Impact of semantic association on information recall performance
(part 2 of 2)

Experiment performed at the 2012 ACCU Conference

First published in CVu vol. 24 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 2012 ACCU conference; the first part was published in the last issue of C Vu.[8] This second part discusses the results from the linear relationship question that subjects answered. The experiment is derived from and takes account of results from previous ACCU conference experiments; in particular the it closely replicates one performed in the 2004 ACCU experiment.[5]

To recap the description of the experiment given in the first article: subjects first saw two, unnested, if-statements and were asked to remember the names of the variables and operators appearing in the control expressions, this information had to be recalled after they had analysed a nested if-statement having the following form (the results of the remember/recall question were covered in part 1):

```
if ((e > a) && (u < a))
   if (u > e)

      ............
   else
      ............
```

subjects were asked to indicate which arm of the nested if-statement they thought would be executed, should the conditional expression in the first if-statement be true; some questions did not have a unique answer, i.e., the first conditional expression did not sufficiently constrain the values of the two variables in the second conditional expression that it was possible to unconditionally deduce whether the expression was true/false. The analysis of subject responses to these questions is the subject of this article.

The relative order of the three variables was randomly chosen for each problem presented (the same identifiers, a, e and u, were always used).

1.1 The hypothesis

Studies that have investigated some of the kinds of relational reasoning that people encounter in everyday life have found patterns in subjects' performance, e.g., the accuracy of answers has depended on how the original relationships were specified (see below).

The two hypothesis tested by this part of the ACCU 2012 experiment is that the accuracy of subjects' (i.e., developers) answers is consistent with the two patterns outlined by De Soto, London, and Handel,[2] described below (also see Table .1).

2 Relational reasoning

The psychology of deduction uses the terms *linear syllogisms* or *linear reasoning* to describe deduction between statements involving relational operators. The term usually used to describe a (sub)expression containing a relational operator, in programming language specifications, is *relational expression*.

Linear syllogisms are part of mathematical logic and the skills associated with making deductions based on relational information are usually assumed to be one of the higher cognitive abilities that humans possess. However, studies have found that a number of animals have the ability to adapt their behavior to a given situation based on relational knowledge they have previously acquired. For instance, aggressive behavior between two animals is sometimes used to determine which one is dominant, relative to the other; aggression can lead to fighting and injury and is best avoided if possible. The ability to make use of relative dominance information (perhaps obtained when watching the interaction between other members of a social group) may reduce the need for aggressive behavior during an encounter between two members of the same group who have not yet established their relative dominance through a face to face encounter (i.e., the member most likely to loose is able to deduce this outcome and behave in a subservient fashion).

A study of Pinyon Jays (a social species of birds) and Scrub Jays (a non-social species) by Pazymino[11] found that individual birds from the social species appeared to make use of relational information to work out their relative dominance while birds from the non-social species did not.

2.1 Relational reasoning in humans

If some animal brains don't possess what are considered higher level cognitive reasoning abilities and yet possess a cognitive mechanism capable of combining and making use of relational information, it is possible that humans also possess a similar mechanism (this is not to say that they don't have any other high level cognitive systems capable of performing the same task). A possible consequence of having such a special purpose, lower level, reasoning mechanism is that it may not handle all relational expressions in the same way (i.e., it is likely to be optimized for handling those situations that commonly occur in it's owners everyday life). Some of the studies of human linear reasoning have found that subjects are slower and make more errors when the operands in a sequence of relational expressions occur in certain orders.

A study by De Soto[2] used a task based on what is known as *social reasoning* (using the relations *better* and *worse*). Subjects were shown two premises, involving the names of three people, and a possible conclusion (e.g., *Is Mantle worse than Moskowitz?*) and given 10 seconds to answer "yes", "no", or "don't know".

Table .1: Eight sets of premises describing the same relative ordering between A, B, and C in different ways (peoples names were used in the study), followed by the percentage of subjects giving the correct answer. Adapted from De Soto, London, and Handel.[2]

	Premises	Percentage Correct Responses		Premises	Percentage Correct Responses
1	A is better than B B is better than C	60.5	5	A is better than B C is worse than B	61.8
2	B is better than C A is better than B	52.8	6	C is worse than B A is better than B	57.0
3	B is worse than A C is worse than B	50.0	7	B is worse than A B is better than C	41.5
4	C is worse than B B is worse than A	42.5	8	B is better than C B is worse than A	38.3

Based on the results (see Table .1) the researchers made two observations (which they called *paralogical principles* (cases 5 and 6 possess both, while cases 7 and 8 possess neither):

1. People process orderings more accurately in one direction compared others. Subjects' gave more correct answers when the ordering direction was better-to-worse (case 1), than mixed direction (case 2, 3), and were least correct in the direction worse-to-better (case 4). This suggests that use of the word *better* should be preferred over *worse* (the British National Corpus[9] lists *better* as appearing 143 times per million words, while *worse* appears under 10 times per million words and it is not listed in the top 124,000 most used words).
2. People end-anchor orderings; that is, they focus on the two extremes of the ordering. In this study people gave more correct answers when the premises stated an end term (better or worse) followed by the middle term, than a middle term followed by an end term.

A related experiment in the same study used the relations *to-the-left* and *to-the-right*, and *above* and *below*. The above/below results were very similar to those for better/worse. The left-right results showed that subjects performed better with a left-to-right ordering than a right-to-left ordering.

The strategy used to solve a given problem has been found to vary between people. A study by Sternberg and Weil[15] found a significant interaction between a subjects' aptitude (as measured by verbal and spatial ability tests) and the strategy they used to solve linear reasoning problems. However, a person having high spatial ability, for instance, does not necessarily use a spatial strategy. A study by Roberts, Gilmore, and Wood[14] asked subjects to solve what appeared to be a spatial problem (requiring the use of a very inefficient spatial strategy to solve). Subjects with high spatial ability used non-spatial strategies, while those with low spatial ability used a spatial strategy. The conclusion made was that those with high spatial ability were able to see that the spatial strategy was inefficient and to select as alternative strategy, while those with less spatial ability were unable to perform this evaluation.

If the evaluation of relational expressions in source code is performed using a cognitive mechanism that has been optimized for certain kinds of frequently occurring, everyday, activities then it is possible that developer performance will be good for relational expressions that match the form of these everyday activities and not so good on relational expressions that don't match. The if-statement conditional expressions used in this study permuted over all possible combinations of operator/operand ordering.

The list of questions for each subject was generated by randomising the eight possible operator/operand orderings, creating questions using this ordering, randomizing the orderings again and repeating until all of the required questions had been generated. This process was repeated to when generating the problem sheets for each subject.

3 Threats to validity

As well as the possible threats to validity listed in part 1, the following are specific to the subject of part 2.

Although subjects were told: "Treat the paper as if it were a screen, i.e., it cannot be written on.", there was nothing to prevent them using the paper on which the questions were written as a temporary work area. Several subjects did write notes on the paper next to a few if-statement problems.

For those questions whose answer was that either arm might be executed some subjects wrote a question mark (i.e., ?) as their answer and some left the answer blank. Both forms of answer were treated as specifying that either arm of the nested if-statement could be executed. It is not possible to check whether this assumption was the intended answer.

Measurements of C source[6] show that the binary less-than operator (i.e., <) occurs twice as frequently as the greater-than operator (i.e., >), compared to the better/worse English words used by De Soto et al which has a frequency ratio of 14. It is possible that the much lower frequency ratio for the relational operators will cause the performance for both of them to be very similar.

Subjects can approach the demands of answering the problems this study presents them in a number of ways, including the following:

- seeing it as a challenge to accurately remember/recall the conditional expression information and be willing to trade-off performance on the relational operand question,
- recognizing that *would refer back* is always an option, but that it is more important to correctly answer the relational operand question,
- making no conscious decision about how to approach the answering of problems,

4 Results

A total of 432 nested if-statement problems were answered by 22 subjects, of which 47 (10.9%, in 2004 the percentage was 4.7%) were incorrect. The mean number of answers per subject was 19.6 (sd 7.7), slightly lower than the 2004 mean of 21.

Subjects had a mean of 15.1 years (sd 10.2) experience writing software professionally.

The number of incorrect answers is very weakly correlated with the number of problems answered (Pearson correlation coefficient 0.24, 95% confidence interval -0.20 to 0.60). While performance on reasoning tasks has been found to decrease with age,[3] subject software development experience (which is likely to be highly correlated with age) is not correlated with percentage of incorrect answers (Pearson correlation coefficient 0.02, 95% confidence interval -0.42 to 0.45) and only very weakly to the number of answers given (0.29, 95% confidence interval -0.16 to 0.64)

The error rates reported by other studies (where subjects read a problem typed on a card) were: De Soto et al[2] 39.2–61.7% (subjects were required to answer within 10 seconds rather than in their own time), Clark[1] 6%, Potts[12] 5%, Mayer[10] 4–36%, Quinton et al[13] not given, Sternberg et al[15] 1.7–3.5%. A study where subjects heard a tape recoding of the problem[4] reported an error rate of 8–19%.

In the following discussion *H* denotes high, *M* denotes middle, and *L* denotes low. So `H > M` denotes "high greater than middle" and `M > L` "middle greater than low". *unk* is used to denote the case where the conditional expression does not uniquely specify the relationship between all three variables.

All of the data and R code used in the analysis is available on the experiments web page.[7]

4.1 Reasoning performance

Table .2 lists the number of correct and incorrect answers for various combinations of relational operators in the outer if-statement, ordered by percentage of incorrect answers (the percentages from the 2004 ACCU experiment are in the last column).

Table .2: For all subjects, the total number of correct/incorrect answers and the percentage of incorrect answers for the combination of relational expressions appearing in the first two columns. Last column is the percentage incorrect in the 2004 ACCU experiment. *H* denotes high, *M* denotes middle, and *L* denotes low. So `H > M` denotes "high greater than middle" and `M > L` "middle greater than low"; *unk* is used to denote the case where the conditional expression does not uniquely specify the relationship between all three variables.

Left condition	Right condition	Correct	Incorrect	Percent	2004 %
M < H	L < M	51	3	0.056	0.059
L < M	M < H	33	2	0.057	0.038
H > M	L < M	40	3	0.070	0.055
M > L	M < H	40	4	0.091	0.056
M < H	M > L	39	4	0.093	0.044
L < M	H > M	34	4	0.105	0.028
M > L	H > M	42	5	0.106	0.066
H > M	M > L	37	6	0.140	0.017
unk	unk	69	16	0.188	NA

If subject behavior in 2012, for this question, was consistent with that in 2004 the relative order of percentage incorrect answers for the two years would be strongly correlated, however a Kendal rank correlation test shows a weak negative correlation (i.e., -0.29).

The following compares the results against those predicted by the two hypothesis proposed in the introduction:

1. Use of the more common operator reduces incorrect answers: Looking at the operands appearing in questions having the lowest and highest percentage of incorrect answers we see that these closely match the predictions made by the hypothesis (see the first two columns of Table .3), what is the probability of this occurring through a random process?

   There are $8!$ ways of arranging the 8 available combinations (the *unk* case is ignored here); there are two ways in which the two less-than operators can occur first, one way a greater-than can occur last and $5!$ different ways of ordering the other possibilities, giving a probability of $\frac{2!5!}{8!} \Rightarrow 0.006$ for this combination occurring at random (well below a p-value of 0.05).

   While the 2012 behavior matches this hypothesis the results from the 2004 experiment do not; in fact for 2004 the lowest incorrect percentage combination and second highest percentage are swapped, almost the opposite of the proposed hypothesis.

2. End-anchoring: The operand ordering that is the complete opposite of this end-anchoring pattern has the lowest percentage of incorrect answers, orderings that follows this pattern have the second lowest percentage and highest percentage of incorrect answers. The *Middle* value appears as the first operand twice (for both left and right relational expressions) in the lowest incorrect percentage and the high incorrect percentage; there is no evidence of any end-anchoring.

Figure .1 shows the percentage of incorrect answers given by each subject, ordered by increasing percentage. Just under half of the subjects do not given any incorrect answers; perhaps the analysis will reach a different

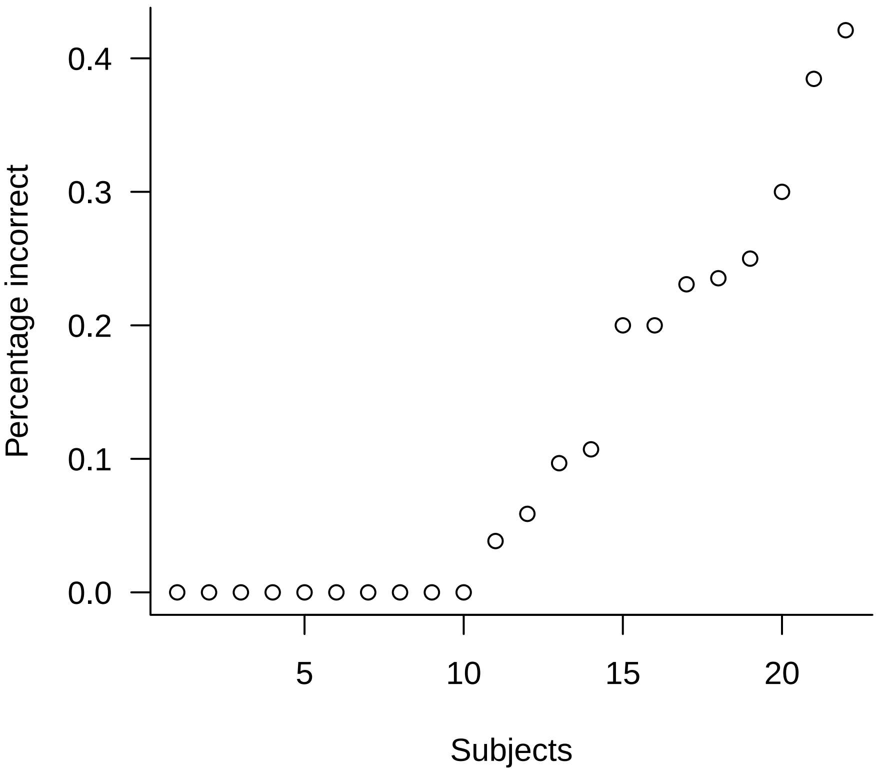

Figure .1: Subjects listed in order of increasing percentage of incorrect answers.

conclusion if subjects who give very few incorrect answers are excluded. Table .3 only includes results from subjects who answers were at least %6 incorrect. There is only one change in the relative ordering, `M>L H>M` moves up to 4th from 7th.

Table .3: A subset of Table .2 created by only including results from those subjects whose percentage of incorrect answers was greater than 6%.

Left condition	Right condition	Correct	Incorrect	Percent
M < H	L < M	22	3	0.120
L < M	M < H	10	2	0.167
H > M	L < M	15	3	0.167
M > L	H > M	19	4	0.174
M > L	M < H	17	4	0.190
L < M	H > M	12	3	0.200
M < H	M > L	13	4	0.235
H > M	M > L	14	6	0.300
unk	unk	32	16	0.333

In those cases where the conditional expression in the outer if-statement did not contain enough information

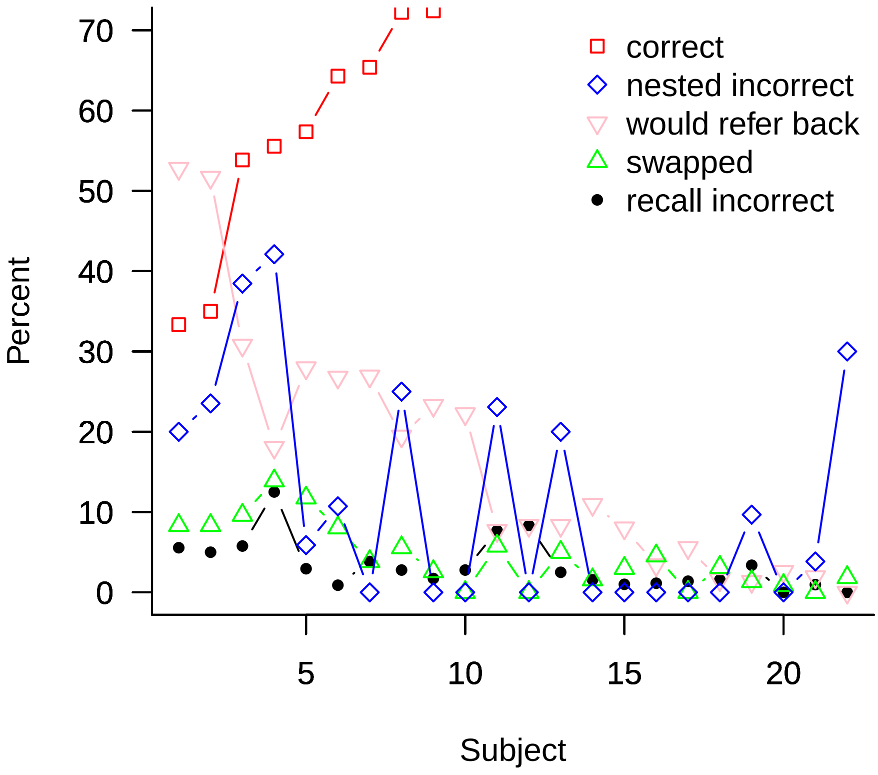

Figure .2: The percentage of operand recall correct (square box), relational incorrect (diamond), *would refer back* (triangle point-down), operand swapped (triangle point-up) and operand recall incorrect answers (bullet) for each subject. Subjects are ordered by percentage of operand recall correct answers given (scale clipped to expand relational view).

to uniquely specify which arm of the nested if-statement would execute, the first arm was incorrectly given in 10 answers and the second arm in 6 answers. If we assume there is an equal probability of either arm being incorrectly specified, then there is a 12% chance of 10 out of 16 incorrect answers specifying one particular arm.

There were 9 answers specifying that either arm was possible but in fact the correct answer to the question was one particular arm (5 for one arm, 4 for the other).

4.2 Interaction between remember/recall and reasoning questions

Having to answer the first part of the problem (i.e., remembering information about the variables in the control expression of two if-statements) ties up cognitive resources (e.g., short term memory decays over time and unless regularly refreshed it is soon lost), leaving less resources to process the nested if-statement problem.

Figure .2 shows that percentage of incorrect answers in the relational question is not correlated with percentage of correct answers in the operand remember/recall question.

However, there is quite a good correlation between incorrect answers and percentage of swapped operand answers to the remember/recall question (Pearson correlation coefficient 0.66, with a 95% confidence interval of 0.34 to 0.85, p-value = 0.00074); the correlation with percentage of incorrect operand answers is not quiet so good (0.55, with a 95% confidence interval of 0.17 to 0.79, p-value = 0.0076).

5 Discussion

A surprisingly high percentage (45%) of subjects gave no incorrect answers. This observation was not noted in 2004 because the low number of incorrect answers given by subjects meant that zero incorrect was not surprising (the 2004 figure was 44% subjects giving no incorrect answer).

Looking at Figure .2 many of the subjects who had a very low percentage of incorrect answers also had a very low percentage of incorrect remember/recall answers. There appears to be a group of subjects whose performance on the two questions in this experiment was much better than other the subjects.

There is a good correlation between incorrect answers to the nested if-statement question and percentage of swapped operand answers in the remember/recall question. Both of these findings are consistent with subjects' having problems processing the relative ordering of identifiers seen in code.

There are two notable differences in subject performance between 2004 and 2012:

- the percentage of incorrect answers is more than twice as high in 2012; the figure is reduced from 10.9% to 7.2% if *unk* answers are excluded,
- there is no correlation between the form of conditional expression and the relative incorrect answer rate in the results from the two years.

and two differences in the questions answered:

- the remember /recall question involved assignment statements in 2004 and the operands of if-statement conditional expressions in 2012,
- in 2004 the nested if-statement relational expression question always had an answer that was one of the two arms, while in 2012 the question could have the answer that either arm might be executed.

Were the differences in the questions used the main contributing factor in the differences in subject responses seen in the two years?

More experiments are needed to find out why nearly half of the subjects gave no incorrect answers (or alternatively why just over half gave incorrect answers) and to find out what caused subject performance to vary on almost the same question (in 2004 and 2012).

6 Conclusion

There were two groups of subjects who exhibited their own consistent behavior in answering the two questions in this experiment:

- Approximately 40% of subjects gave a very low percentage of incorrect answers to both questions (i.e., zero or one incorrect answer),
- approximately 25% of subjects showed some tendency to mix up the order of operands appearing in both questions.

7 Further reading

For a readable introduction to human reasoning see *Reasoning and thinking* by Ken Manktelow. *The Cognitive Animal* edited by M. Bekoff, C. Allen, and G. M. Burghardt contains 57 short, wide ranging, essays (of varying quality) on animal cognition.

7.1 Acknowledgments

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

References

1. H. H. Clark. Linguistic processes in deductive reasoning. *Psychological Review*, 76(4):387–404, 1969.
2. C. B. De Soto, M. London, and S. Handel. Social reasoning and spatial paralogic. *Journal of Personality and Social Psychology*, 2(4):513–521, 1965.
3. A. S. Gilinsky and B. B. Judd. Working memory and bias in reasoning across the life span. *Psychology and Ageing*, 9(3):356–371, 1994.
4. J. Huttenlocher. Constructing spatial images: A strategy in reasoning. *Psychological Review*, 75(6):550–560, 1968.
5. D. M. Jones. Experimental data and scripts for short sequence of assignment statements study. http://www.knosof.co.uk/cbook/accu04.html, 2004.
6. D. M. Jones. The new C Standard: An economic and cultural commentary. Knowledge Software, Ltd, 2005.
7. D. M. Jones. Experimental data and scripts for impact of semantic association on information recall performance. http://www.knosof.co.uk/dev_experiment/accu12.html, 2012.
8. D. M. Jones. Impact of semantic association on information recall performance. *C Vu*, 24(6):3–8, Jan. 2013.
9. G. Leech, P. Rayson, and A. Wilson. *Word Frequencies in Written and Spoken English*. Pearson Education, 2001.
10. R. E. Mayer. Qualitatively different encoding strategies for linear reasoning premises: Evidence for single association and distance theories. *Journal of Experimental Psychology: Human Learning and Memory*, 5(1):1–10, 1979.
11. G. Paz-y-Miño C, A. B. Bond, A. C. Kamil, and R. P. Balda. Pinyon jays use transitive inference to predict social dominance. *Nature*, 430:778–781, Aug. 2004.
12. G. R. Potts. Storing and retrieving information about ordering relationships. *Journal of Experimental Psychology*, 103(3):431–439, 1974.
13. G. Quinton and B. J. Fellows. 'Perceptual' strategies in the solving of three-term series problems. *British Journal of Psychology*, 66:69–78, 1975.
14. M. J. Roberts, D. J. Gilmore, and D. J. Wood. Individual differences and strategy selection in reasoning. *British Journal of Psychology*, 88:473–492, 1997.
15. R. J. Sternberg and E. M. Weil. An aptitude × strategy interaction in linear syllogistic reasoning. *Journal of Educational Psychology*, 72(2):226–239, 1980.