Effect of risk attitudes on recall of assignment statements (part 2 of 2)

Experiment performed at the 2011 ACCU Conference
First published in CVu vol. 2x no. x

Derek M. Jones
derek@knosof.co.uk

1 Introduction

This is the second of a two part article describing an experiment carried out during the 2011 ACCU conference. The first part was published in the previous issue of C Vu.[9] This second part discusses the remember/recall assignment statement component of the experiment; see part 1 for a discussion of the experimental setup.

Coding guidelines sometimes recommend that words rather than non-words be used in identifiers and sometimes recommend that prefixes be added to identifiers to denote some property (e.g., `E_` to denote a member of an enumerated type). The identifiers used in the assignment problem had letter sequences designed to investigate subject performance differences caused by these two factors.

The format of the task subjects' were asked to perform in this part of the experiment is identical to the memory for assignment statements portion of the experiment performed at the 2006 ACCU conference,[7] with the one difference that four assignment statements rather than three were used as the information to be remembered. See the write-up of that experiment for more details that are omitted here.

One of the results from the 2006 experiment (and earlier experiments measuring memory performance) was that subjects did not make enough mistakes to find any statistically significant correlations between identifier attributes and recall performance. It was hoped that increasing the number of assignment statements from three to four would increase the load on short term memory and result in more mistakes being made.

2 Characteristics of human memory

The human *short term memory* subsystems are a gateway through which all conscious input data must pass. They have a very limited capacity and because new information is constantly streaming through them the accuracy of a particular piece of information rarely is rarely maintained for very long. Information in short term memory is either quickly lost or stored in another, longer term, memory subsystem.

The following are some of the factors that studies have been found to effect subject recall performance of recently seen lists of information. These factors are expected to have some impact on subject performance in this experiment.

- People pay particular attention to the initial part of a word[3, 10] (this enables them to start looking up a word in the mental lexicon while its remaining sounds are being heard).
- A decrease in word list recall performance for similar sounding words.[2, 4] It is believed that the similarity causes confusion between the various word sound sequences and a subsequent failure to correctly retrieve the original information.
- The extent to which the information to be remembered is already stored in longer term memory subsystems (i.e., known letter sequences such as words).
- The time delay between seeing the information and having to recall it (because the remembered information degrades over time),
- A capacity limit on the total amount of information that can be remembered and shortly afterwards recalled or recognized,

The above findings suggest that subject performance could exhibit one or more of the following patterns (some of which drive performance in different directions):

- performance will be better on problems where the identifiers have different initial letters, compared to problems where the initial identifier letters are the same.

  This could occur both because people have been found to pay more attention to the start of a word and if subjects attempt to optimise performance by remembering just the first letter there is often a STM capacity advantage; while both the spoken form of single letters are represented by a single syllable (except *w* which contains two) and each of the letter sequences used was pronounceable as a single syllable the sound duration of the word syllable is longer.

- performance will be worse on problems where the identifiers have similar spoken forms, compared to problems where the identifier have dissimilar spoken forms,
- performance will be better on problems where the identifiers are known words, compared to problems where the identifiers are non-words.

Source code identifiers contain a variety of different kinds of character sequences. Some are recognizable words or phrases, some abbreviated forms of words or phrases, while others have no obvious association with any language known tot he reader (e.g., they may be acronyms that are unknown to the reader). Reading involves converting these character sequences to sounds and it is to be expected that subjects' memories of an identifier will be sound based, rather than vision based.

Part of the problem subjects are asked to answer involved indicating the identifier that did not appear in a previously seen list of assignment statements. This is a recognition problem, while remembering the value assigned in a recall problem. Studies have found that recognition and recall memory have different characteristics.[1]

3 The assignment problems

The problems and associated page layout were automatically generated using a C program and various awk scripts to generate troff, which in turn generated postscript. The identifier and constant used in each assignment statement was randomly chosen from the appropriate set and the order of the assignment statements (for each problem) was also randomized. The source code of the C program and scripts is available from the experiments web page.[8]

Due to a fault in the generation script the first 10 problems for each subject all used sets of identifiers where the last two letters of each set of identifiers were the same. The intent was that the randomisation algorithm be applied to the choice of identifiers used for all problems.

3.1 Selecting identifiers and integer constants

A sufficient number of letter sequences were created so that most subjects would not encounter the same sequence more than once. In all 50 different words and 50 different nonwords were used (see the following list). Given this list the same identifier sequence will repeat after every set of 20 problems seen by a subject.

The identifier letter sequences were designed to investigate differences in subject performance caused by two factors:

- all identifiers in the assignment list being words or all pronounceable English non-words (no checks were made for English nonwords that were words in other languages and possible known to subjects),
- all identifiers in the assignment list starting with or not starting with the same letter (in the latter case all subsequent letters were the same and so the words rhymed).

For simplicity identifiers consisted of a sequence of three letters having the pattern CVC. The following lists the sets of five identifiers used in the assignment problem. Assignment problems were created in groups of 20. Each group of 20 used all of the identifiers appearing in one row. The identifiers used in each assignment problem were selected by randomly choosing a row previously unused for a given subject. Four of the identifiers in the row were randomly selected to be used in the list of the four assignment statements to be remembered. The fifth identifier was used as the *not seen* identifier.

```
cat mat hat pat bat
hen pen men ben yen
hop pop top mop cop
din pin sin kin tin
cub rub tub hub pub
dat lat wat gat tat
gen ren sen ven nen
```

```
dop gop vop rop pop
nin rin zin cin lin
fub lub wub mub bub
dig dog dad den dot
lot lip led lap lid
pin pat pod peg pen
sat sir sum sod sad
wag wit won web wig
fot fis fup fep fik
kam kig kus kos kuk
ras rit rus roz ral
tid tol tep tul teb
vib vok vup vek vot
```

The word "pop" accidently appeared in both a word and nonword sequence. The answers to problems where "pop" appeared in a non-word list were not included in the analysis described below.

The nonwords have a variety of characteristics, including: the nonword *cin* sounding like the word *sin*, *fot* could be remembered as *foot* + CVC pattern, *roz* rozzer slang for policeman (at least in British English) or abbreviation for the name Rosalyn.

3.2 Selecting integer constants

The experimental factor under investigation involves attributes of identifiers and the impact of other kinds of information (mostly involving integer constants) on subject performance needs to be minimized. A good approximation to short term memory requirements is the number of syllables contained in the spoken form of the information. Choosing single digit integer constants containing a single syllable minimises their impact on short term memory load.

The integer constants chosen were 4, 5, 6, 8, and 9 (the digit 7 was not used because its English spoken form has two syllables). To within an order of magnitude they all have the same frequency of occurrence in source code.[6]

4 Threats to validity

Experience shows that software developers are continually on the look out for ways to reduce the effort needed to solve the problems they are faced with. Because each of the experimental problems seen by subjects has the same format it is possible that some subjects will detect what they believe to be a pattern in the problems which they then attempt to use to improve their performance.

While the general format of the problem used commonly occurs during program comprehension, the mode of working (i.e., paper and pencil) is rarely used these days; source code is invariably read within an editor and viewing is controlled via a keyboard or mouse. Referring back to previously seen information (e.g., assignment statements) requires pressing keys (or using a mouse) and having located the sought information additional hand movements (i.e., key pressing or mouse movements) are needed to return to the original source location. In this study subjects were only required to tick a box to indicate that they *would refer back* to locate the information. The cognitive effort needed to tick a box is probably a lot less less than would be needed to actually refer back. Studies have found[5] that subjects make cost/benefit decisions when deciding whether to use the existing contents of memory (which may be unreliable) or to invest effort in relocating information in the physical world. It is possible that in some cases subjects ticked the *would refer back* option when in a real life situation they would have used the contents of their memory rather than expending the effort to actually refer back.

Each identifier appeared once per set of 100 assignment statements. Based on expected subject performance, it was anticipated that most identifiers would be seen once, with only a few identifiers being seem twice by the faster subjects. Thus any learning of individual problem identifier sets by the faster subjects was not expected to have a significant impact on the results.

While subjects were told they are not in a race and that they should work at the rate at which they would normally process code, it is possible that some subjects ignored this request. A consequence of this is that the distribution in the number of problems answered, and perhaps the accuracy of the results, may be different than would occur if all subjects followed the instructions given to them.

If subjects randomly guess answers to questions that they cannot recall answers to, then (given that only five possible numeric values were used and no value occurred more than once in the same problem):

- if a subject knew no answers and randomly guessed the four answers, then an average of 0.88 of a question would be guessed correctly (total 0.88 correct per problem),
- if a subject knew one answer and randomly guessed the other three answers, then an average of 0.875 of a question would be guessed correctly (plus one known answered correctly; total 1.875 correct per problem),
- if a subject knew two answers and randomly guessed the other two answers, then an average of 0.67 of a question would be guessed correctly (plus two known answered correctly; total 2.67 correct per problem),
- if a subject knew three answers and randomly guessed the other one answers, then an average of 0.5 of a question would be guessed correctly (plus three known answered correctly; total 3.5 correct per problem).

4.1 Ecological validity

For the results of this experiment to be applicable to professional developer performance it is important that subjects work through problems at a rate similar to that which they would process source code in a work environment. Subjects were told that they are not in a race and that they should work at the rate at which they would normally process code. Experience from previous experiments has shown that the competitive instinct in some developers causes them to ignore the work rate instruction and attempt to answer all of the problems in the time available. To deter such behavior during this experiment the problem pack contained significantly more problems (28 in total) than subjects were likely to be able to answer in the available time (one subject answered all problems and two subjects all but one).

The structure of the problem follows a pattern that is often encountered when trying to comprehend source code: see information (and try to remember some of it), perform some other task and then perform a task that requires making use of the previously seen information.

Considering the experimental context as a whole, the constant repetition of exactly the same kind of activity rarely occurs in program development. The constant repetition provides an opportunity for learning to occur, e.g., subjects have the opportunity to tune their performance for a particular kind of problem. The issue of learning and problem solving strategies used by subjects is discussed below.

5 Results

It was estimated that each subject (20 expected, on the day 30) would be able to answer 20 problem sets (on the day 20.0, sd= 7.7) in 20-30 minutes (on the day 20 minutes). Based on these estimates the experiment would produce 2000 (on the day 2980) individual answers. Table .1 gives a summary of the kinds of the results.

Table .1: Summary of results for this and the 2006 experiment. The Total column is summed over all answers while the "By subject" column gives the subject mean (standard deviation in brackets) values. The Correct/Incorrect recall and *would refer back* percentages are calculated using the total number of answers in those three cases. The "Not seen (incorrect)" values are calculated using the total number of the *not seen* answers.

	Total	By subject	2006 Total	2006 By subject
Correct recalls	1379 (57.2%)	59.5% (29)	747 (57.9%)	60.4% (26)
Incorrect recalls	397 (16.5%)	15.5% (15)	143 (11.1%)	9.6% (10)
Would refer back	634 (26.3%)	25.0% (28)	400 (31.0%)	30.0% (26)
Not seen (incorrect)	54 (9.5%)	8.4% (10)	24 (6.0%)	5.7% (8)

The average amount of time taken to answer a complete problem was 60 seconds. The format of the experiment means that no information is available on the amount of time invested in trying to remember information, answering the questionnaire sub-problem, and then thinking about the answer to the recall sub-problem (i.e., the effort break down for individual components of the problem). While STM recall performance drops very quickly after the information is no longer visible (studies have found below 10% correct within around 8 seconds in many situations[2]). Even the fastest subject took over 25 seconds per complete problem and so recency effects are likely to be minimal.

5.1 Subject strategies

Feedback from subjects who took part in the previous experiments highlighted the use of a variety of strategies to remember information for each assignment problem. The analysis of the threats to validity in some of those experiments has discussed the question of whether subjects traded off effort on the filler task in order to perform better on the assignment problem, or carried out some other conscious combination of effort allocation between the subproblems. To learn about strategies used during this experiment, after 'time' was called on problem answering, subjects were asked to list any strategies they had used (a sheet inside the back page of the handout had been formatted for this purpose).

The responses to the strategies question generally contained a few sentences. Only a few responses involved the questionnaire part of the problem, with subjects saying they answered honestly.

The strategies listed consisted of a variety of the techniques people often use for remembering lists of names or numbers. For instance, number word associations, continuous repetition of information, creating memorable sentences merging words, reordering the sequence presented into a regular pattern (e.g., alphabetical), inventing short stories involving the words and numbers and associating the information with musical sound patterns. One subject gave remembering the first letter as a strategy.

From the replies given it was not possible to work out if subjects give equal weight to answering both parts of the problem, or had a preference to answering one part of the problem.

No subject listed a strategy that was based on the visual appearance of the identifiers or numbers, although several subjects said they tried to associate images with the identifiers and numbers.

5.2 Recall/would refer back performance

This subsection treats the value recall and *would refer back* answers as a single subproblem for the purpose of analyse. The *not seen* answers are treated as a different subproblem and is discussed in the following subsection.

Figure .1 is a scatter plot of the percentage of correct/incorrect recall and *would refer back* answers given by subjects for each of the four kinds of identifiers. The straight line is the set of points along which the values on the two axis sum to 100%.

The ideal subject behavior is for all points in the correct recall vs. *would refer back* scatter plot to lie along the straight line, i.e., subjects would either give the correct answer or they *would refer back*; see top-left of Figure .1.

The scatter plot of correct vs. incorrect recall answers shows some points along the 100% line, i.e., in some cases subjects were either right or wrong and never gave a *would refer back* answer (perhaps these subjects are willing to take more risks); see bottom-left of Figure .1.

The scatter plot of incorrect recall against *would refer back* answers shows some points along the 100% line, i.e., in some cases subjects either give incorrect or *would refer back* answers (perhaps these subjects are willing to take more risks); see top-right of Figure .1.

Self-knowledge, metacognition, is something that enables a person to evaluate the accuracy of the memories they have. Subjects who give many incorrect answers do not accurately evaluate the state of their own memories of previously seen information (i.e., they overestimated the accuracy of their memories). It is also possible that subjects who gave many *would refer back* answers also showed poor metacognitive performance (i.e., they underestimated the accuracy of their memories and would have mostly given correct answers had they risked a numeric answer). However, it is not possible to evaluate this possibility based on the available

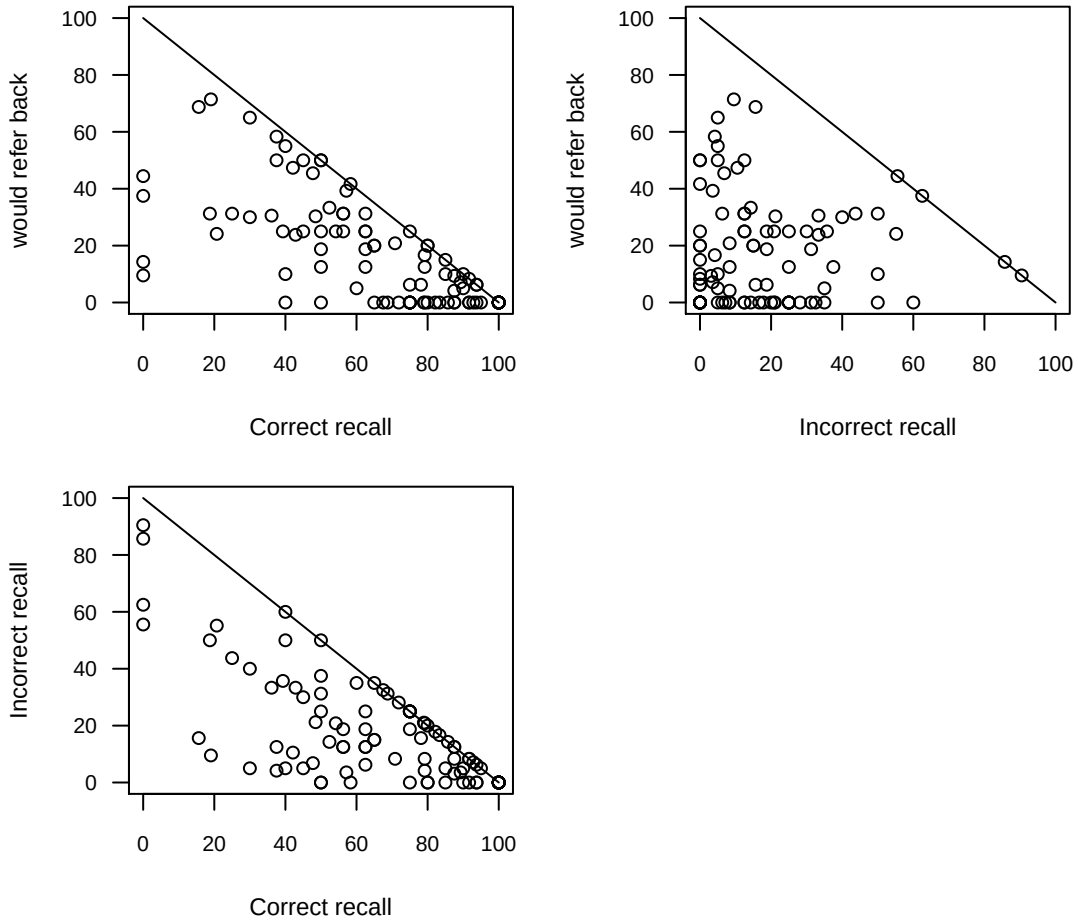

Figure .1: Scatter plots of various combinations of percentage correct recall, incorrect recall and *would refer back* answers plotted against each other. The straight lines are the set of points along which the values on the two axis sum to 100%.

data.

5.3 Not seen performance

This subsection treats the *not seen* answers as a single subproblem for analyse. There were 570 *not seen* answers of which 9.5% were incorrect. Subjects who randomly chose an answer would achieve a 80% failure rate.

Figure .2 is a scatter plot of the percentage of correct/incorrect recall and *not seen* answers given by subjects for each of the four kinds of identifiers. The results are dominated by the high percentage of correct answers.

5.4 Effects of different identifier naming patterns

This experiment has a two factorial design with two levels. The factors are "is word" and "same first" and the levels are TRUE/FALSE. All four combinations of identifiers were used, enabling the interaction between them to be investigated.

The statistical technique used to analyse the results is ANOVA (analysis of variance). For details of the analysis see the source code of the R program used to analyse the data available for download[8] along with the (anonymous) data extracted from subject answers.

The subject data was split into two groups of response variables, with "is word" and "same first" used as the

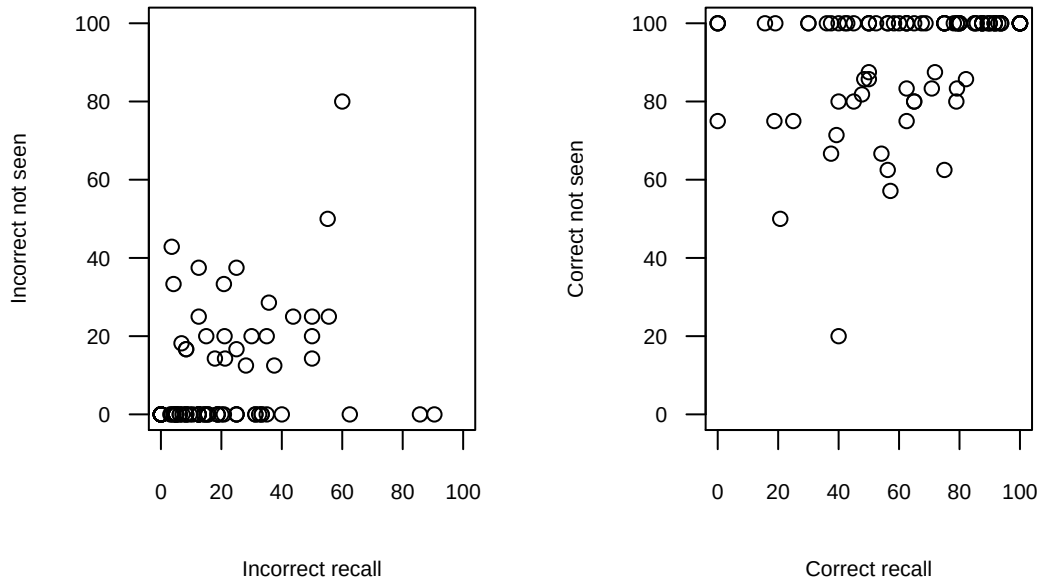

Figure .2: Scatter plots of percentage of incorrect *not seen* vs. percentage incorrect recall answers and correct *not seen* vs. percentage correct recall answers.

predictor variables in both cases. One group of response variables was percentage of recall correctness and percentage of *would refer back* answers and the other response variable was percentage of incorrect *not seen* answers.

For some combinations of identifier attributes some subjects gave a small number of answers. Answer counts were converted to percentages and to prevent a small number of answers potentially giving a non-representative value any subject data set containing less than five (recall + *would refer back*) answers or less than three *not seen* answers were excluded from the analysis.

5.4.1 Recall and would refer back

Analysis of the data found that differences in the identifier attributes "is word" or "same first" were not good predictors of the percentage of correct, or incorrect, recall answers or percentage of *would refer back* answers (p-values for correct recall were "is word" 0.34 and "same first" 0.80).

5.4.2 Not seen

Analysis of the data found that differences in the identifier attribute "same first" was a good predictor of subject *not seen* performance; ANOVA p-value=0.038, below the widely used maximum of 0.05 for level of significance.

For *not seen* answers the mean percentage of incorrect answers was over twice as high when all of the assignment identifiers started with the same letter compared to when they all started with different letters; 5.9% (sd 6.8%) and 2.6% (sd 4.5%) error rates respectively. This was difference statistically significant.

Analysis of the data found that differences in the identifier attribute "is word" was not a good predictor of *not seen* subject performance; ANOVA p-value=0.90. The interaction between the two predictor variables had a p-value of 0.62.

The mean subject percentage of incorrect *not seen* answers was slightly higher when all assignment identifiers contained words compared to all non-words (4.8% vs. 3.6%, with sd of 7.3% and 5.5% respectively), but this difference was not statistically significant.

5.5 Comparison of 2011 results with 2006

How do the results of the 2006 and 2011 experiment compare? Both ran for 20 minutes and subjects completed subjects completed almost the same number of problems.

The summary in Table .1 shows that recall and *would refer back* percentages are similar. There is a higher percentage of incorrect recalls in 2011, as might be expected with the increased amount of information that has to be remembered.

An analysis of the 2011 and 2006 subject answers finds a statistically significant difference in the mean percentage of "incorrect recall" answers (Student's t-test gives p-value=0.012; the 95% confidence interval for the 2011 mean percentage is between 1.3 and 10.1 greater than the 2006 value) but not for "correct recall" or *would refer back* (p-values of 0.55 and 0.30 respectively).

If subjects' percentage of incorrect recall answers increases, the percentage for correct recall and/or *would refer back* must decrease. The fact that the increase is statistically significant implies that it occurred over a large fraction of subjects, while the non-significant finding for the answers that should have decreased implies that there was a lot of variability in subject performance for these two kinds of answer.

Repeating the ANOVA analysis on the 2006 answers fails to find any significant predictors for any of the responses of interest, i.e., the factor "same first" was not a good predictor of *not seen* performance (which it was for the 2011 answers).

The 2004 ACCU experiment had the same remember/recall format as 2006/2011 but compared different assignment identifier attributes (i.e., number of syllables and word/nonword). The results for 2004 were: correct recall around 60%, incorrect recalls around 10% and *would refer back* 30%.

The filler problems used in the 2006/2011 experiments both involved providing answer to a short list if questions, i.e., making use of existing knowledge to solve a problem that only required a small amount of information to be held in STM.

6 Conclusion

The 2011 results suggest that when developers have to recall information about a recently seen list of identifiers they make more misidentification mistakes when those identifiers all start with the same letter compared to when they start with different letters; the 2006 results do not replicate this finding. If authors of coding guidelines practice feel it is worthwhile adding a fix sequence of letters to an identifier to denote some attribute, the number of mistakes made when reading these identifiers might be reduced if these characters were added somewhere other than at the start of the identifier (e.g., at the end).

No significant performance difference was found between assignment lists using identifiers that were all words or all non-words.

The results were consistent with the finding of two previous experiments where the breakdown of subjects answers was approximately: *would refer back* 25%, recall correct 60% and recall incorrect 15% (see Table .1). Increasing the number of to-be remembered assignment statements from three to four results the number of incorrect recall answers increasing.

Future experiments might confirm and extend these findings to identifiers containing more than three letters and investigate subject performance when handling identifiers having other attributes.

7 Further reading

Statistics Explained by Perry R. Hinton provides a very good introduction to statistics, including ANOVA.

For a readable introduction to human memory see "Essentials of Human Memory" by Alan D. Baddeley. A more advanced introduction is given in "Learning and Memory" by John R. Anderson. An excellent introduction to many of the cognitive issues that software developers encounter is given in "Thinking, Problem Solving, Cognition" by Richard E. Mayer.

8 Acknowledgments

The author wishes to thank everybody who volunteered their time to take part in the experiment and those involved in organising the ACCU conference for making a conference slot available in which to run it.

References

1. J. R. Anderson. *Learning and Memory: An Integrated Approach*. John Wiley & Sons, Inc, second edition, 2000.
2. A. D. Baddeley. How does acoustic similarity influence short-term memory? *Quarterly Journal of Experimental Psychology*, 20:249–264, 1968.
3. H.-F. Chitiri and D. M. Willows. Word recognition in two languages and orthographies: English and Greek. *Memory & Cognition*, 22(3):313–325, 1994.
4. V. Coltheart. Effects of phonological similarity and concurrent irrelevant articulation on short-term-memory recall of repeated and novel word lists. *Memory & Cognition*, 21(4):539–545, 1993.
5. W.-T. Fu and W. D. Gray. Memory versus perceptual-motor tradeoffs in a blocks world task. In *Proceedings of the Twenty-second Annual Conference of the Cognitive Science Society*, pages 154–159, Hillsdale, NJ, 2000. Erlbaum.
6. D. M. Jones. The new C Standard: An economic and cultural commentary. Knowledge Software, Ltd, 2005.
7. D. M. Jones. Developer beliefs about binary operator precedence. *C Vu*, 18(4):14–21, Aug. 2006.
8. D. M. Jones. Experimental data and scripts for effects of risk attitude on recall of assignment statements study. http://www.knosof.co.uk/dev-experiment/accu11.html, 2011.
9. D. M. Jones. Effects of risk attitude on recall of assignment statements. *C Vu*, 23(6):19–22, Jan. 2012.
10. M. Taft and K. I. Foster. Lexical storage and retrieval of polymorphemic and polysyllabic words. *Journal of Verbal Learning and Verbal Behavior*, 15:607–620, 1976.