Australian Government
Department of Defence
Defence Science and
Technology Organisation

Literature Review on Mental Models and Linear Separability

Susannah J. Whitney

Land Operations Division
Defence Science and Technology Organisation

DSTO-GD-0741

ABSTRACT

The mental models theory suggests that people make reasoning errors because they construct partial – and inaccurate – mental models. It predicts that where people are required to consider false information, they are more prone to making errors than when they are only required to consider true information. Findings consistent with this theory have been demonstrated across a number of studies, particularly the work of Johnson-Laird. However, researchers at DSTO suggested that these findings are better explained by a linear separability effect. That is, problems are easier to solve when they are linearly separable than when they are nonlinearly separable. That is, the simplicity and precision with which correct and incorrect answers can be separated determines the extent to which they will be solved correctly. This literature review examines research on mental models and linear separability published between 2000 and 2012, to establish if this explanation has been proposed by other researchers. Results indicate that no other researchers have proposed this, or similar, explanations, hence the linear separability hypothesis has the potential to make a novel contribution to the literature.

Published by

Land Operations Division
DSTO Defence Science and Technology Organisation
PO Box 1500
Edinburgh South Australia 5111 Australia

Telephone: 1300 DEFENCE
Fax: (08) 7389 6567

AR-015-591
April 2013

APPROVED FOR PUBLIC RELEASE

Literature Review on Mental Models and Linear Separability

Executive Summary

Researchers such as Johnson-Laird have consistently demonstrated that people have difficulty solving complex reasoning problems, such as:

Only one statement about a hand of cards is true:
1. There is a King or Ace or both.
2. There is a Queen or Ace or both.
Which is more likely, King or Ace?

While the majority of people will respond that the Ace is more likely to occur, this is logically incorrect. As only one statement about the hand of cards is true, the Ace can never occur, hence the King is more likely. Johnson-Laird suggests that the reason people make such mistakes is that they construct partial mental models to assist in reasoning. However, these models are flawed as they do not represent false information, e.g. that if Statement 1 is true in the problem above, then Statement 2 must be false, and vice versa.

Defence Science and Technology Organisation (DSTO) researchers have suggested that the concept of linear separability is more accurate at explaining Johnson-Laird's findings. Categories are linearly separable when a single line (for categories with two dimensions, such as height/weight, colour/shape) can be drawn that differentiates between categories. Categories are nonlinearly separable when they cannot be differentiated using a single line.

In a preliminary study conducted in 2003-2004, DSTO researchers demonstrated that the linear separability explanation for Johnson-Laird's findings was plausible and supported by the data. In order to examine this more fully, research was commenced in 2011 under the auspices of Land Operations Division's Enabling Research project.

One key component of this work was a review of relevant literature on mental models and linear separability published between 2000 and 2012 to identify related studies and ensure that no similar or competing research had been conducted. Results indicated that while a number of similar studies have been conducted, the Enabling Research Project still represents a novel contribution to the body of research.

This page is intentionally blank

Contents

This page is intentionally blank

Glossary

DSTO	Defence Science and Technology Organisation
ERP	Enabling Research Project
LOD	Land Operations Division
LS	Linearly separable
NLS	Nonlinearly separable

Boolean Algebra terms

AND	All values are true, e.g. *A AND B* means that both A and B are true
NOT, ¬	The value is not true
OR	One, some, or all values are true, e.g. *A OR B* is true when A is true, B is true, and when A and B are true
XOR	Exclusive Or; only one value is true, e.g. *A XOR B* means that either A or B, but not both, is true

This page is intentionally blank

1. Introduction

1.1 Overview

Researchers have repeatedly demonstrated that humans have difficulties solving complex reasoning problems. This has implications both for cognitive theories and for applied areas such as the way military intelligence reports are presented. Defence Science and Technology Organisation (DSTO) researchers have suggested that the concept of linear separability can explain some complexities in problem solving, which have previously been explained by the mental models theory [1]. Some preliminary research in this area was conducted in 2003-2004 by DSTO vacation students [2], and the work was recommenced in 2011 under Land Operations Division's (LOD) Enabling Research Program (ERP).

As part of the ERP work, a literature review was conducted to identify articles published since the preliminary research was conducted. This was done to ensure that the ERP work still makes a novel contribution to the body of research. The literature review begins with an overview of the logical principles relevant to the study, describes the predominant theory and some key findings, then discusses major studies conducted in the area from 2000-2012. Overall, the literature review reveals that while some studies conducted in this timeframe have some similarities to the ERP study, there are sufficient differences such that the ERP study still represents a novel piece of research. This report is intended to be read in conjunction with [3], which describes the methodology, results, and implications of the study in more detail.

1.2 Relevant logical principles

Consider the following logical problem.

Problem 1

If the server is full, then memory is busy.
The server is full. What, if anything, can be deduced about memory?

It is reasonably straightforward to deduce that the answer to this problem is that if the server is full, then it follows that memory must be busy. This problem is likely to be solvable without any formal training in logic. However, what if the server is not full? What can be deduced about the state of memory in this situation?

In order to explain this in enough detail to understand the fundamentals of the research project, it is necessary to cover some basics of logical reasoning. The problem above is an example of a logical statement of the form *if A then B*. Under formal logical rules[1], this means that B, known as the *consequent*, always follows in the presence of A, known as the *antecedent*. The relationship between A and B is of a type known as a *conditional*, where the presence of A implies the presence of B.

[1] *Modus ponens;* see, for instance http://en.wikipedia.org/wiki/Modus_ponens for more detail.

One way of representing if A, then B is through a truth table. Table 1 shows the possible values (true or false) for A and B, and the resultant values for the statement if A, then B. The first two lines are reasonably intuitive; if A and B are both true, then the statement is true, and if B is false, then the statement is false. The third and fourth lines demonstrate an important logical principle: if the antecedent is false, then the consequent can be true or false, and the statement will still be logically true. Using the example above, if the server is *not* full, then memory can be either busy or not busy without the statement being logically false.

Table 1: Truth table for If A, then B

A	B	If A, then B
True	True	True
True	False	False
False	False	True
False	True	True

There are a number of other logical relationships, or operators, that are relevant to this research project, including:

- *AND* – all values are true, e.g. *A AND B* means that both A and B are true.
- *OR* – one, some, or all values are true, e.g. *A OR B* is true when A is true, B is true, and when A and B is true.
- *Exclusive OR* (*XOR*) – only one value is true, e.g. *A XOR B* means that either A or B, but not both, is true.
- *NOT* – this means that a value is not true.

For more detail on these operators, or the principles of Boolean algebra underpinning them, the reader is referred to [4]. To ensure clarity, in this literature review, whenever logical problems are included they are italicised. When terms such as "and" and "or" are used in the logical sense (as opposed to the naturalistic sense), they will be italicised or written in uppercase or both.

1.3 Johnson-Laird's mental model theory

Johnson-Laird and colleagues suggest [5, 6] that in order to solve reasoning problems, people construct mental models. However, as the complexity of the problem increases, people omit information from the models, neglecting to represent explicitly false information. While this keeps the problem within the limits of working memory, it inadvertently introduces errors into the reasoning process. In Problem 1 above, for instance, Johnson-Laird suggests people would omit the last two lines of the table, where the antecedent is false.

To examine how this introduces errors into reasoning processes, consider the following problem [5]:

Problem 2

Only one statement about a hand of cards is true: 1. There is a King or Ace or both. 2. There is a Queen or Ace or both. Which is more likely, King or Ace?"

Johnson-Laird suggests that in solving this problem, people will construct a mental model similar to Table 2[2]. The first line represents the first statement (*King OR Ace OR both*), and the second line represents the second statement (*Queen OR Ace OR Both*). The ellipses in the third line represent all other possibilities, which are not represented in the model to conserve working memory space. Based on this model, it is intuitive, and appealing, to conclude that the Ace is more likely since it occurs twice in the model, compared to only once for the King.

Table 2: Partial Problem 2 mental model

King		Ace
	Queen	Ace
	…	

However, this answer is incorrect. Johnson-Laird attributes this to the use of the partial mental model, which does not take into account the fact that only one statement about the hand of cards can be true[3]. That is, if Statement 1 is true (*King OR Ace OR Both*), Statement 2 (*Queen OR Ace OR Both*) must be false. A complete model of the problem is shown in Table 3 below, with parentheses around an item indicating that it cannot occur. The first three lines represent the possible cards in the hand if the first statement is true, while the last three lines represent the possible cards if the second statement is true. It is clear from the table that an Ace can never occur, only the King or Queen.

Table 3: Complete Problem 2 mental model

King	(Ace)	(Queen)	(Ace)
(King)	Ace	(Queen)	(Ace)
King	Ace	(Queen)	(Ace)
Queen	(Ace)	(King)	(Ace)
(Queen)	Ace	(King)	(Ace)
Queen	Ace	(King)	(Ace)

Over a series of experiments, Johnson-Laird and colleagues [5, 6] have demonstrated a robust effect, where the majority of participants were unable to correctly solve problems such as Problem 2, while they were able to solve problems that did not require representation of

[2] Note that Johnson-Laird also uses the term "model" to describe each line in the table, e.g. using his terminology Table 2 includes 3 models, King and Ace, Queen and Ace, and All Other Possibilities.

[3] This model also does not include separate lines for the different combinations "King", "Ace", "King AND Ace", but this does not alter the outcomes.

explicitly false information. For instance, Problem 2 was solved correctly by only 21% of participants, whereas the following problem was solved correctly by 79% of participants:

Problem 3

> If there is a King in the hand then there is an Ace in the hand,
>
> or if there is a Queen in the hand, there is an Ace in the hand.
>
> Which is more likely, King or Ace?[5 p.77]

The correct answer to this problem is "Ace". Table 4 shows the partial mental model. The only combination not shown in this table is the absence of the King and the Queen. As discussed previously, the Ace can logically occur if the King or Queen is absent; but it is not necessary to resolve this in order to solve the problem correctly.

Table 4: Partial Problem 3 mental model

King		Ace
	Queen	Ace
King	Queen	Ace
	…	

Johnson-Laird's findings were replicated in a DSTO study. In [2], military and civilian participants were presented with problems that did or did not require representation of explicitly false information. An example of the former is:

Problem 4

> Only one of the following statements about an impending enemy attack is true:
>
> 1. The enemy will approach from Wade Valley or Swain Valley or both.
>
> 2. The enemy will approach from Swain Valley and artillery fire will warn of their approach.
>
> Is it possible for the enemy to come from Swain Valley and for artillery fire to warn of their approach?

An example of the latter is:

Problem 5

> Only one of the following statements about a road convoy is true:
>
> 1. There is an Armoured Personnel Carrier in the convoy or there is a Tank in the convoy or both
>
> 2. There is a Mine Clearance Vehicle in the convoy and a Tank in the convoy
>
> Is it possible for there to be an Armoured Personnel Carrier and a Tank in the convoy?

The correct answer to Problem 4 is "no"; the enemy can never approach from Swain Valley with artillery fire warning of their approach. This meets the conditions for Statement 2 to be true, but the enemy approaching from Swain Valley also fulfils the requirements for Statement 1 to be true (Wade Valley or Swain Valley or both). However, the problem explicitly states that only one statement is true.

The correct answer to Problem 5 is "yes". If there is an Armoured Personnel Carrier in the convoy, the first statement is true. If there is a Tank in the convoy (but no Mine Clearance Vehicle), the second statement is false. Hence, it is possible for there to be an Armoured Personnel Carrier and a Tank.

Results from this study were consistent with those obtained by Johnson-Laird and colleagues. That is, participants were more likely to correctly answer questions that did not require representing explicitly false information. There were no significant differences in error rates between military and civilian participants, although the military participants were significantly faster to respond.

1.4 Linear separability explanation for Johnson-Laird's findings

Galanis and colleagues [1] suggest that the concept of linear separability provides a better, more comprehensive explanation for Johnson-Laird's findings than the mental models theory. This concept comes from the domain of category learning; the extent to which humans and other species are able to learn that objects belong to certain categories.

As noted by Blair and Homa [7], when objects to be categorised are plotted on x- and y-axes, categorisation is linearly separable when a single line can be drawn that differentiates between category membership. Otherwise, categorisation is nonlinearly separable. To illustrate this, consider that there is a group of objects that are either triangular or circular in shape, and red or blue in colour. These objects clearly vary on two dimensions, colour and shape. Next, consider that these objects belong to either Category A, or Category B, and that category membership is determined by some combination of the two dimensions.

Figure 1 shows hypothetical categorisation if Category A comprised objects that were red or a triangle or both, and Category B comprised objects that were not red or a triangle or both. This can be expressed as a Boolean function, where Category A membership is *(Red OR Triangle)*, and Category B membership is *NOT (Red OR Triangle)*. As the figure shows, it is possible to draw a single line differentiating between Category A and Category B. This is known as Linearly Separable (LS) categorisation.

However, consider a different hypothetical categorisation, as shown in Figure 2. In this instance, an item is Category A if it was red and a triangle, or blue and a square, otherwise it was Category B. When written as a Boolean function, Category A membership is *(Red AND Triangle) OR (Blue AND Square)*. This equation can be simplified to *(Red XOR Square)*[4]. Similarly, Category B membership simplifies to *NOT (Red XOR Square)* In this instance, the two members of Category A, the red triangle and blue square, cannot be delineated from the

[4] See, for instance, http://www.allaboutcircuits.com/vol_4/chpt_7/7.html

two members of Category B using a single line, as shown in Figure 2. This is known as Nonlinearly Separable (NLS) categorisation. In general, categories using *XOR* to determine membership are NLS, while categories using *OR* function are LS.

In these examples, in order to correctly categorise an object it is necessary to consider both the colour and the shape of an object. It would not be possible to make a decision on the basis of a single dimension. This is known as an unreducible decision. In contrast, if it were possible to decide on the basis of a single dimension, for instance, if all circles were Category A, and all squares were Category B, and was unnecessary to consider colour, this would be known as a reducible decision.

Categorisation problems have important implications for theories of learning and memory. In particular, researchers suggest that the use of NLS and LS categorisation tasks can illuminate areas such as the way in which people learn. Some studies have shown that NLS categories are more difficult to learn than LS categories [8, 9], while others have shown that NLS and LS categories are learned equally easily [7, 10]. However, it is acknowledged that categorisation problems have important implications for theories of learning and memory and whether or not there are any limits on the amount of categorisation information that can be learned [11]. In addition, categorisation problems using *XOR* have been described as some of the most important NLS functions [12].

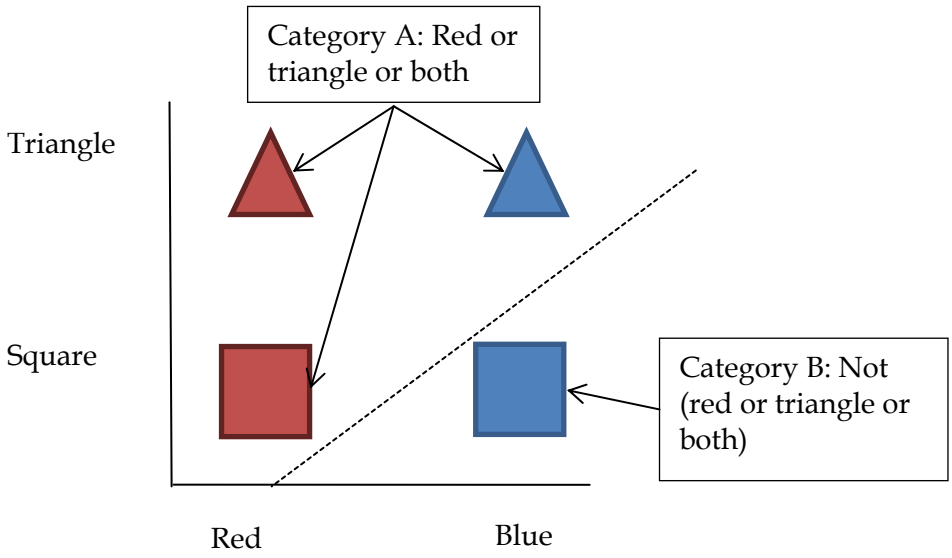

Figure 1: Linearly separable categorisation

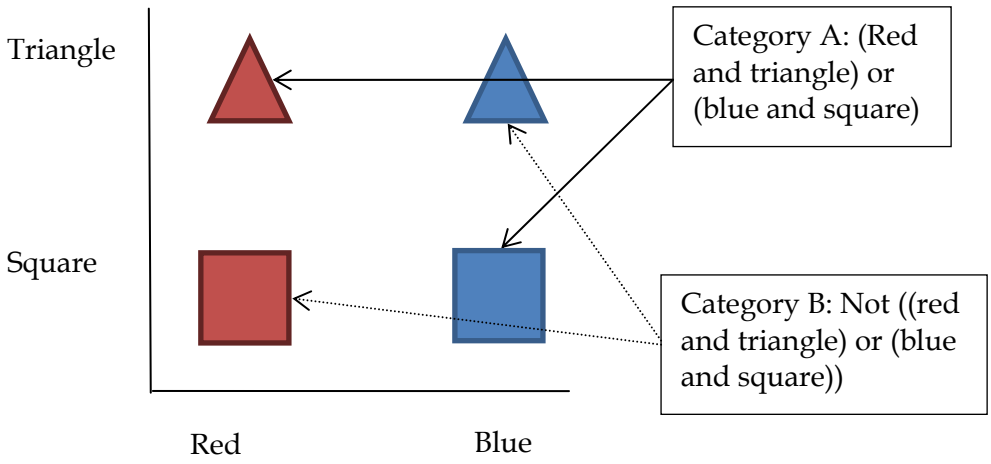

Figure 2: Nonlinearly separable categorisation

In proposing a linear separability explanation for Johnson-Laird's findings, Galanis and colleagues [1] observed that problems which the majority of people solve incorrectly tend to be NLS, whereas problems the majority of people solve correctly are LS. For instance, Problem 2, when written as a Boolean function, is *(King AND Ace) XOR (Queen AND Ace).*

This is shown in Figure 3. In the figure the symbol "¬" represents "*NOT*", and the starbursts indicate answers that are logically true. That is, as discussed on p3, the only possible combinations of cards are a King or a Queen (but not both), and the Ace can never occur. As noted by Blair and Homa [7], as this problem has three variables rather than two, in order for it to be LS, it would be necessary to draw a two dimensional plane separating true from false answers. As the figure shows, this is not possible. This problem was solved correctly by only 21% of participants in [5].

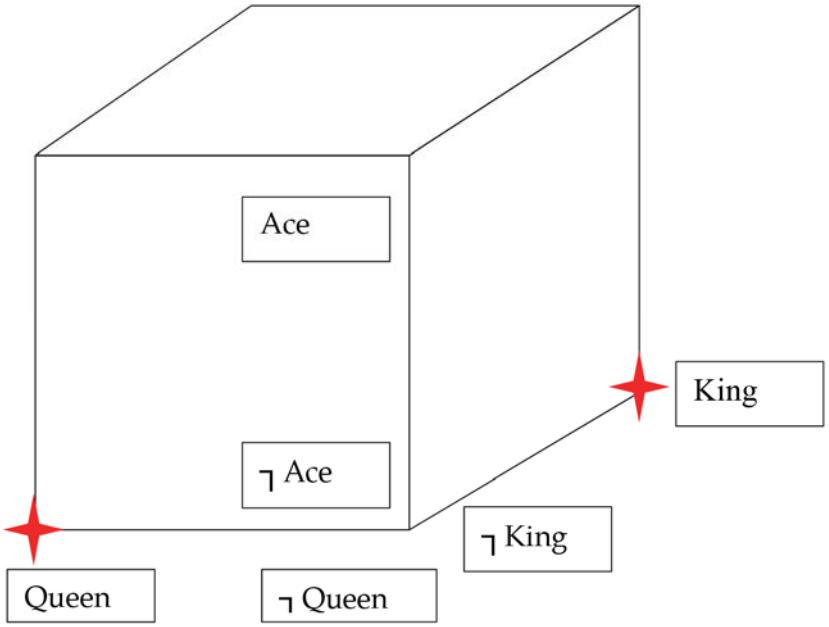

Figure 3: Problem 2, starbursts indicate true answers

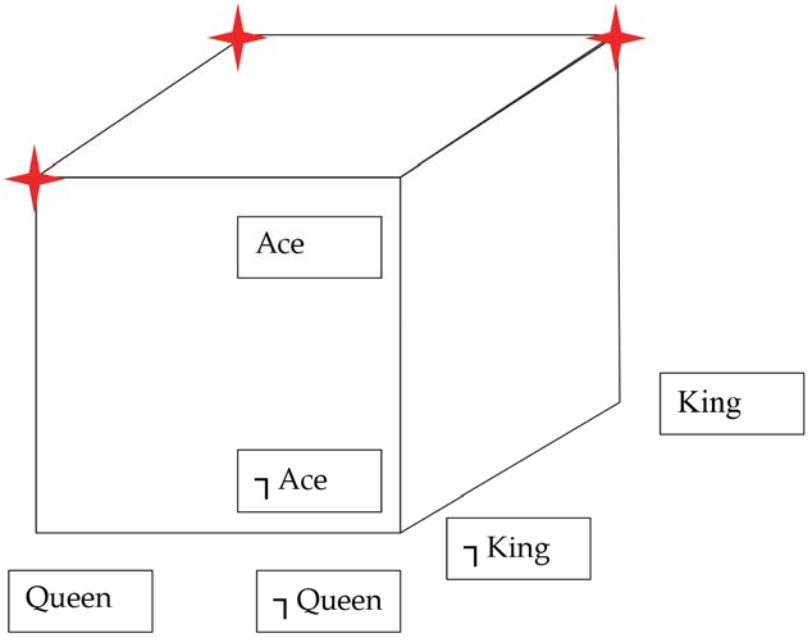

Figure 4: Problem 3, starbursts indicate true answers

In contrast, consider Problem 3. When written as a Boolean function, this becomes *(King AND Ace) OR (Queen AND Ace)*. This is represented graphically in Figure 4. It is clear that this is an LS problem, as a single plane can be drawn through the figure separating true and false answers. This problem was solved correctly by 79% of participants in [5]. Hence, Galanis et al.'s [1] linear separability explanation of Johnson-Laird's findings is plausible.

1.5 The DSTO ERP study

The aim of the DSTO ERP study was to explore linear separability as an alternative explanation for Johnson-Laird's findings. Rather than using written problems, the study used combinations of three light switches. One of the reasons for doing this is because some researchers [13] have suggested that Johnson-Laird's findings can be attributed to participants misreading or misunderstanding the questions; this will be discussed in more detail in the following sections.

In the study, each participant was trained and tested on 1 separable and 1 inseparable function. Each function contained 3 variables, displayed as a combination of shaded and unshaded shapes (see Figure 5). Participants were told that these shapes were switches controlling a hypothetical light, and that a certain combination of shaded and unshaded shapes switched the light on. Using onscreen buttons, participants judged if the light is on or off. Immediate onscreen feedback (CORRECT or INCORRECT) was provided. Participants saw 8 combinations of switches (representing all possible combinations) presented 8 times. Participants were not given explicit strategies about how they should attempt to learn the function, although a post-experimental survey was given to determine if participants were attempting to deduce the rule, memorise correct combinations, or use some other strategy.

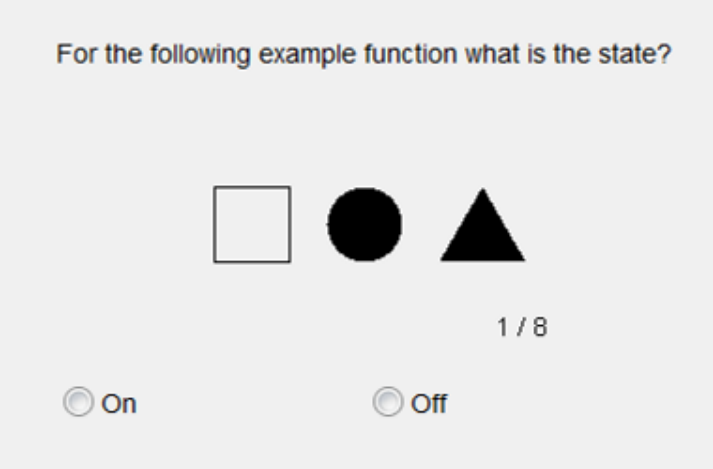

Figure 5: Screen shot from experiment, training phase

Following this, participants were tested on their knowledge of the function. They were presented with a combination of one or two shapes and the light state (on or off), as in Figure 6, and asked what can be deduced about another shape. Each train-and-test sequence was repeated 5 times. After completing training and test for one function, the training and test procedure was repeated for the second function. It was hypothesised that NLS functions would take more trials to learn, and result in poorer comprehension performance, than LS functions. The methodology and results for this study are described in more detail in [3].

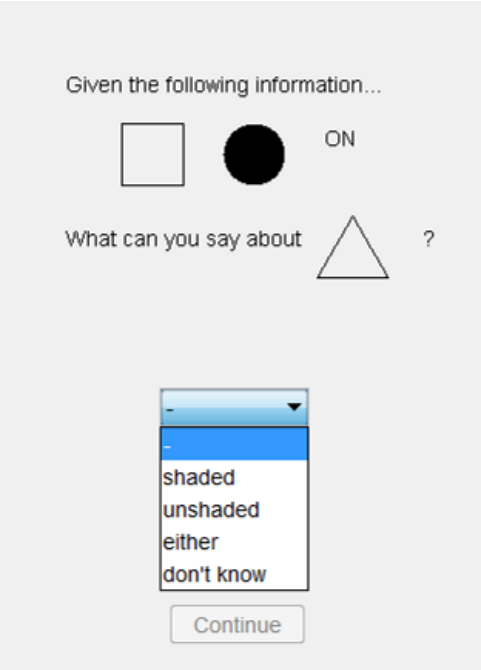

Figure 6: Screenshot from experiment, test phase

The work conducted by DSTO vacation students in 2003-2004, e.g. [2] included an initial literature review. However, as approximately 8 years has passed since that work was conducted, an additional literature review was conducted to identify any key studies that had been published in that time. In order to ensure that all relevant studies were identified, the timeframe for the literature review was set at 2000-2012.

2. Literature Review

Searches were conducted using the PsychINFO database and Google Scholar. A broad search strategy was employed. Initial searches used two key methods. Firstly, searches were conducted using key phrases and categories, such as "category learning", "Johnson-Laird", and "linear separability". Secondly, searches were conducted to identify recent publications citing some of the key papers in this field, including Johnson-Laird's early papers [5, 6], and the previous research conducted at DSTO [2]. As relevant papers were identified, further searches were conducted to identify papers that cited them.

The results of the literature review are summarised in two categories. Firstly, Section 2.1 reviews recent research related to Johnson-Laird's work on mental models. Secondly, Section 2.2 reviews research on category learning, including learning NLS categories such as *XOR* functions.

2.1 Mental models and representation of false information

Since 2000, Johnson-Laird and colleagues have continued to publish prolifically. In addition, a number of other researchers have conducted studies and proposed theories that either support or contradict theories relating to mental models and representation of false information.

A number of studies conducted by Johnson-Laird and colleagues have examined problem solving relating to the *XOR* function. They use the term "Exclusive Disjunction" to refer to the *XOR* function, and "Inclusive Disjunction" to refer to the *OR* function. However, the use of different terms does not alter the intended meaning of the function.

In [14], Barres and Johnson-Laird examined the extent to which people drew incorrect inferences for four types of problems, Conjunction (*AND*), Conditional (IF…THEN), Exclusive Disjunction (*XOR*), and Inclusive Disjunction (*OR*). Participants were given a rule, such as *Either there is an A or else there is a 1, but not both* (exclusive disjunction) or *If there is an A then there is a 1* (conditional), and were asked to write down combinations of letters and numbers that would make the rule true or false. For instance, the combination *not-A, 1* would make the exclusive disjunction rule true, but the combination *A, 1* would falsify it.

Results indicated that participants were significantly better at generating combinations that would make the rule hold true than they were at generating combinations to falsify the rule. No direct comparison was conducted between results for different types of problems, although the summary included in the paper (e.g. [14, Table 3, p9]) suggested that generating examples to make the rule true for Exclusive Disjunction problems was more difficult (72% correct) than Conjunction problems (81% correct), but easier than Conditional problems (58% correct). In contrast, generating examples to make the rule false for Exclusive Disjunction problems (23% correct) was more difficult than for Inclusive Disjunction (30% correct) and Conditional problems (49% correct), but marginally easier than Conjunction problems (20% correct).

A similar study was conducted by Johnson-Laird and Hasson [15]. In this study, participants were given a logical premises, such as *Either Dan is in Madrid or else Bill is in Seoul but not both*, and a conclusion, that either did or did not follow logically from the premise, such as *Therefore, Dan is in Madrid and Bill is in Seoul*, [15, p1106]. This is an exclusive disjunction, or *XOR*, premise, and in this instance the conclusion does not follow logically from the premise; if Dan is Madrid, then Bill is not in Seoul. Participants were presented with a variety of premises, using exclusive disjunctions, conditionals, and other relationships, and were asked to write down whether or not the conclusion was logical, and their reasoning. Results suggested that participants were poorer at evaluating the conclusions to exclusive disjunction premises than conditionals. However, Johnson-Laird and Hasson do not report if this difference was statistically significant, as the focus of their study was comparing the types of strategies used to judge the logic of the conclusions rather than the difficulty in judging conclusions.

A similar study was conducted by Mackiewicz and Johnson-Laird [16]. Participants were given logical inferences of the form *A is taller than B or else B is taller than C. A is taller than B*, and asked to draw a valid conclusion. In this case, the valid conclusion is that B is not taller

than C. The inferences were either disjunctions (as per the example), or conditionals, such as *A is taller than B if and only if B is taller than C. A is taller than B.* A valid conclusion to this inference is that B is taller than C. Results indicated that participants were significantly better at drawing valid conclusions from conditionals than from disjunctions.

In [17], Khemlani and Johnson-Laird examined participants' ability to solve exclusive and inclusive disjunction problems. An inclusive disjunction problem, which they hypothesised would be easier to solve, took the form:

Problem 6

> Suppose that at least one of the following assertions is true, and possibly both:
>
> 1. You have the marshmallows.
>
> 2. You have the truffles or the jelly beans, and possibly both.
>
> Also, suppose you have the marshmallows. What, if anything, follows? Is it possible that you also have either the truffles or jelly beans? Could you have both? [17, p.618]

A sample exclusive disjunction problem, which Khemlani and Johnson-Laird hypothesised would be more difficult to solve, was:

Problem 7

> Suppose that only one of the following assertions is true:
>
> 1. You have the mints.
>
> 2. You have the gum or the lollipops, but not both.
>
> Also, suppose you have the mints. What, if anything, follows? Is it possible that you also have either the gum or the lollipops? Could you have both?

The correct answer to Problem 6 is that, if you have the marshmallows (Assertion 1 is true), it is logically possible that also have either the jelly beans or the truffles, or both (Assertion 2 is true). It is also logically possible that you have neither the truffles nor the jellybean (Assertion 2 is false). The correct answer to Problem 7 is that if you have the mints (Assertion 1 is true), then Assertion 2 must be false, and you can have both the gum and the lollipops, but cannot have only one.

The mental models theory predicts that error rates will be higher for the exclusive disjunction problem, and this was supported by the results. Inclusive disjunction problems such as Problem 6 were solved correctly by 100% of participants, compared to only 17% for exclusive disjunction problems such as Problem 7. This effect was lessened, but not eliminated, when participants were explicitly instructed that one assertion in the problem was false.

Further evidence on the difficulty of answering logical problems using exclusive disjunction, or *XOR* problems, is provided by Johnson-Laird and colleagues in [18]. Their study used four basic exclusive disjunctions:

Either A, or else not both B and C.
Not both A and B, or else C.
Either A, or else A and B and C.
Either A and B and C, or else C.

For each disjunction, four assertions were developed. Participants were presented with the disjunction and an assertion, and asked if both could possibly be true. For instance, the assertions for the first disjunction (either A, or else not both B and C) were:

Not-B and not-C
A and not-B
Not-A and B and C
A and B and C

The assertions were constructed so that, according to the mental models theory, partial models would lead to either correct answers (control problems) or incorrect answers (illusory problems). In addition, the correct answer could be either "consistent" (both statements could be true) or "inconsistent" (both statements could not be true). Results indicated that illusory problems were significantly more likely to be answered incorrectly than control problems. However, there was no significant difference in responses to "consistent" or "inconsistent" problems.

A similar study was conducted by Walsh and Johnson-Laird [19]. They used the disjunction *A or B, but not both*, and the conditional *A or B or both*. They tested four assertions for each disjunction: *A*, *not-A*, *B*, and *not*-B and asked participants what followed logically from the disjunction or conditional and the assertion. In addition, the disjunctions referenced either or one two individuals, who were conducting either the same or different activities. For instance, an example referencing one individual is: *Sarah is sitting in the armchair, or Sarah is opening the front door but not both. Sarah is opening the front door. What follows?* An example referencing two individuals conducting different activities is: *Brian is standing by the fireplace or Joanne is looking at the mirror but not both. Joanne is not looking at the mirror. What follows?* Results indicated that participants were significantly better at solving problems when they referred to only one individual. However, there were no significant differences between responses to disjunctions and conditionals.

Even more complex disjunctions were examined by Santamaría and Johnson-Laird [20]. In their study, participants were confronted with problems that contained disjunctions within disjunctions, such as:

Problem 8

> Only one of the two following assertions is true about John:
>
> 1. John is a lawyer or an economist, or both.
>
> 2. John is a sociologist or an economist, or both.
>
> He is not both a lawyer and a sociologist. Is John an economist?

The mental model theory predicts that constructing a partial mental model will lead to incorrectly answering "Yes". The correct answer is that John can never be an economist, for the same logical reasons that the Ace can never occur in Problem 2. Participants were asked to solve a variety of control and illusory problems. In addition, Santamaría and Johnson-Laird explored if the way the problem was worded affected results. Participants were either given the problem as worded above (which the authors term a 'logical disjunction' problem), or a problem where the wording made it clear that only one statement could be true. An example of this type of problem, which the authors termed a 'physical disjunction' problem, is:

Problem 9

> John was reading the newspaper looking for a job. There were two ads on that page but he cut out only the one that matched his qualifications:
>
> Job 1 was for a lawyer or an economist, or both.
>
> Job 2 was for a sociologist or an economist, or both.
>
> He is not both a lawyer and a sociologist. Is John an economist?

Results indicated that control problems were significantly more likely to be answered correctly than illusory problems. In addition, physical disjunction problems were significantly more likely to be answered correctly than logical disjunction problems. The interaction was significant, such that illusory problems were significantly easier to answer correctly if they contained physical rather than logical disjunctions, but no such effect was evident for the control problems.

The suggestion that the wording of the problem can reduce the likelihood of succumbing to an illusory problem is an interesting one, and one that has been explored by other researchers. For instance, Barrouillet and Lecas [13] suggested that Johnson-Laird's findings could be attributed to participants failing to correctly understand disjunctions. As an example, they used the following problem, originally reported in [6]:

Problem 10

> Suppose that you are playing cards and that you get two cards. You know that 'if the first card is a king, then the second card is an ace, or else if the first card is not a king, then the second card is an ace'. You look at the first card and you see that it is a king. What can you conclude about the second card?
>
> a. The second card is an ace.
>
> b. The second card is not an ace.
>
> c. Cannot say whether it is an Ace or not an Ace.

In [13], 100% of participants answered that the second card was an Ace. This answer is incorrect. In this problem, the use of the phrase 'or else' is intended to act as an XOR, identifying that only one statement about the hand of cards is true. That is, it is supposed to direct participants to the fact that the problem takes the logical form (*King and Ace) XOR (Not-King or Ace)*. When the problem is expressed this way, it is clear that the Ace can never occur, for the same logical reasons it cannot occur in Problem 2. Johnson-Laird assumes that the high error rates are due to incorrect mental models. However, Barrouillet and Lecas suggest that the high error rates instead occur because participants are confused about the meaning of 'or else' and incorrectly read the problem as *(King XOR Not-King) AND Ace*. When the problem is formulated in this way, it is true that the second card is an Ace.

In order to test if high error rates are caused by participants misunderstanding the logical form of the problem, Barrouillet and Lecas examined response rates to Problem 10 as presented above, and when it was preceded by the following vignette, which more explicitly states that problem:

> Suppose that you are playing cards and that you get two cards. The people sitting next to you say that they know how the cards have been shuffled and that they can predict the value of the second card from the value of the first. Paul, on your left, says that 'if the first card is a King, then the second card is an Ace'. Louis, on your right, says that 'if the first card is not a King, then the second card is an Ace'. You know that what one of them says is true, and that what the other says is false, but you don't know who is right and who is wrong.

Results show that the rates of participants answering incorrectly decreased significantly when participants were given the vignette in addition to the premise. However, the rates of participants answering correctly did not increase correspondingly. Rather, there was a significant increase in participants answering 'Cannot say'. A later study by Newsome and Johnson-Laird succeeded in increasing the rate of correct responses to illusory problems, but only when participants were given much more explicit instructions to think about the ways in which the premise could be falsified [21].

The implication of these studies is debated by their respective authors. While Barrouillet and Lecas [13] suggest these findings refute the mental models theory, Johnson-Laird disagrees [22]. The disagreement relates to the structure of the partial mental model that is constructed, rather than the validity of the finding that illusory problems are more difficult to solve.

In addition, neither Johnson-Laird nor Barrouillet and Lecas acknowledge a potentially confounding factor. That is, adding the vignette makes the problem less abstract and more concrete. It has been previously demonstrated with other reasoning problems that changing the problem from abstract to concrete results in higher rates of correct answers [23-25].

Another explanation for Barrouillet and Lecas' findings [13] is that, as suggested by Galanis et al, [1] participants treat the *XOR* as an *OR*. If they incorrectly identify the location of the disjunction, as Barrouillet and Lecas suggest [13], the problem becomes *(King OR not-King) AND Ace*[5]. Alternatively, if the location of the disjunction is correctly identified, the problem becomes *(King AND Ace) OR (not-King AND Ace)*. Irrespective of where the disjunction is located, if the *XOR* is treated as an *OR*, then it is logically true that the second card must be an Ace. Either way, Galanis et al.'s suggestion is a plausible explanation for the findings.

Recently, Johnson-Laird has begun to explore Boolean concepts in relation to the mental models theory [26, 27]. In [26], Goodwin and Johnson-Laird considered that Boolean functions, like other forms of logic, generate mental models. In accordance with the mental models theory, they proposed that explicitly false information would not be represented in the models. This can lead to susceptibility to illusions, particularly with the *XOR* function, where only one clause can be true at any time.

For instance, consider the function (*A AND B) XOR B*. According to the mental models theory, this will produce the partial mental model shown in Table 5. The first line represents the first clause, (*A AND B)*, and the second line represents the second clause, *B*.

Table 5: Partial mental model for (A AND B) XOR B

A	B
	B

This model does not account for the fact that if the first clause is true, the second must be false. That is, if (*A AND B)* is true, *B* must be false; and if (*A AND B*) is false, *B* must be true. Therefore, *A* can never occur, and the only possible value that makes the function true is *B*. The mental model theory predicts that people will incorrectly answer that it is possible for A to occur.

To test this, Goodwin and Johnson-Laird [14] generated a variety of two-value Boolean functions. These were designed so that partial mental models would yield the same answer as the complete mental model (control problem) or would not yield the same answers as the complete mental model (illusory problem). Participants were told that each function described a set of objects, and that they should write down all possible descriptions of the objects. For instance, given the function *red and square or else not square* (the equivalent of the function discussed above), the only logically correct answer is that the objects are square and not red, although it is an illusory and plausible answer that the objects may also be red and square. Results indicated that illusory problems were significantly more likely to be answered incorrectly.

[5] Note that if the problem is constructed in this way, *(King OR not-King)* is removed from the equation under the law of excluded middle, leaving only the *Ace*.

In [27], Goodwin and Johnson-Laird describe two experiments that have strong similarities to the ERP. In both experiments, participants were tested on their ability to learn nine Boolean functions, each containing three variables. Participants were given a computer-based display of three light switches. Whether the light was on or off was controlled by a Boolean function. For each function, participants were given five minutes to try and establish the conditions under which the light would be switched on, by testing different combinations of switches. Once the five minutes had elapsed, or once participants felt they understood the Boolean function, they were asked to write down their explanation. The variables that were of interest to Goodwin and Johnson-Laird were the accuracy of explanations relative to the number of mental models produced by each function.

In order to illustrate this, Table 6 is a partial reproduction of Table 5 from [27]. The second column lists the functions used in Experiment 1. In Experiment 2, the functions were altered by replacing every instance of *B* with *not-B* and vice versa. According to the mental models theory, each function will produce a separate instantiation (or, using Johnson-Laird's terminology, a separate model) for each combination of values that make the function true. For instance, the first function produces only one model, (*A AND not-B)*, as this is the only combination of values that make the function true. The second function produces two models, (*A AND B)* and (*not-A AND not-B)*, where each combination of values will make the function true. The third column lists the number of models produced by each function.

Table 6: Functions and accuracy from Goodwin and Johnson-Laird (2011)

Function no.	**Function**	**No. models produced**	**% correct Exp. 1**	**% correct Exp. 2**
1	A AND ┐B	1	96	100
2	A XOR ┐B	2	82	89
3	A OR ┐B	2	57	92
4	A AND (┐B XOR C)	2	79	100
5	(A XOR ┐B) AND (A XOR C)	2	89	81
6	(A AND ┐B) OR (B AND C)	2	59	48
7	A XOR (┐B AND C)	3	39	67
8	A OR (┐B XOR C)	3	64	58
9	(A XOR ┐B) OR (A XOR C)	4	68	100

The final two columns show the percentage of correct answers in Experiments 1 and 2. Results indicated that with one exception, Function 9, accuracy decreased significantly as the number of mental models produced by the function increased. In addition, the length of time it took to describe the function and the number of tests participants performed increased significantly as a function of the number of models.

Function 9, the exception to these patterns, is the only function producing four models where the light is switched on. During analysis, the researchers noted that in Experiment 1, this function produces two mental models of cases where the light is switched off. Participants were learning these two cases, rather than the four where the light was switched on, presumably because the smaller number made it easier to memorise. A similar pattern was observed in Experiment 2. Hence, Function 9 was excluded from analysis.

While both this study and the ERP project use light switches and Boolean algebra concepts, there are some differences. Most notably, Goodwin and Johnson-Laird [27] do not consider concepts of LS and NLS, although my own analysis of their functions indicates that only Functions 1 and 3 are LS, and the rest are NLS. It is not appropriate to compare response rates to LS and NLS functions in this study due to the potentially confounding effect of the number of models produced by each function. In addition, in Goodwin and Johnson-Laird's study, a number of the functions are reducible; that is, some terms are irrelevant in determining whether or not the light is switched on. For instance, in the first three functions, the value of C is irrelevant in determining the state of the light switch. All problems used in the ERP project are unreducible, where all values must be considered in order to reach a correct answer.

2.2 Category learning

As discussed in Section 1.4, previous research into the extent to which NLS categorisation can be learned has produced mixed results. A literature review by Ashby and Maddock [11] supported this, identifying a number of studies where NLS categories were more difficult to learn than LS, but also studies where both categories were learned equally well. Some recent studies have further examined the extent to which NLS categorisation can or cannot be learned.

In 2001, Blair and Homa [7] suggested that the reason some studies had failed to find difficulties in learning NLS categories was because they had used only two categories, with a small number of category members. They felt that this did not adequately challenge participants, and that learning NLS categories may be more difficult with a larger number of categories and category members. To test this, they conducted a study using four categories, with three or nine members per category. Two of the categories were LS and two were NLS. Results showed that LS categories were easier to learn than NLS for both 3-member and 9-member conditions. A small number of participants were totally unable to learn the NLS categories; this effect was more pronounced in the 9-member conditions than in the 3-member conditions.

In the same year, a study was conducted by Ashby and colleagues [28]. They examined the extent to which participants were able to categorise simple visual stimuli, straight lines. As a between-subjects variable, there were either two or four categories. In both conditions, category membership was determined by quadratic functions. Ashby and colleagues note that the category boundaries were inexact, meaning that some items belonged to more than one category. Consequently, both the two and four category functions were NLS, as it was impossible for a single straight line to differentiate between category memberships.

Results indicated that participants were more accurate at categorising items in the two category condition compared to the four category condition. Ashby and colleagues [28] analysed the results in more detail using a variety of decision-making models, to identify the optimum, and more frequently used, decision strategies. They found that while NLS functions such as *XOR* gave best separation between categories, participants instead tended to base decisions on LS functions. While these may have been easier, they resulted in less optimum

categorisation. The authors conclude that these findings support the view that there is a NLS constraint on categorisation and decision-making.

A similar study was conducted by Maddox and colleagues [29]. In their study, participants categorised lines into two or four categories. As with [28], category boundaries were inexact, so that the functions were NLS. Results indicated that participants were better at learning categorisation in the two category condition than in the four category condition. In addition, similar to [28], analysis of the results using decision-making models found that approximately half the participants in the 4 category condition used suboptimal rules to guide their decision-making. Maddox and colleagues do not explicitly state if these suboptimal rules were LS.

The extent to which inexact category boundaries affects category learning was further explored by Ell and Ashby [30]. In their study, participants learned two categories, with varying degrees of overlap. Results indicated that higher degrees of overlap resulted in poorer learning. In the conditions with the highest degrees of overlap, performance on the final trials was not significantly greater than chance. That is, participants' performance was no better than what would be predicted if they were guessing.

The extent to which NLS categories could be learned was further explored by Hoffman and Redher [31, 32]. They examined the extent to which four different category types could be learned. The first category was LS, while the second category was a NLS category derived from an *XOR* function. The third and fourth categories are not directly relevant to the literature review. As an additional area of interest, this study used eye tracking equipment to measure the direction and duration of participants' gaze. Results indicated that the NLS took twice as many trials (14 vs. 7) to learn vs. the LS category, adding to the suggestion that NLS categories are more difficult to learn.

The eye tracking data in conjunction with the error rates across time provide some interesting insights into the way that participants approached the task. Where participants consistently looked at all aspects of the stimuli across the experiment, the researchers suggested that this was consistent with participants memorising the correct responses. In contrast, where participants abruptly ceased looking at aspects of the stimuli that were irrelevant to categorisation, and confined their gaze to relevant aspects, the researchers concluded that these participants had deduced the underlying rule that determined categorisation, and were proceeding to apply the rule. In the first group, which the researchers termed 'memorisers', error rates slowly declined, whereas in the second group, termed 'rule learners', there was a sudden and pronounced drop in error rates almost to floor level [32]. The tendency for rule learning behaviour to occur was more pronounced in the LS category than in the NLS. This suggests that learning the underlying rules was more difficult in the NLS category, and hence participants were resorting to memorising the correct answers, which was a suboptimal strategy as it was susceptible to forgetting.

Research has also examined LS and NLS category learning in humans and other animals. For instance, Smith and colleagues [12] compared learning rates for *XOR* categorisation in humans and monkeys. Both groups of participants were presented with stimuli comprising two, three, or four dimensions, and were asked to label them as belonging to one of two categories. Categorisation was determined by an *XOR* function, and immediate feedback was given following each trial. Results indicated that both species found *XOR* categorisation

difficult to learn, although the human participants outperformed the monkeys. The authors suggest that human participants were learning the underlying rules for each category, rather than simply memorising correct responses. This study follows previous research by Smith and colleagues examining category learning in pigeons and humans [33].

There has also been research in a number of other areas, for instance, several studies have used eye-tracking instruments to measure what aspects of a stimulus participants study, and how long they study it, in making classification decisions [34, 35]. In addition, studies by Lewandowsky and colleagues [36, 37] have examined the role that working memory capacity plays in categorisation decisions. Their studies have demonstrated that categorisation performance is correlated with working memory capacity. Their studies also showed that there are individual differences in categorisation: when presented with categorisation tasks, some participants will tend to use rule-based strategies, while others will use exemplar-based strategies.

One area of research that is relevant to the ERP is the distinction between classification and inference. As described by Markman and Ross [38], classification is the process of determining which category an item belongs to, given certain dimensions. In contrast, in inference, the category and some dimensions are given, and an inference must be made about values on another dimension. The examples given by Markman and Ross are predicting a person's political affiliation given their stance on certain political issues (classification task), or predicting a person's stance on a given political issue given their political affiliation and stance on other issues (inference task). Although not empirically tested, Markman and Ross suggested that NLS categories are more difficult to learn through inference rather than through classification.

A study conducted by Yamauchi and colleagues [39] examined the extent to which NLS categories could be learned through classification and inference. Their study used two NLS categories, with four dimensions (triangle vs. circle, green vs. red, small vs. large, and left vs. right). Participants learned either through classification, or through inference. In the classification condition, participants were given values on all four dimensions, and asked to predict which category the stimulus belonged to. In the inference condition, participants were given values on three of the four dimensions, as well as the category membership, and were asked to predict the value on the fourth dimension. Results indicated that inference learning was significantly more difficult than category learning. In the inference condition, fewer participants reached criterion levels of performance, and those who did reach criterion levels took significantly more trials than participants in the classification condition. On a test examining transfer of learning, participants who learned through classification also performed significantly better than participants who learned through inference.

3. Discussion

At the time the DSTO ERP project was developed, it had several novel elements that distinguished it from previous research. These included:

- A Boolean algebra approach to the Johnson-Laird problem,
- Testing the ease of learning and comprehending LS and NLS functions, and
- Using pictures rather than written problems.

In the years since the DSTO ERP project was first proposed, a variety of research on mental models and linear separability has been published. Some of these studies are similar to the ERP project, and are discussed in this section.

Johnson-Laird's research has continued to demonstrate that people are susceptible to errors in solving logical problems which, according to the mental models theory, require representation of false information. A number of these studies have demonstrated that problems including an *XOR* function are more difficult to solve correctly than functions including an *OR* function [14-20]. This is consistent with the hypothesis of the ERP project. However, these studies are distinct from the ERP project in two main ways. Firstly, Johnson-Laird does not consider concepts of LS in his work. Secondly, he does not test learning or comprehension of the functions, only the extent to which they can be solved on a single presentation.

Johnson-Laird's recent work includes one study which uses similar methodology to the ERP. In Goodwin and Johnson-Laird's study [27], participants solved problems where Boolean functions using a combination of light switches determined whether or not a light was switched on. However, this research is quite distinct from the ERP project for a number of reasons. Firstly, the variable of interest in their study was the number of mental models generated by each function. Secondly, Goodwin and Johnson-Laird do not consider concepts of linear separability. While the functions used in their study were a mixture of LS and NLS functions, no attempt was made to compare response rates to the two function types, and any comparison would be confounded by the fact that the LS functions were reducible. Finally, there was no test of comprehension.

A number of studies have examined difficulty in learning NLS functions, or solving problems based on NLS functions. These studies have demonstrated that, in general, NLS functions are more difficult to learn than LS functions, particularly as the number of categories increases [7, 28-30]. Some studies have specifically examined *XOR* functions as a category of NLS, and found that these functions are more difficult to learn than the LS *OR* function [12, 31, 32]. However, these studies differ from the ERP project in a number of ways. Firstly, while these studies did test the extent to which the functions could be learned, they did not test comprehension of the functions. Secondly, the types of categories used in these studies were quite different to the types of categories used in the ERP project; lines and abstract shapes rather than light switches. In addition, unlike the ERP project, in a number of these studies there were overlaps between category boundaries; that is, some items belonged to more than one category [28-30].

The literature review found only one study that tested participants' ability to comprehend rather than learn functions. In Yamauchi's study [39], participants either classified items into categories, or made predictions about item dimensions given a category. The latter, known as inference, is similar to the comprehension phase of the ERP project, where participants are asked to make predictions about the state of a particular light switch, given the state of the light (on or off) and the state of one or two of the remaining switches (shaded or unshaded). Consistent with the hypothesis or the ERP project, the study found that inference was difficult for NLS functions. However, this study did not compare inference for LS and NLS functions.

In conclusion, a number of studies have been conducted that touch on areas similar to those proposed by the ERP project. However, there are still sufficient unique elements to the ERP project such that it makes a novel contribution to the body of knowledge.

4. Acknowledgements

The enabling research project was originally developed by George Galanis and Armando Vozzo. I would like to thank them for providing background and context to this literature review.

5. References

1. Galanis, G., et al. (2011) Errors in human reasoning: Exploration of the implication of Johnson-Laird's falsity model
2. Sparkes, J. and Huf, S. (2003) *Mental models theory and military decision-making: A pilot experimental model [DSTO-GD-0368]*. Department of Defence, Editor. Edinburgh, SA. Department of Defence,.
3. Whitney, S., Galanis, G. and Vozzo, A. (in prep) *Linear separability in category and inference (DSTO-TR-XXXX)*. Edinburgh. DSTO.
4. Gregg, J. R. (1998) *Ones and zeros: Understanding Boolean algebra, digital circuits, and the logic of sets*. New York, IEEE Press
5. Johnson-Laird, P. N. and Savary, F. (1996) Illusory inferences about probabilities. *Acta Psychologica* **93** (1-3) 69-90
6. Johnson-Laird, P. N. and Savary, F. (1999) Illusory inferences: A novel class of erroneous deductions. *Cognition* **71** (3) 191-229
7. Blair, M. and Homa, D. (2001) Expanding the search for a linear separability constraint on category learning. *Memory & Cognition* **29** (8) Dec 1153-1164
8. Wattenmaker, W. D., et al. (1986) Linear separability and concept learning: Context, relational properties, and concept naturalness. *Cognitive Psychology* **18** (2) 158-194
9. Medin, D. L., Dewey, G. I. and Murphy, T. D. (1983) Relationships between item and category learning: Evidence that abstraction is not automatic. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **9** (4) 607
10. Medin, D. L. and Schwanenflugel, P. J. (1981) Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory* **7** (5) 355
11. Ashby, F. G. and Maddox, W. T. (2005) Human category learning. *Annual Review of Psychology* **56** 149-178
12. Smith, J. D., Coutinho, M. V. C. and Couchman, J. J. (2011) The learning of exclusive-or categories by monkeys (*Macaca mulatta*) and humans (*Homo sapiens*). *Journal of Experimental Psychology: Animal Behavior Processes* **37** (1) 20
13. Barrouillet, P. and Lecas, J. F. (2000) Illusory inferences from a disjunction of conditionals: A new mental models account. *Cognition* **76** (2) 167-173
14. Barres, P. and Johnson-Laird, P. (2003) On imagining what is true (and what is false). *Thinking & Reasoning* **9** (1) 1-42
15. Johnson-Laird, P. and Hasson, U. (2003) Counterexamples in sentential reasoning. *Memory & Cognition* **31** (7) 1105-1113
16. Mackiewicz, R. and Johnson-Laird, P. N. (2012) Reasoning from connectives and relations between entities. *Memory & Cognition* 1-14
17. Khemlani, S. and Johnson-Laird, P. (2009) Disjunctive illusory inferences and how to eliminate them. *Memory & Cognition* **37** (5) 615-623
18. Johnson-Laird, P., Lotstein, M. and Byrne, R. M. J. (2012) The consistency of disjunctive assertions. *Memory & Cognition* 1-10
19. Walsh, C. R. and Johnson-Laird, P. (2004) Co-reference and reasoning. *Memory & Cognition* **32** (1) 96-106
20. Santamaría, C. and Johnson-Laird, P. (2000) An antidote to illusory inferences. *Thinking & Reasoning* **6** (4) 313-333
21. Newsome, M. R. and Johnson-Laird, P. (2006) How falsity dispels fallacies. *Thinking & Reasoning* **12** (2) 214-234

22. Johnson-Laird, P. (2000) Illusions and models: A reply to Barrouillet and Lecas. *Cognition* **76** (2) 175-178
23. Manktelow, K. I. (1981) Recent developments in research on Wason's selection task. *Current Psychological Reviews* **1** (3) 257-268
24. Atran, S. (2001) A cheater detection module? Dubious interpretations of the Wason Selection Task. *Evolution and Cognition* **7** (2) 1-7
25. Stenning, K. and Van Lambalgen, M. (2004) A little logic goes a long way: Basing experiment on semantic theory in the cognitive science of conditional reasoning. *Cognitive Science* **28** (4) 481-529
26. Goodwin, G. P. and Johnson-Laird, P. (2010) Conceptual illusions. *Cognition* **114** (2) 253-265
27. Goodwin, G. P. and Johnson-Laird, P. N. (2011) Mental models of Boolean concepts. *Cognitive Psychology* **63** 34-59
28. Ashby, F. G., et al. (2001) Suboptimality in human categorization and identification. *Journal of Experimental Psychology: General* **130** 77-96
29. Maddox, W. T., Filoteo, J. V. and Hejl, K. D. (2004) Category number impacts rule-based but not information-integration category learning: Further evidence for dissociable category-learning systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **30** (1) 227
30. Ell, S. W. and Ashby, F. G. (2006) The effects of category overlap on information-integration and rule-based category learning. *Attention, Perception, & Psychophysics* **68** (6) 1013-1026
31. Hoffman, A. B. and Rehder, B. (2006) Linear separability and concept learning: Eyetracking individual differences. In: *47th Annual Meeting of the Psychonomics Society,* Houston, TX: 19 November
32. Rehder, B. and Hoffman, A. B. (2005) Eyetracking and selective attention in category learning. *Cognitive Psychology* **51** (1) 1-41
33. Cook, R. G. and Smith, J. D. (2006) Stages of abstraction and exemplar memorization in pigeon category learning. *Psychological Science* **17** (12) 1059
34. Blair, M. R., et al. (2009) *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, Austin, TX, Taatgen, N. A. and van Rijn, H. (eds.)
35. Watson, M. R. and Blair, M. R. (2008) Attentional allocation during feedback: Eyetracking adventures on the other side of the response
36. Craig, S. and Lewandowsky, S. (2012) Whichever way you choose to categorize, working memory helps you learn. *Quarterly Journal of Experimental Psychology* **65** (3) 439-464
37. Lewandowsky, S. (2011) Working memory capacity and categorization: Individual differences and modeling. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **37** (3) 720
38. Markman, A. B. and Ross, B. H. (2003) Category use and category learning. *Psychological Bulletin* **129** (4) 592
39. Yamauchi, T., Love, B. C. and Markman, A. B. (2002) Learning nonlinearly separable categories by inference and classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **28** (3) May 585-593

DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION
DOCUMENT CONTROL DATA

1. PRIVACY MARKING/CAVEAT (OF DOCUMENT)

2. TITLE

Literature Review on Mental Models and Linear Separability

3. SECURITY CLASSIFICATION (FOR UNCLASSIFIED REPORTS THAT ARE LIMITED RELEASE USE (L) NEXT TO DOCUMENT CLASSIFICATION)

Document	(U)
Title	(U)
Abstract	(U)

4. AUTHOR(S)

Susannah J. Whitney

5. CORPORATE AUTHOR

DSTO Defence Science and Technology Organisation
PO Box 1500
Edinburgh South Australia 5111 Australia

6a. DSTO NUMBER	6b. AR NUMBER	6c. TYPE OF REPORT	7. DOCUMENT DATE
DSTO-GD-0741	AR-015-591	General Document	April 2013

8. FILE NUMBER	9. TASK NUMBER	10. TASK SPONSOR	11. NO. OF PAGES	12. NO. OF REFERENCES
2012/1067434/1	LOD.3 - LHS ER&D	CLOD	24	39

13. DSTO Publications Repository

http://dspace.dsto.defence.gov.au/dspace/

14. RELEASE AUTHORITY

Chief, Land Operations Division

15. SECONDARY RELEASE STATEMENT OF THIS DOCUMENT

Approved for public release

OVERSEAS ENQUIRIES OUTSIDE STATED LIMITATIONS SHOULD BE REFERRED THROUGH DOCUMENT EXCHANGE, PO BOX 1500, EDINBURGH, SA 5111

16. DELIBERATE ANNOUNCEMENT

No Limitations

17. CITATION IN OTHER DOCUMENTS Yes

18. DSTO RESEARCH LIBRARY THESAURUS

Decision making, literature surveys, cognitive processes

19. ABSTRACT

The mental models theory suggests that people make reasoning errors because they construct partial – and inaccurate – mental models. It predicts that where people are required to consider false information, they are more prone to making errors than when they are only required to consider true information. Findings consistent with this theory have been demonstrated across a number of studies, particularly the work of Johnson-Laird. However, researchers at DSTO suggested that these findings are better explained by a linear separability effect. That is, that the distinction between problems that are easy to solve and difficult to solve is between is better predictated by whether they are linearly separable or inseparable than whether they require consideration of false information. This literature review examines research on mental models and linear separability published between 2000 and 2012, to establish if this explanation has been proposed by other researchers. Results indicate that no other researchers have proposed this, or similar, explanations, hence the linear separability hypothesis has the potential to make a novel contribution to the literature.