entire CQB

To estimate the probability that students can rote-learn a battery of up to 10,000 CQB questions and answers, projections can be made from the sample of students used in our experiment. Previous literature has suggested that the relationship between number of questions and performance is non-linear, and likely to be a decaying power function (Underwood, 1957). To this end a probable power function was projected onto the data of original questions from our experiment (i.e. not the reworded questions, which were used to test transfer of knowledge onto similar situations), see Figure 16 and Table 4.

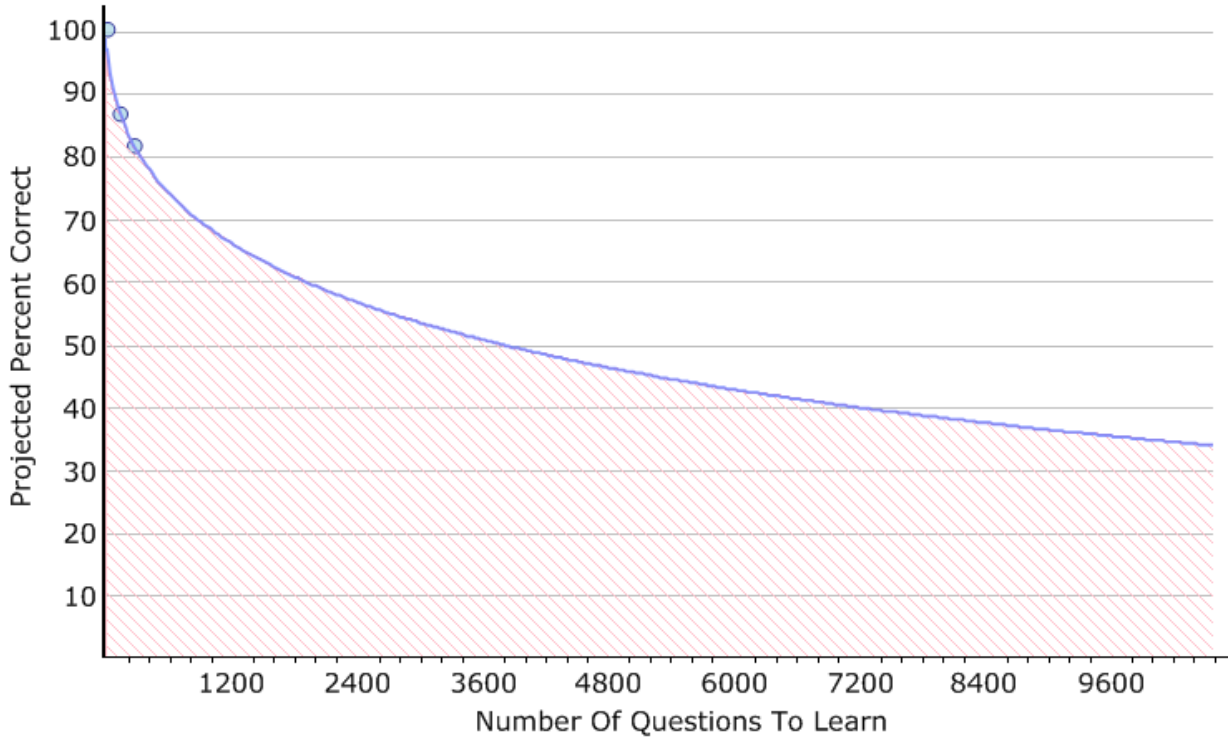

Figure 16: A projection of the number of questions to be learned and percent correct from the viewpoint of rote-learning.

No. Questions	Projected Percent Correct
1	100.0
1000	68.2
2000	59.2
3000	53.3
4000	49.0
5000	45.5
6000	42.6
7000	40.2
8000	38.1
9000	36.3
10000	34.6

Table 2: Approximate values for the fitted functions of Figure 16.

To fit the curve in Figure 16, a least-squares function fit of power functions was applied using several curve-fitting techniques (i.e. Fourier, Power, Exponential and Polynomial regression). The aim was to find the best-fitting curves through the three data points found from the experiment to demonstrate how a probability curve may appear that is based on the premise that performance decreases as a power function (see Section 3.8 for examples in the literature). A suitable curve would be one that declines steadily in a similar fashion to the Section 3.8 curves and fits the data points nearly exactly. Specifically, this curve is a Freundlich power curve, but many competing and similar curves can be fit equally-well through the three data points; some of which taper above 35% around 10,000 questions and some below (all fit above 25%). The r^2 goodness of fit approximates 1 as the curve was selected a-posteriori (after the data was collected) to fit the data average data points. This curve is an *estimated projection*, because it is fundamentally impossible to predict data with complete certainty as to what the future values will be. It is shown to give the reader an impression of how the probability to rote learn will likely decrease as the number of questions increase, but there are large variations between individuals as to how their individual curves would appear.

The power function gives an indication of where the average likely performance will occur if similar conditions are given for the flight students as those which the University students studied under. The University students spent one day per 136 questions and averaged over 80% for both tests without any previous knowledge of the subject area. This curve can be fitted using the formula **$y = 70.3 * \exp(68.98/(x+194.53))$**, but it should be noted that there is always a wide variance between subjects that will cause actual values to fluctuate around this curve. This projection is based on a rote-learned sample of 272 questions in two days, and more precise values are not possible within the limitations of this controlled study. Table 2 shows the values for multiples of 1,000 questions to be learned, such that in the best-case scenario students might approximate a pass mark of 75% for the 10,000 question CQB, and in the worst they may achieve a score as low as around 35%. This curve rests on several assumptions given in the discussion and represents what could happen if the examinations for the ATPL(A) are given in a short space of time, necessitating memorization of the bulk of the CQB.

If the probability curve does settle near 35% for 10,000 questions, then this figure is only 10% above what would be expected in a pure-chance scenario. As there are only ever four possibilities, and only one correct answer, then the statistical probability of guessing the correct answer is 1 in 4, or 25%. Therefore, it would not be much above what would be expected by purely guessing the answers. Also, there are several variables that can cause performance to vary such as time spent for learning, motivation, difficulty of the questions, learning strategies, previous knowledge, test resits, etc.

5.3 Statistical probability to memorize the CQB per module

According to a leading provider of JAR FCL theoretical examination preparation software and workbooks, Dauntless-soft, the examinations can be undertaken at a students' request, but are typically completed in five to six sittings. Therefore, the average student will complete about two modules per sitting. The length of examinations can vary greatly between modules, with some modules taking only 30 minutes to complete, while others can take upwards of three hours. If the average CQB module contains around 800 questions at a minimum (800 x 13 = 10,400), then a student could attempt to memorize around 1,600 questions per examination sitting (based on two modules per sitting).

Figure 17 shows the probability curve to learn 1,600 questions based on the same assumptions laid out in the previous section, namely that each question is read five to six times and is based on purely rote-learned material (pre-existing knowledge cannot be factored into this estimation). If the same conditions are met, then it would take around 12 continuous days (11.76 days) to read all of the questions five to six times as the students did. However, we know from previous studies that the rate of learning and memorization is not linear, and as new knowledge is stored it will interact with previous knowledge. Twelve days represents a highly conservative estimate and the actual figure is probably a lot higher, as the trend will decay towards chance values unless compensatory learning is undertaken.

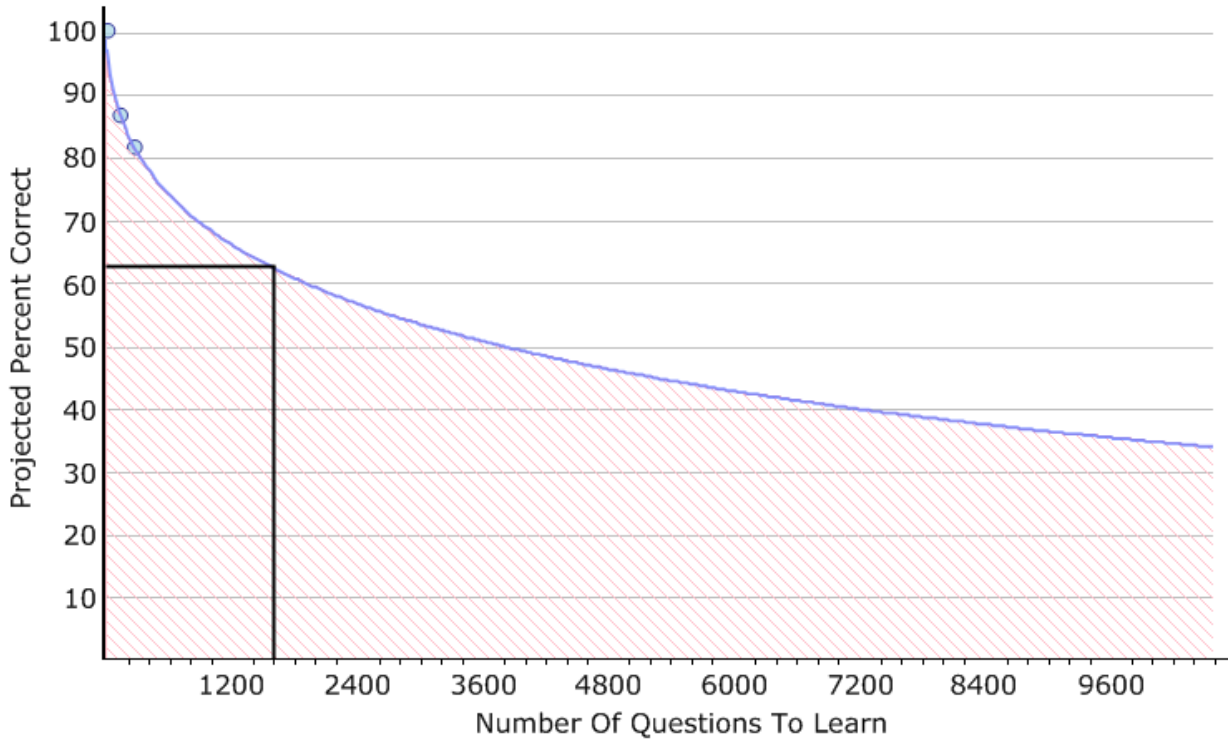

Figure 17: When learning is restricted to just two modules, the average projection of the curve is estimated to be around 65% correct.

To extend these estimates, it is theoretically possible that a candidate could undertake two-module examination sittings every three weeks using the rote-learning techniques outlined by this study and have all thirteen examinations finished in five months. As has been previously stated, the students' results for the 272-questions to be learned in just two days of study were exemplary, with the majority of candidates (75%) exceeding a 75% mark. However, these estimates rest on some fairly serious assumptions, for example that a potential flight student could attempt to learn a meteorology module without understanding the computations that underlie the computed questions (see Chapter 3).

The central question surrounds whether candidates will have sufficient time between examinations to learn an entirely new set of question and answer pairs while effectively forgetting previous examination preparations so that it does not interfere with learning the new knowledge. The results from full 10'000 question curve relied on the assumption that the examinations would be close together, and that memorization of the full CQB would be necessary to pass the full JAR FCL licenses. Yet the advent of self-chosen examination dates and the possibility of resits make the issue less clear.

5.4 Comparison with the NASA study results

In Casner *et al.'s* (2004) study the authors concluded that with regards to test performances "the FAA data clearly suggests that memorization is at work" (p. 3). The principal reason for this was an examination of test completion times, which were often completed "in far less time than would be required for the average human to even read the questions and answers on the test." The FAA (2003, c.f. Casner *et al*., 2003) reported that some questions were answered, on average, in an astonishing half a second – and sometimes even quicker – for calculated answers. In light of such compelling evidence, there does exist a real risk that students will attempt to memorize the tests.

However, it should be noted that the FAA examination, at the time, was vastly shorter than the European equivalent. The JAA required 500 hours of classroom study, whereas the FAA examination required only 35 hours of classroom study (Verheijen, 2002) and the final FAA examination was 75-questions long. So while there was strong evidence that students were memorizing questions and answers, this may be an artifact of the size of the question pool, which is much more manageable for the examination that was used for Casner and colleagues' study. The size of the dataset to be memorized is a determining factor as to whether an individual would either attempt memorization, or whether it is even within the realms of possibility. As the data set size increases, the potential payoff for memorization arguably decreases.

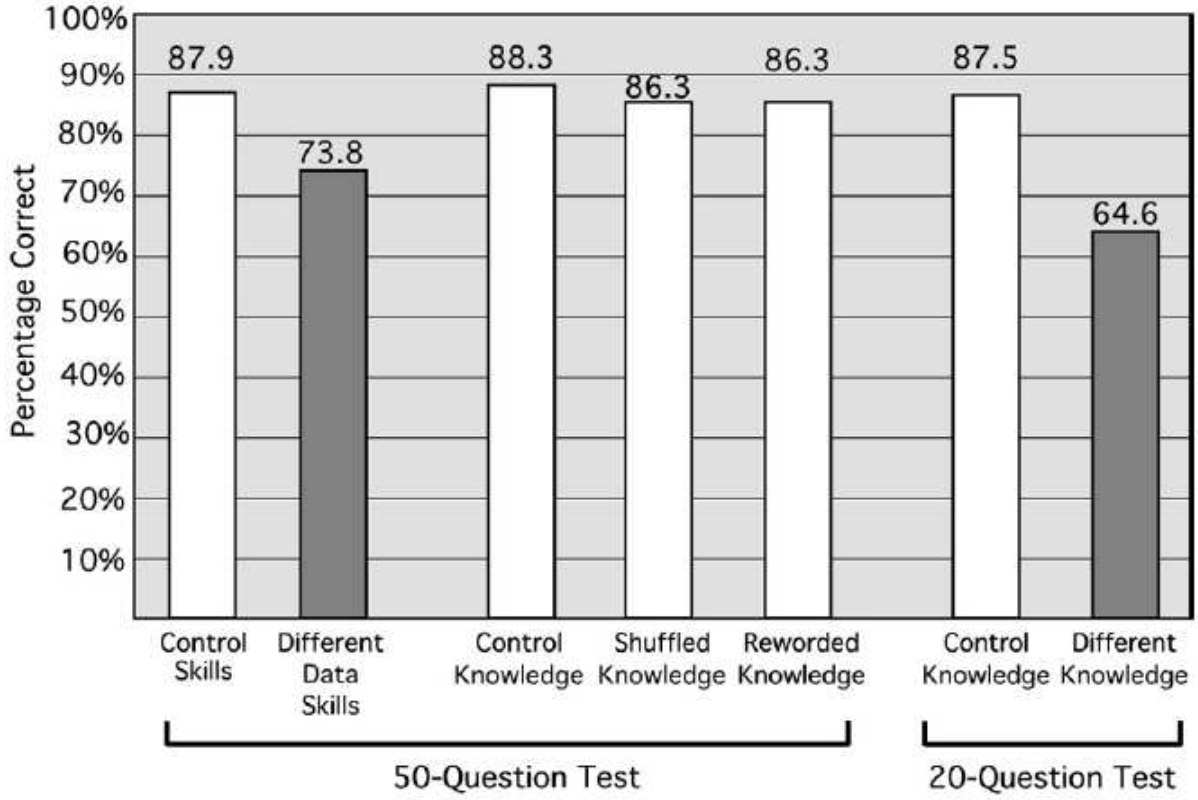

Figure 18: Casner *et al.'*s (2003) findings. For the 50-question test the average marks were in excess of the pass mark.

There are significant limitations to the findings found by the NASA study when examined in this present context. Firstly, the researchers used a sample of flight school students who had recently sat the actual FAA examination, and thus there was no control over learning techniques or the sample itself. The researchers found that the control skills questions (the same as those given for the FAA examination) were answered very well, but that the results for completely different data skills also exceeded the pass mark in the larger test and 60% in the smaller, 20-question test. It should be concluded, then, that students were aided somewhat by having access to the questions, but would have passed the examination even if it hadn't been published. The students could apply their knowledge to a completely unknown test showing a deep understanding and transfer of knowledge. Similarly worded questions and shuffled answers were also completed to a very high degree. Knowledge questions were questions relating to facts, whereas skills questions were more deductive and involved working out an answer.

The results from our study were remarkably similar in scores, even with the differences in protocol, sample and test size. Although the students could rote-memorize more than 100 questions per day and achieve scores above 80%, it is improbable that they would have performed even nearly as well on a different dataset. This is because the knowledge was accrued for a very specific circumstance, unlike the flight students used for the NASA study, whose learning was designed to be applied to many varying circumstances and explains the high scores in the different dataset condition.

6 Discussion

6.1 Potential caveats of the study

The scientific study conducted to aid in the decision-making process, whether or not to publish the CQB, was successfully completed and has provided a benchmark to project what typical university students with no knowledge of aviation can achieve. By doing this, a separation between rote-memorization from other methods of learning was achieved, which is a useful parallel to the NASA study (2003) that contained a mixture of learning strategies. Although our study provided an invaluable resource to compliment the research derived from the literature, there are limitations to any scientific study that need to be acknowledged.

Firstly this study, like all scientific studies, could be enhanced with a larger sample, more resources to investigate similar or competing hypotheses, or alterations made to the existing paradigm such as a follow-up period months later, or a larger question set. The sample size had to be limited to a representative sample of participants that did not exceed the amount that could be paid and the time with which the study had to be completed. Given that the students were required to expend over ten hours of learning and testing to this study, they had to be paid accordingly. It would not be feasibly possible to obtain such a sample of non-flight students without offering financial remuneration. A request to obtain a larger, supplementary internet-based sample by opening up the web-based learning management system to anyone in the world was rejected by the question copyright owners, who understandably did not want the questions to become known to their current flight students or copied by other entities. The maximum number of paid students we could recruit was limited to twenty to keep within the budget allowance and timeframe of the study.

Secondly, with any prediction of human performance comes an understanding that performance between and within individuals is subject to variance, and an estimation of error is itself liable to fluctuations depending on the circumstances. For the results, this unknown variance is mathematically estimated as either a confidence interval (lower to upper bound) or as a standard deviation. Factors that are known to affect human performance are listed in Table 3 below. Most relevant to pilots will be their learning strategies and time spent learning the material, as a measure of time output. The higher the weighting that a variable has on performance, the more likely it is to explain the variances. This brief table is given to provide the most likely causes of inter-individual variation, not including specific examination-day variables like sleep loss, nutrition, or stress.

Variable	Interpretation
Learning strategy	Depending on the examination, certain strategies may be predisposed towards good examination results and long retention of learned materials. Certain learning strategies favour different individuals, although it is generally considered that deep learning of material is more beneficial than surface learning for long-term retention.
Time spent for learning	The longer the time that is spent for learning, the better a candidate is likely to perform. However, this is mediated depending on whether the learning is spread over different learning sessions or 'crammed' during few. Also, individuals may require more or less time comparatively than other individuals in order to achieve similar marks depending on how quickly material is stored in memory.
Motivation to learn	Motivation can be intrinsically or extrinsically rewarded. An enthusiastic and keen candidate is more likely to seek out learning methods and strategies that promote a wider appreciation of the subject matter. Motivation varies between individuals and may explain, at least in part, why some candidates study longer than others.
Test resits	Candidates with prior exposure to the examination may be better prepared when resitting the examination than on their first attempt. This could be a function of accumulated knowledge in the interim and better preparedness for the examination contents.
Base knowledge and intelligence	One of the foremost differentiations between candidates may include existing flight or piloting knowledge and whether a candidate is studious, including previous examination records in other fields of study.
Classroom attendance	Higher classroom attendance may indicate a greater accumulation of knowledge, or indicate higher levels of enthusiasm and dedication.
Personality factors	Personality factors such as introversion/extraversion, conscientiousness, determination, and motivation have been implicated in the ability to study (Eysenck. 1997).

Table 3: Variables that may alter a student's examination outcome

6.2 Results and discussion regarding experimental hypotheses

We begin the discussion of results by providing the results regarding the hypotheses specified in section 4.1.:

1. If rote-learning is an effective strategy, then the students will be able to achieve the pass mark of 75% without having in-depth knowledge of the field.

   University students, who were carefully screened for having no knowledge of aviation, managed on average to exceed the pass mark for the examinations, even when the questions were reworded and the potential answers shuffled. Therefore, rote learning was an effective strategy to pass the examinations in the experiment conducted for this report.

2. If students only learn the exact question and answer pairs, then there will be marked difference between the original and reworded questions.

There was a statistically-significant difference between the performances for the original and the reworded questions. Despite this, the differences could not conventionally be described as "marked." Performance did worsen, but the absolute differences were small (around 4% difference for the first test and around 3% difference for the second test).

3. If rote-learned material is not stored effectively in memory, then there will be an effect of forgetting between the two test periods spaced a week apart.

   There was a statistically-significant difference between the questions that were given a week apart, but again the absolute difference was small at around 3%. Therefore the effect of forgetting is quite small for a week apart.

4. If rote-learned material is not stored effectively in memory, the decay in performance in the post-test will be larger for the reworded questions (hypothesis of interaction between type of questions and test phase on performance)

   There was no statistical interaction between the type of question and the time period. Therefore it appears that the information that has been stored in memory is not differentially impacted by the presentation of the questions. This indicates that the memorized information can be recalled even when the questions and answers are different from the presentation during learning.

5. If there is an effect of test size, then performance for the second test will be worse than the first despite proportional learning time given to both. The difference has to be larger than the difference in forgetting to be valid as half of the questions had been learned a week prior to the final test.

   Students answered the questions in the larger test less accurately than for the smaller test. This result did exceed the rate of forgetting results, but only by around 2% worse (3% for rate of forgetting and 2% attributable to test size). Students perform worse when there are more questions in the examination.

6.3 Discussion and interpretation of results

Overall, the results showed that typical university students were capable of exceeding the pass-mark of 75% in each of the conditions in our experiment. One day of concentrated study was enough to achieve these high scores, where the students memorized a total of 272 questions by rote. The results were marginally worse for the reworded questions, but the average still exceeded 75%. Rather impressively, seven subjects managed an overall score above 90% for the full 544 question battery off just two days of study.

This experiment gives an insight into how well the CQB can be rote-learned if the questions and answers are made available prior to testing, enough time is allocated and the amount of items to remember is reasonable. As we discuss, there will be a difference in the probabilities to rote memorize questions depending on the examination format and the time between examinations. Although half of the data points for the 272-questions were learned a week prior to the examination, it would be expected that learning would be distributed over time rather than concentrated exclusively the day prior to examination.

The rate of forgetting is shown to be a difference of 86.58% correct for the first test and then 83.16% for the same questions given a week later ($d = 0.53$), which is statistically significant. That the students could achieve over 80% a week later in the examination, based on only one day of learning, lends credence to the idea that rote-learning can be used to pass such an examination if the questions and answers are identical. However, the reality of the situation is less clear given the list of potential variables that can impact performance.

A week after learning the original set of questions, and having learned a new set of questions in the interim period, the students still managed to score on average above 80% correct. This shows that, at least in the short-term, the knowledge is committed to memory efficiently and can be recalled nearly as accurately as the day after learning. How much this statement would apply to longer time periods (from one month up to years) is impossible to know without a follow-up study.

Rewording and reshuffling the questions had little performance impact overall, with the mean differences in correctness ranging from 3.67% in the first test and only 2.54% in the second test. This indicates that simply rewording questions is not an effective strategy to discourage rote-learners as rote-learning alone produces knowledge that can be applied to these similar questions. However, for computed questions rote-learning will not result in knowledge that can be applied to similar situations unless the formulas themselves are memorized and can the student can apply them, even if the examination format does not specifically test for this if the answers are known prior to examination. So while knowledge transfer has arguably occurred, the examination format is not a practical application of the knowledge, and so transfer may ultimately be limited to MCQ and not necessarily to real-world applications.

A common theme in the post-test questionnaire filled out by the subjects was the topic of interference, whereby questions in the second set of questions were difficult to learn because of the similarity to the first set of questions. As the amount of questions to be learned increases, the effect of interference is bound to prevail to some degree, depending on various factors. In addition, when we asked the students about their learning strategies nearly all students reported having used mnemonics to some extent to remember and associate the correct answer to the question. Given that the reworded questions were answered nearly as well as the original questions, it diminishes the argument that students had learned by mnemonics alone. The knowledge was committed to memory and could be used in similar situations, but does not necessarily show that the knowledge could be transferred for the calculated answers.

Calculated answers differ from knowledge questions in that the student is expected to apply a formula to the question to derive the answer. However, rote memorization of question-answer pairs does not require knowledge of the equations. The students in the experiment managed to correctly answer many of the calculated questions without having been taught how to use the relevant formulae, which are not given in the questions or answers but are expected to have been learned in the classroom or from a book. The knowledge questions, by comparison, do show a transfer of knowledge as the learned material was used to correctly answer similarly worded questions. Transfer of knowledge refers to the notion that learned material (whether rote or otherwise) can be applied to similar situations. In this respect, rote learning can produce a level of learning that is useful in everyday scenarios outside of the learning environment. The close accordance in results between the original and reworded questions suggests that rote memorization of facts may lead to a quality of learning that is applicable in many situations.

Subjective performance assessment can be deduced by the confidence ratings that the students assigned to their performance. Figures 14 and 15 showed high correlations between overall test performance and the students' own predicted performances ($r = 0.81$ for the first test, $r = 0.83$ for the second test). Those students who were better prepared for the examination knew that they had performed well, whereas those students who predicted that they had performed poorly were correct in their self-assessment. It is plausible that the students who had formed better learning strategies were more confident in their capabilities as a result of this. All learning was standardised in the learning management system such that time spent learning could not be a determining factor for the results.

Furthermore, by asking the students about their hypothetical learning strategy and if they could use text books as well as the questions and answers for which they would be tested, students reported that they would mainly work with the multiple choice questions to orient their learning and use the textbooks just to consult some misunderstandings, relying basically on a rote-learning strategy to possess a higher probability to pass the exam. When rote learning is combined with textbook study or classroom study then the student is engaging with the material to understand it fully, and will result in better-retained and understanding of the syllabus than rote learning alone. The students used in the experiment would have liked to have read a book as well as rote learning to understand the material better. This qualitative finding showed an insight into the typical student mindset when faced with a MCQ examination rather than a written, essay-style examination. The learning strategies subtly change depending on what the demands for the examination are.

More concerning, perhaps, is the rapid rate of forgetting for such materials after the examinations have been passed. Casner, Heraldez and Jones (2006) found that learned materials from the PPL were only remembered well years later if they were relevant to the daily flight requirements of the pilots. That is to say, although pilots have to "master a formidable amount of aeronautical knowledge" (Casner, Heraldez and Jones, 2006, p. 72), if it is not used then it is quickly lost and replaced by the most pressing knowledge.

To estimate the probability that students can in fact rote-learn a battery of up to 10,000 CQB questions and answers, projections have been made from the sample of students used in this experiment. Previous literature has suggested that the relationship between number of questions and performance is non-linear, and likely to be a decaying power function. The likelihood of memorization is different depending on how close together the examinations are spaced and whether there are test resits or not. If the full CQB is to be memorized because the examinations are close together, then the probability of passing the test is very low. However, we have estimated that if learning is spread across a sufficiently large amount of time then it may be possible to pass the examinations by rote by learning around two modules at a time.

In the case of resits, candidates with prior exposure to the examination may be better prepared when resitting the examination than on their first attempt. This could be a function of accumulated knowledge in the interim and better preparedness for the examination contents. The availability of this option, though, varies from different member states where the maximum duration of time between resits and the number of resits available is different. For example, one state may allow students to re-take the same examination in "anytime or no fixed time limit" to as long as waiting "three calendar months". In Brian et al (1998)'s study, 88% of students managed to answer a specific post-test question correctly after a learning period , but after five months the students performed markedly worse on the same question- only 48% managed to answer it correctly. This shows that learning can be applied readily when testing and learning are close together, but over time the same knowledge disintegrates (the forgetting curve) until it cannot be applied as reliably as when the material was first learned.

In as much that our research has shown the efficacy of rote-learning, the study is limited by the timeframes used. To extend the findings even to 1,000 questions learned would require an outlay of time that is at least four-times longer than the 6-7 hours of study the University students undertook, not including the 1.5-hours for the first examination and 3-hours for the second examination. In order to maintain a stable level of examination performance, increasingly large amounts of study time would be required such that the time taken to learn 1,000 questions is not double the time taken to learn 500 questions if stable performance is to be achieved. The central question surrounds whether candidates will have sufficient time between examinations to learn an entirely new set of question and answer pairs while effectively forgetting previous examination preparations so that it does not interfere with learning the new knowledge. The results from Section 5.3 relied on the assumption that the examinations would be close together, and that memorization of the full CQB would be necessary to pass the full ATPL(A) license. Yet the advent of self-chosen examination dates and the possibility of resits make the issue less clear, which was examined in Section 5.4. In fact, a determined student may be able to pass these examinations by rote under the right conditions and mental preparation.

Students who employ rote-learning strategies are able to apply this knowledge to pass a MCQ examination, perhaps even when the questions and answers are presented differently than the presentation during learning. Some aspects of the examination promote and require verbatim recall of specific statistics, and knowledge of aviation that simply has to be memorized. In these instances, rote-learning of the CQB will produce hard-wired knowledge that can be recalled in MCQ format. For other examination styles such as essay-writing, then a wider knowledge of the subject material is required, but rote-learning of facts may assist in understanding and often is an inescapable part of learning. The publication of the CQB will assist students to rote-learn facts and materials, but there are several mitigating circumstances that can be used to discourage rote-learning as being the focus of study.

In terms of regulations and knowledge of procedures and essential flight statistics, the study supports the conclusion that some meaningful learning does occur with rote-learning. That the learning applied to reworded questions supports the idea that material is learned flexibly and lasts in memory up to a week, even when competing knowledge is introduced during that week. Rote learning of the CQB will result in memorization of vital flight information when presented in MCQ format.

7 Comparative risk analysis

Based on the results of the literature review and the scientific study, the Moebus team has conducted a comparative risk assessment from a consultancy point of view.

7.1 Introduction

Taking into account the results and discussion in the previous sections, the following aspects are important for a comparative risk analysis:

1. Whether the exact examination questions and answers are made available prior to testing,
2. Whether students have enough time for study, and
3. Whether students are able to take two to three-module examinations at a time over spreading the examination in several sittings (e.g. 5-6).

Suppose the full CQB is open to public and the exact questions and answers are used in actual theoretical examination, there are two variables which control the success of rote-learning; 1) "availability of enough study time prior to testing" and 2) "how the examination is conducted". In summary;

- The concept of publishing the full CQB questions and answers without any mitigating measures , using the exact same published questions and answers to actual Part FCL licensing examination and continuously exercising non-standardised examination procedures among the EASA States, could be considered as a potential risk.

7.2 List of possible risks based on the study results

As mentioned in the previous section, we have decided to consider these two main variables of "availability of study time prior to testing" and "examination procedure" as our main focus when we discuss publication of the CQB.

Based on the study results, students may face difficulty in achieving the pass score of 75% when they are faced with more than 600 questions to rote learn at a time. We have also decided to take into consideration increasing the number of questions in the CQB as a way to prevent students from successfully passing the examination by plain memorization of matching questions and answers as memorization capabilities decrease with increase dataset volume.

Furthermore, we have learned that there is a statistically significant effect (although absolute differences were rather small) of answering correctly when the original questions and answers are modified by re-wording, shuffling values and datasets and re-locating the correct answer.

During our literature search, we have looked into the EASA Part 66 theoretical examination procedure. In there, we have realised that some of Part 66 module examination employ partial essay questions. From this idea, we have also decided to explore the possibility of employing partial essay questions to Part FCL examination and whether it can be an effective way to prevent students from conducting only "surface learning".

Publication of full CQB with a standardised examination procedure

A first possibility is to apply a standardised examination procedure among all of the EASA member States for pilots' examinations. As discussed thoroughly in our scientific study, a longer study time being available between modulated testing schedules may eventually cast a risk. For example, when the time between module examinations is far enough apart, it gives students the opportunity to only rote-learn the given materials.. Although we could not investigate all of the EASA states and their individual procedures, some of the results of the survey reveal that each state employs different intervals. As a result, we should consider assessing different degrees of time intervals to be applied.

Based on the results of the literature review and the experimental study, Table 4 shows our recommendation for the maximum time period in which students should be allowed to sit their examinations to avoid the possibility of rote-memorization of answers. At 500 questions there should only be one day to sit the examination(s), whether it is two 250-question examinations or one 500-question examination. However, when the amount learned increases to 1,000 questions then there should not be longer than 1 week between examinations. This would deter students from learning a certain amount for one examination and then spending the next two weeks memorizing the questions for the next examination. We estimate that it would take longer than 3 months for a student, undertaking intensive study, to memorize 10,000 questions even if the examinations are split by modules and the student can self-select their examination time. Thus, the maximum time period in which students should be allowed to sit their examinations should not exceed 3 months.

Table 4: Recommended length of maximum time between examinations

Overall Examination Questions	Suggested Time Period
500	1 day
1,000	1 week
5,000	1 month
10,000	3 months
20,000	9 months

As with all of these recommendations, a shorter timeframe would further deter students from attempting rote learning, but these are given as the maximum recommended time between examinations. One day was estimated from the results as a recommended timeframe as it would require more than three days of concentrated rote-learning to attempt to memorize the questions, so if these questions were given in one day then students couldn't rote learn the remaining questions in the intervening period. One week would require more than seven days to memorize 1,000 questions based on the findings from the experiment, while a similar extension (136 questions/day) takes 36 days to memorize material. As has been shown throughout the document, the relationship between memorization and question size is not linear. Therefore it would require significantly more time than 136 questions read five times per day to memorize up to 5,000 questions. In light of this, one month for the examinations would provide a strong deterrent to rote learn. The estimations for 10,000 and 20,000 questions are based on similar principles such that students would require more than 100 days of intensive learning to attempt to memorize 10,000 questions.

Publication of full CQB with increased number of questions

The second possibility is to increase the number of questions in the CQB. According to our experimental results, a student may be able to achieve 75% of correct answers if he or she only has to memorise up to 600 question and answer pairs, and there is enough study time. Suppose a modulated examination environment exists in one of the EASA member States where students are allowed to take only 2-3 module-exams at a time and there are more than 100 hours of study time available, there is a potential risk of possibly memorizing all question and answer pairs from the fully published CQB.

However, based on the study results, we can speculate that students may not be able to reach a passing score of 75%, if the number of questions increased to 2 times or more from the current number of questions. The increase should be proportional to the current catalogue of questions for each module and each licence type. This estimation was derived based on below assumption:

1. Suppose students are allowed to take modulated examination for Part FCL theoretical examination and plan to take subjects 010 *Air Regulation* and 021 *Aircraft General Knowledge* at a scheduled date. For examination i.e. ATPL this would equal to a low set of questions being tested (see table below) for which the total CQB set would be rather easy to memorize in a short period of time. By increasing the amount of possible questions pure Rote learning can be prevented.

2. As previously mentioned the CQB contains roughly 3 to 4 times more questions than are being tested. In this example a student would only need to memorise between 372 to 496 question and answer pairs from the CQB. However, when the number of possible questions would be doubled, a student would have to memorise between 744 and 992 questions; according to our study results, rote-learning in excess of 744 questions may possibly prevent students from reaching passing score of 75%.

	Total Q	4 x actual Q	5 x actual Q	6 x actual Q	7 x actual Q	8 x actual Q	9 x actual Q	10 x actual Q	11 x actual Q	12 x actual Q
010	44	172	220	264	…	..	….	…	..	….
021	80	320	400	480	…	…	..	…	..	..
Subtotal	**124**	**502**	**720**	**744**	…	…	…	…	…	..
022	60	240	300	…	…	…	…	…	…	…
Subtotal	**184**	**740**	**1020**	…	….	…	…	…	..	…

Table 5: ATPL(A) subjects 010 and 021 question numbers and possible additional numbers of questions

3. Since the amount of questions varies from module to module small ones such as subject 031 and 032 will allow students to dedicate their studies to achieve a passing grade in between sittings.
4. As we have "rote learning up to 600 questions could possibly reach 75% correctness in testing", we have continued multiplying questions numbers until we reach 600 questions and more.
5. The result was that we have to increase actual question numbers for small modules by several factors order to finally pass 600-question level.

	Total Q	actual Q	2 x actual Q	3 x actual Q	4 x actual Q	5 x actual Q	6 x actual Q	7 x actual Q	8 x actual Q	9 x actual Q
031	**25**	100	200	300	400	500	600	700	800	900
032	**31**	124	248	372	496	620	744	868	992	1116
SubTotal	**56**	224	448	672	896	1120	1344	1568	1792	2016

Table 6: ATPL(A) subject 031 and 032 question numbers and possible additional numbers of questions

Publication of full CQB with reworded and change of datasets

The third consideration is to re-word or re-shuffle and exchange the datasets in the CQB when the questions and answers are open to the public. When we discussed about the FAA pilot license study by NASA, the research has shown that FAA pilot license students perform significantly worse in examination when data skills and knowledge is switched from the items they had learned to similar items. In our study, we have also found that publishing the CQB and using those questions and answers with re-wording or shuffling the dataset or choice answers could lower the students' possibility of passing examination by rote-learning (although absolute differences were rather small).

An example of answer choice shuffled will be such as:

(Original)
Q: FL 80, an OAT +6°C is measured. What will the temperature be at FL 130, if you consider the temperature gradient of the Standard Atmosphere?

a) -4°C **Correct**
b) +2°C Incorrect
c) 0°C Incorrect
d) -6°C Incorrect

(Answer choice re-shuffled)
Q: FL 80, an OAT +6°C is measured. What will the temperature be at FL 130, if you consider the temperature gradient of the Standard Atmosphere?

a) +2°C Incorrect
b) -6°C Incorrect
c) -4°C **Correct**
d) 0°C Incorrect

As seen in Table 4, the total questions in the examination modules 010 and 020 are 44. We have been also discussing that the CQB are said to contain 3-4 times more questions than actual number of examination questions. In other words, a student who is preparing to take two modularised exams of subjects 010 and 020 need to memorise around 502 question and answer pairs from the CQB. According to paragraph 5.2 of the Inferential Statistics section, our experimental results show that students who are encountered with re-worded or shuffled datasets questions and answers in their examination scores perform about 4 % lower on the 1st time and 3% lower at the 2nd attempt.

In other words, when 90% of 502 questions are exactly the same questions and answers and only 10% are modified:

(at 90% original, 10% modified)
502 * 0.9 = about 80% correct answers (according to Fig 16) = 360
502 * 0.1 = about 76% correct answers = 40
Total 400 questions correct or 80% of 502 questions or passing score

(at 70% original, 30 % modified)
502 * 0.7 = about 85% correct answers (reference Fig 16) = 297
502 * 0.3 = about 81% correct answers = 121
Total 418 questions correct or 83% of 502 questions or passing score

In summary, this option does not appear to prevent students from passing the examination by rote-learning, however in the interests of risk evaluation we decided to leave this criterion in the later risk assessment table.

Publication of the full CQB with partial essay questions

Lastly, in our literature review and search, we have learned that 'surface learning' is an important first step in the learning strategy to further 'deep' learn materials, as for example, memorization of information is critical for a flight student to master the wide variety of knowledge that is necessary to safely operate an airplane in the future.

However, when students are asked how to prepare for Multiple Choice Question (MCQ) examinations, they answer that they would mainly work with the multiple choice questions to orient their learning and use textbooks just to consult some misunderstandings. In other words, they would rely essentially on 'surface' learning strategy rather than a 'deep' one. This qualitative finding showed an insight into the typical student mindset when faced with a MCQ examination rather than a written, essay-style examination. The learning strategies subtly change depending on what the demands for the examination are.

The EASA Part 66 license examination employs partial essay questions in 3 modules out of 13 modules in B1 Turbine engine license, for example. This translates to a student who is spending 9.2% of time or 80 minutes out of total 865 hours working for the essay questions. Suppose the same percentage of essay questions is applied to Part FCL CPL (A) examination, it would mean that students spend 73 min out of 795 minutes in answering essay questions and the rest for MCQ.

1. Suppose we include essay questions that only counts 9.2% of examination mapping just like the EASA Part 66 theoretical examination, the rest of MCQ are the exactly the same questions and answers from the publicly available CQB in the future and modulated examination procedures are taken places at some EASA member states.

2. By memorizing the rest of 90.8% of questions from the CPL (A) 010 and 021 module examination or 456 questions (See table 5. 502 questions * 90.8% = 456 questions) using rote-learning, a student may be able to achieve as high as 78%.

3. Suppose the student completely failed the essay questions and achieved 78% correct in MCQ, it would mean that the student failed the examination.

Total question in 010 and 021	= 124
MCQ % (90.8%)	= 113
Essay % (9.2%)	= 11
MCQ score achieved 78%	= 88 questions correct
Essay score achieved 0%	= 0 questions or points correct
Total score achieved	88 out of 124 or scored 70%

4. However, IF the student can achieve a higher score in the essay, there is a potential risk that examination even including essays can be accomplished using rote learning.

MCQ score achieved 78%	= 88 questions correct
Essay score achieved 25%	= 2.75 questions or points correct
Total score achieved	90.75 out of 124 correct or scored 73%
MCQ score achieved 78%	= 88 questions correct
Essay score achieved 50%	= 5.5 questions or points correct
Total score achieved	93.5 out of 124 correct or scored 75%

As a result of the above estimation, when a student scores less than 50% correct in the essay questions, then it is less likely that the students can pass the examination by simply rote-learning the question and answer pairs.
In our evaluation, we decide to evaluate the only option of including essay questions according to the current practice done in the EASA Part 66 theoretical examination.

7.3 Risk Assessment methodology

In the following, two risk evaluation criteria of 'threat' and 'likelihood' are used and their assessment value and detailed description are discussed here.

Threat level	Description
3 (High)	There is a high likelihood that students will benefit from using of the publicly available questions and answers for rote-learning.
2 (Medium)	There is a medium likelihood that students will benefit from a public CQB to rote-learn questions and answers.

1 (Low)	There is a low likelihood that students will benefit from a public CQB to rote-learn questions and answers.
0 (None)	There is no likelihood that students will benefit from a public CQB to rote-learn questions and answers.

Table 7: Threat "Students to rote learn" classification

Likelihood level	Description
3 (High)	There is a high likelihood that students will pass the examination based solely on rote-memorization.
2 (Medium)	There is a medium likelihood that students will pass the examination by rote-learning.
1 (Low)	There is a low likelihood that students will pass the test by rote-learning.
0 (None)	There is no likelihood of students passing the exams by rote-learning.

Table 8: Likelihood "Students to pass exams by rote-learning" classification

Rank	Description
6 (Highest)	This is the highest rank of the risk category where the students may engage in rote learning, and at the same time, pass the examination by implementing this option.
5 (2nd highest)	This is the 2nd highest rank of the risk category where the students may engage in rote learning, and at the same time, pass the examination by implementing this option.
4 (3rd highest)	This is the 3rd highest rank of the risk category where the students may engage in rote learning, and at the same time, pass the examination by implementing this option.
3 (middle risk)	This is the middle risk category where the students may engage in rote learning, and at the same time, pass the examination by implementing this option.
2 (2nd lowest risk)	This is the 2nd lowest rank of the risk category where the students may engage in rote learning, and at the same time, pass the examination by implementing this option.
1 (Lowest risk)	This is the lowest rank of risk category where the students may engage in rote learning, and at the same time, pass the examination by implementing this option.

Table 9: Risk Ranking interpretation

7.4 Risk assessment table

To further determine the risk of pure rote learning, threat and likelihood were weighted and added to produce a ranking using expert opinion based on the scientific study.

	Observation	Threat	likelihood	Total	Rank
1	**Publishing full CQB** under current non-standardised examination practices among the EASA States. Exact Q & A that are published will be used for actual exams.	3	3	6	**6**
2	**Publishing full CQB** with new standardised examination practices among the EASA States. The "time-to-complete 13-module exams" in 2-3 days	2	1	3	**3**
3	**Publishing full CQB** with new standardised examination practices among the EASA States. The "time-to-complete 13-module(as example is ATPL) exams" in 7 days is for a point of discussing	2	2	4	**4**
4	**Publishing full CQB** with new standardised examination practices among the EASA States. The "time-to-complete 13-module exams" in more than 7 days	3	3	6	**6**
5	**Publishing full CQB**. But questions and answers are later re-worded, changed of datasets and shuffled of answer choices when actual exams are given.	3	2	5	**5**
6	**Publishing full CQB with partial essay questions.** Percentage is similar to EASA Part 66 license and the score student achieves from the essay questions are less than 50%.	2	2	4	**4**
7	**Publishing full CQB with increased #**s of questions (up to 3 times more or around 37,500)	1	1	2	**1**

Table 10: Risk Assessment Table

7.5 Risk assessment results

Our evaluation was conducted on the basis of fully publishing the CQB. From there, we assigned different set of variables to further evaluate likeliness of students attempting to rote learning as well as passing theoretical examination through only rote-learning.

After rankings were assigned to table 9, we can re-organise that table according to high risks to no risk as following:

Risk (high to low)	Observation	Threat	Likelihood	Total
1 Highest Risk	**Publishing full CQB** under current non-standardised examination practices among the EASA States. Exact Q & A that are published will be used for actual exams.	3	3	6
1 Highest Risk	**Publishing full CQB** with new standardised examination practices among the EASA States. But "time-to-complete i.e. ATPL 13-module exams" in **8 days or more (time is initial point of discussion).**	3	3	6
2 High Risk	**Publishing full CQB**. But questions and answers are later **modified** by such as re-worded, changed of datasets and shuffled of answer choices when actual exams are given.	3	2	5
2 High Risk	**Publishing full CQB with increased #**s of questions (up to 2 times more or around 20'000 to 30'000) in order to prevent pure rote learning	1	1	2
3 Medium Risk	**Publishing full CQB** with new standardised examination practices among the EASA States. The "time-to-complete i.e. ATPL 13-module exams" in **7 days (time is initial point of discussion)**	2	2	4
3 Medium Risk	**Publishing full CQB with partial essay questions.** Percentage is similar to EASA Part 66 license and the score the student achieves from the essay questions are less than 50%.	2	2	4
4 lowest Risk	**Publishing full CQB** with new standardised examination practices among the EASA States. The "time-to-complete i.e. ATPL 13-module exams" **within 2 days (time is initial point of discussion)**	2	1	3

Table 11: Re-organised risk assessment table according to risk value

8 Conclusions

Under the current prevailing examination procedures throughout the EASA member states, no clear recommendation for either publishing or not publishing can be given without any further accompanying actions. However, the mitigating actions described below are weighted and represent a possible way forward in the decision making process whether to publish the CQB or not;

Our risk assessment was conducted according to the results of the literature review and the scientific study that provided us with the following important findings:

Overall, the results showed that typical university students were capable of exceeding the pass-mark of 75% in each of the conditions in our experiment. One day of concentrated study was enough to achieve these high scores, where the students memorized a total of 272 questions by rote. The results were marginally worse for the reworded questions, but the average still exceeded 75%. Rather impressively, seven subjects managed an overall score above 90% for the full 544 question battery off just two days of study.
This experiment gives an insight into how well the CQB can be rote-learned if the questions and answers are made available prior to testing, enough time is allocated and the amount of items to remember is reasonable. However, there will be a difference in the probabilities to rote memorize questions depending on the examination format and the time between examinations.
The size of the dataset to be memorized is a determining factor as to whether an individual would either attempt memorization, or whether it is even within the realms of possibility. As the data set size increases, the potential payoff for memorization arguably decreases.
To estimate the probability that students can in fact rote-learn a battery of up to 10,000 CQB questions and answers, projections have been made from the sample of students used in this experiment. Previous literature has suggested that the relationship between number of questions and performance is non-linear, and likely to be a decaying power function. The likelihood of memorization is different depending on how close together the examinations are spaced and whether there are test resits or not. If the full CQB is to be memorized because the examinations are close together, then the probability of passing the test is very low. However, we have estimated that if learning is spread across a sufficiently large amount of time then it may be possible to pass the examinations by rote by learning around two modules at a time.
Another interesting result was found by asking the students about their hypothetical learning strategy and if they could use text books as well as the questions and answers for which they would be tested. Students reported that they would mainly work with the multiple choice questions to orient their learning and use the textbooks just to consult some misunderstandings, relying basically on a rote-learning strategy to possess a higher probability to pass the exam. When rote learning is combined with textbook study or classroom study then the student is engaging with the material to understand it fully, and will result in better-retained and understanding of the syllabus than rote learning alone. The students used in the experiment would have liked to have read a book as well as rote learning to understand the material better. This qualitative finding showed an insight into the typical student mindset when faced with a MCQ examination rather than a written, essay-style examination. The learning strategies subtly change depending on what the demands for the examination are.

In terms of regulations and knowledge of procedures and essential flight statistics, both the literature review as well as the experimental study support the conclusion that some meaningful learning does occur with rote-learning. That the learning applied to reworded questions supports the idea that material is learned flexibly and lasts in memory up to a week, even when competing knowledge is introduced during that week. Rote learning of the CQB will result in memorization of vital flight information when presented in MCQ format.Based on the above findings, we have listed 8 risk evaluation options to be assessed whether there is a potential risk of employing rote-learning as a strategy for flight students to score enough correct answers to pass the examination. Applied was the condition of the full CQB published, no changes in actual examinations implemented and non-standardised examination procedures as it is currently the case among the EASA member states.

After assigning the values such as 3 for high risk, 1 for low risk or 0 for no risk, we have come to the following risk ranking based on those 7 options; whereas the highest risk is listed first.

1. **Publishing full CQB** under current non-standardised examination practices among the EASA States. Exact Q & A that is published will be used for actual exams.

2. **Publishing full CQB** with new standardised examination practices among the EASA member States. But "time-to-complete 13-module exams i.e. ATPL" in **8 days or more (time is initial point of discussion).**

3. **Publishing full CQB** but questions and answers are later **modified** by re-worded, changed of datasets and shuffle of answer choices when actual exams are given.

4. **Publishing full CQB** with increased number of questions (up to 2 times more or around 20'000 to 30'000).

5. **Publishing full CQB** with new standardised examination practices among the EASA States. The "time-to-complete 13-module exams i.e. ATPL" in **7 days (time is initial point of discussion).**

6. **Publishing full CQB** with **partial essay questions**. Percentage is similar to EASA Part 66 license and the score student achieves from the essay questions are less than 50%.

7. **Publishing full CQB** with new standardised examination practices among the EASA States. The "time-to-complete 13-module exams i.e. ATPL" **within 2.5 days (time is initial point of discussion).**

Based on the results of our study, it will be highly improbable for a student to attempt memorising all of the ca. 10,000 to 15,000 question and answer pairs at a time even with an abundance of time being allocated to study. However, we have also concluded from our study that, it is possible for a student to pass the Part FCL theoretical examination when the full CQB is published AND each EASA Member State practices non-standardised examination procedures by allowing modulated subject examination within variable time frames.

As mentioned in the beginning of this chapter, under the actual current examination practices, any straight recommendation either way is difficult to make. However, when applying mitigating measures such as standardisation of testing procedures as well as a limited time frame in which a student is to complete all subjects, the risk of regurgitation of information only is greatly reduced. Likewise, the CQB could be increased to such an extent at which memorization of the data available becomes futile and students are discouraged from rote-learning only.

To this extent, we recommend the agency to develop such EASA-wide standardised examination procedures such as enrolment only through a certified flight training organization, and a very limited time frame in which the examination is to take place regardless of modular or integrated training. Further the CQB should be increased to at least double the volume and should be centralized so as to ensure that each EASA member state is sourcing the same test questions. These and other proposed standards should be released for consultation by the member states which is to be eventually implemented throughout all EASA member states prior to publishing the CQB. This may require some further evaluation of current national supervisory agencies procedures and their certification process in order to derive a harmonisation of EASA-wide standards fulfilling the mitigating requirements to publish the CQB.

10 Applicable Documents (AD) and Reference Documents (RD)

APPLICABLE DOCUMENTS (AD)

Specifications attached to the Invitation to Tender EASA.2008.OP.23 Impact assessment of the publication of questions of theoretical examinations for Part 66 and Part FCL

Appendix 1 to Terms of reference: Requirements regarding theoretical knowledge examinations for the issue of pilot licences

Reference Documents (RD)

Anderson, L.W., Krathwohl, D.R:, Airasian, P.W., Cruikshank, K.A., Mayer, R.E., Pintrich, P.R., Raths, J., & Wittrock, M.C. (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. New York: Longman.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 2, pp. 89–195). New York: Academic Press.

Ausubel, D.P., Novak, J.D., & Hanesian H. (1978). Educational psychology: A cognitive view, (2nd ed.). New York: Holt, Rinehart & Winston.

Baddeley, A.D. (1999). Essentials of human memory. Philadelphia: Psychology Press.

Baddeley, A.D., & Hitch, G. (1974). Working memory. In Ormrod J.E., Human Learning. Pearson Education. 4th ed.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, 37(3), 379-384.

Baxter, G.P., Elder, A.D., & Glaser, R. (1996). Knowledge-based cognition and performance assessment in the science classroom. *Educational Psychologist*, 31:2, 133-140.

Beckwith, J.B. (1991). Approaches to learning, their context and relationship to assessment performance. *Higher Education,* 22:11, 17-30.

Billing, D. (2007). Teaching for transfer of core/key skills in higher education: Cognitive skills. *Higher Education*, 53, 483-516.

Boyd, K.T. (1990). Airline Transport Pilot, second edition, Iowa State University Press.

Brady, T.F., Konkle, T., Alvarez, G.A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences,* 105:38, 14325-14329.

Brian, E.M., Kathryn, L.L., & Karen, S.O. (1998). Why Johnnie can't apply neuroscience: Testing alternative hypotheses using performance-based assessment. *Advances in Health Sciences Education*, 3, 165-175.

Bruer, J.T. (1994). Schools for thought. MIT Press.

Carpenter, S.K., Pashler, H., Wixted, J.T., & Vul, E. (2007). The effects of tests on learning and forgetting. In press - *Memory & Cognition*.

Casner, S.M., Jones, K.M., Puentes, A., & Irani, H. (2003). FAA pilot knowledge tests: learning or rote memorization? *NASA*, TM—2004–212814.

Casner, S.M., Heraldez, D., & Jones, K.M. (2006). Retention of aeronautical knowledge. *International Journal of Applied Aviation Studies*, 6:1, 71-99.

Cerretta, T.R. (2000). US air force pilot selection and training methods. *Report No. AFMC 99-273 ASC 99-1463*.

Challis, B. H. (1993). Spacing effects on cued-memory tests depend on level of processing. Journal of Experimental Psychology: Learning, Memory, and Cognition, 19: 22, 389-396.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, 49, 997-1003.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior,* 11, 671-684.

Crowder, R. G. (1976). Principles of learning and memory. Oxford, England: Lawrence Erlbaum.

Dauntless-soft:

Eysenck, H.J. (1997). Dimensions of personality. Transaction Publishers.

Frick, R.W. (1996). The appropriate use of null hypothesis significance testing. *Psychological Methods*, 1:4, 379-390.

Flouris, T. (2001). The impact of ground schools in a collegiate aviation program on FAA written examination scores*. University Aviation Association Annual Conference*, Nashville, TN.

Gazzaniga, M.S., Ivry, R., & Mangun, G.R. (2002). Cognitive Neuroscience: The Biology of the Mind. W.W. Norton, 2nd ed.

Greene, R. L. (1989). Spacing effects in memory: Evidence for a two-process account. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15:3, 371-377.

Gunter, B., Barry, C., & Clifford, B.R. (1981). Proactive interference effects with television news items: Further evidence. *Journal of Experimental Psychology: Human Learning and Memory*, 7, 480-487.

Haas, J. (2006). Occupational Licensing Versus Company-led Training*. XVI ISA World Congress of Sociology*, Durban, Republic of South Africa, July 23-29.

Hunter, D.R. (2003). Measing general aviation pilot judgment using a situational judgment technique. *The International Journal of Aviation Psychology*, 13:4, 373-386.

Jensen, R. S. (1996). *Pilot judgment and crew resource management*. Brookfield, VT: Avebury.

Jensen, R. S., & Benel, R.A. (1977). *Judgment evaluation and instruction in civil pilot training*. (Tech. Rep. No. FAA-RD-78-24). Washington, DC: Federal Aviation Administration.

Kember D. (1996). The intention to both memorise and understand: Another approach to learning? *Higher Education*, 31, 341-354.

Landauer, T.K. (1986). How much do people remember? Some estimates of the quantity of learned information in the long-term. *Cognitive Science*, 10, 477-493.

Lundenberg, M.A. & Fox, P.W. (1991). Do laboratory findings on test expectancy generalize to classroom outcomes? *Review of Educational Research*, 61:1, 94-106.

Luria, A.R. The mind of a mnemonist. New York: Avon.

Marton, F., Dall'Alba, G., & Kun, T. L. (1996). Memorizing and understanding: The keys to the paradox? In D. A. Watkins & J. B. Biggs (Eds.), *The Chinese learner: Cultural, psychological, and contextual influences* (pp. 69–84). Hong Kong: Comparative Education Research Centre.

Mavis, B.E., Lovell, K.L., Ogle, K.S. (1998). Why Johnnie can't apply neuroscience: Testing alternative hypotheses using performance-based assessment. *Advances in Health Sciences Education*, 3, 165-175.

Underwood, B.J. Interference and forgetting. *Psychological Review,* 1957, *64*, 49-60.

Mayer, R.E., & Wittrock, M.C. (1996). Problem-solving transfer. *Handbook of Educational Psychology*.

Mayer, R.E. (2001). Rote versus meaningful learning. *Theory Into Practice*, 41:4.

Mozer, M.C., Howe, M., & Pashler, H. (2004). Using testing to enhance learning: a comparison of two hypotheses. *Proceedings of the Twenty Sixth Annual Conference of the Cognitive Science Society,* 975-980. Hillsdale, NJ: Erlbaum Associates.

Novak, J.D., & Gowin, D.B. (1984). Learning how to learn.

Norman, G.R., & Schmidt, H.G. (1992). The psychological basis of problem-based learning: a review of the evidence. *Academic Medicine*, 67, 557-562.

O'Hare, D. (2001). Aeronautical decision making: Metaphors, models, and methods. In P. S. Tsang & M. A. Vidulich (Eds.), *Principles and practices of aviation psychology* (pp. 201–237). Mahwah, NJ: Lawrence Erlbaum Associates, Inc, 2003.

Ormrod, J.E. (2004). Human Learning. Peterson Education. 4th ed.

Ormrod, J.E., Ormrod. R.K., Wagner, E.D., & McCallin, R.C. (1988). Reconceptualizing map learning. *American Journal of Psychology,* 101, 425-433.

Ramsden, P. (1992) Learning to teach in higher education. London: Routledge.

Sax G., & Collet, L.S. (1968). An empirical comparison of the effects of recall and multiple-choice tests on student achievement. *Journal of Educational Measurement*, 5:2, 169-173.

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education*, 35, 453-472.

Scouller, K. M. & Prosser, M. (1994). Students' experiences in studying for multiple choice question examination. *Studies in Higher Education*,19:3, 267-279.

Shulman, L.S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review,* 57*,* 1-22.

Simorn, H.A., & Feigenbaum, E.A. (1964). An information-processing theory of some effects of similarity, familiarization, and meaningfulness in verbal learning. *Journal of Verbal Learning & Verbal Behavior*, 3, 385-396.

Sternberg, Robert J. (2006). Cognitive psychology fourth edition. Thomson Wadsworth, 219.

Theoretical Knowledge Manual (2001). Flight Performance and Planning 1, Oxford Aviation and Jeppesen.

Thomas, P.R., & Bain, J.D. (1984) Contextual dependence of learning approaches: the effects of assessments, *Human learning*, 3, 227-240.

Tweed, R.G., & Lehman, D.R. (2002). Learning considered within a cultural context: Confucian and Socratic approaches. *American Psychologist*, 57:2, 89-99.

Verheijen, F.M. (2002). Flight training and pilot employment. Msc Thesis in Air Transport Management, London.

Ward, P.J. & Walker, J.J. (2008). The influence of study methods and knowledge processing on academic success and long-term recall of anatomy learning by first-year veterinary students. *Anatomical Sciences Education,* 1:2, 68-74.

Webb, G. (1997). Deconstructing Deep and Surface: Towards a Critique of Phenomenography. *Higher Education,* 33:2, 195-212.

Wiegmann, D A., & Shappell, S.A. (1997). Human factors analysis of postaccident data. *International Journal of Aviation Psychology,* 7, 67–82.

Wingfield, A., & Byrnes, D.L. (1999). The psychology of human memory. Harvest Books.

Wilson, P.T., & Anderson, R.C. (1986). What they don't know will hurt them: The role of prior knowledge in comprehension. In J. Orasanu (Ed.), *Reading Comprehension: From Research to Practice* (31-48). Hillsdale, New Jersey: Lawrence Erlbaum.