Lying about intentions

Lara Natasja Warmelink

May 2012

The thesis is submitted in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy of the University of Portsmouth.
I would like to dedicate my thesis to my grandmothers.

To mrs. Warmelink-Minkjan, in the hope she will enjoy reading the fruits of my labour and to mrs. Grovenstein-Tijsma, who will now not be able to.

Before the throne of the Almighty, man will be judged not by his acts but by his intentions. For God alone reads our hearts.

(Mohandas Karamchand Gandhi)
Abstract

This thesis aims to explore whether lying about intentions can be detected using the same methods that are used to detect other types of lie. Part I, the introduction, contains a literature review of research into future thinking and intentions.

In Part II non-verbal cues and thermal imaging are investigated as cues to deception. In Experiment 1 participants of twelve different ethnicities were interviewed at an airport about their forthcoming trip. Half of the participants were instructed to answer truthfully, the other half to answer deceptively. The participants’ veracity affected only three of the 17 non-verbal cues: struggling, thinking hard and controlling behaviour. These cues were reduced to two factors, cognitive load and animation. These two factors could correctly classify 52.3% of truth tellers and 62.6% of liars. The thermal imaging showed that liars tended to get warmer during the interview, while the truth tellers’ temperatures remained constant. This temperature rise was a cue to deception at above chance rates.

In Part III two experiments are described, investigating verbal cues to deceit in longer and shorter interviews. In Experiment 2 participants were interviewed, answering truthfully or deceptively, about a forthcoming trip. The interview contained general questions that were expected by the participants and three types of unexpected questions. Half of the participants had been to the travel destination previously, while the other half had not been there. The results showed that truth tellers generated more details on the unexpected questions than liars, while liars provided more details on expected questions than truth tellers. This pattern was more pronounced with participants who had been to their travel destination before than for
those who had never been to that destination. A positive predictive value of .69
(truth) and a negative predictive value of .75 (lie) were obtained. Experiment 3
investigated verbal cues to deceit in a short one question interview about a
forthcoming trip. The interview question was asked in two forms, one prompting for
time detail and one control. The results showed that truth tellers and those answering
the time prompting question were more likely to provide temporal information. The
results showed that 47.6% of truth tellers and 81% of liars could be classified
correctly on the basis of the time prompt question.

Part IV investigated reaction times as a cue to deception. In Experiment 4,
participants were asked to perform a mission, either immediately or two weeks later.
Half of the participants was instructed to answer any questions asked by
experimenters en route truthfully, the other half was asked to answer deceptively.
Using a computer, participants classified adjectives as positive or negative and were
primed by words related to the real intention of truth tellers (to buy a present), which
was also the cover story for liars. Contrary to the hypothesis, in the immediate
condition liars answered faster on positive than negative adjectives, while the truth
tellers responded equally fast to both adjectives. In the delayed condition no
difference was found between truth tellers and liars.

Part V presents the general discussion. The main findings of the thesis are
summarised, practical and theoretical implications are discussed, the methodological
and ethical limitations are considered and ideas for future research are given.
Overall, the conclusion of the thesis is that false intentions contain cues to deceit that
observers should be able to spot.
Contents

Abstract 3
Contents 5
Declaration 10
List of Tables 11
List of Figures 13
List of Abbreviations 14
Acknowledgements 15
Dissemination 16
Part I: Introduction and Literature Review 17
   Chapter 1: Introduction to Thesis 17
      Introduction. 17
      Thesis outline. 19
   Chapter 2: Episodic Future Thought Literature Review 22
      Intention and mental time travel. 22
      The clarity and level of detail of the mental image. 23
      Intentions and non-intended future thought. 25
Part II: Non-Verbal Behaviour and Thermal Imaging as Lie Detection Cues 27
   Chapter 3: the Method for Experiment 1 and the Questionnaire Results of Experiment 1 27
      Method. 30
         Participants. 30
         Design. 30
         Procedure. 32
Dependent variables. 69

Results. 70

Mean skin temperatures as a function of veracity. 70

Truth – Lie Classifications, Between-Subjects Analyses. 71

Truth – Lie Classifications, Within-Subjects Analyses. 73

Discussion. 76

Thermal Imaging Results. 76

Interviewers’ Assessments. 77

Issues to Consider. 79

The Practical Application of Thermal Imaging. 83

Part III: Verbal Cues. The Effect of Interviewees’ Previous Experience and Time Prompting

Chapter 6: Using Verbal Cues to Detect Deception 84

Literature Review

Chapter 7: Experiment 2. The Effect of the Interviewees’ Previous Experience 92

Method. 96

Participants. 96

Design. 96

Procedure. 98

Data analysis. 100

Results. 102
Expectedness of the questions. 102

Questionnaire results. 103

Interview content. 104

Discussion. 113

Chapter 8: Experiment 3. The Effect of Time Prompting 119

Method. 121

Participants. 121

Design. 121

Procedure. 121

Data analysis. 123

Results. 123

Discussion. 124

Part IV: The Effect of True and False Intentions on Reaction Times 127

Chapter 9: Detecting Lies about Intentions Using a Lexical Decision Reaction Time Task 127

Method. 131

Participants. 131

Design. 132

Procedure. 133

Materials. 135

Data analysis. 137

Results. 138

Discussion. 141

Part V: General discussion 144

Chapter 10: General Discussion 144
Summary of findings. 144
Theoretical implications. 146
Practical implications. 149
Limitations. 151
  Methodological considerations. 151
  Ethical considerations. 153
Future Research. 155
Conclusion. 158
References 160
Appendix 1 174
Declaration.

Whilst registered for the above degree, I have not been registered for any other research award. The results and conclusions embodied in this thesis are the work of the named candidate and have not been submitted for any other academic award.
List of tables

3.1 The mean (SD) of difficulty per phase. 35
4.1 The correlation between the first and the second coder. 45
4.2 The means (and standard errors) for Ethnicity on those variables where Ethnicity had a significant effect. 49
4.3 The means (and standard errors) for English on the variables where English had a significant effect. 51
4.4 The means (and standard errors) for Phase. 52
4.5 The six components with eigenvalues over 1 extracted from the factor analysis. 54
4.6 The means and (SD) of Ethnicity for the animation factor. 55
5.1 Skin Temperature as a Function of Phase. 71
5.2 Area under the ROC Curves and Related Statistics and Discriminant Analyses Regarding Truth/Lie Classifications. 75
6.1 The CBCA Criteria. 86
6.2 The scales of the JMCQ. 89
7.1 The correlation between the first and the second coder. 102
7.2 The three significant multivariate effects for those variables with a significant three way interaction effect. 105
7.3 The mean and (SD) of the amount of total details per questions category, experience level and veracity. 106
7.4 The mean and (SD) of the amount of visual details per questions category, experience level and veracity. 107
7.5 The mean and (SD) of the amount of spatial details per questions category, experience level and veracity. 109
7.6 The mean and (SD) of the amount of temporal details per questions category, experience level and veracity.

7.7 The mean and (SD) of the amount of explanations per questions category, veracity and experience level.

7.8 The three way interaction effects for the 12 variables where this effect was non-significant.

8.1 The percentage of participants from each condition that did or did not mention a time in their answers.

9.1 The stimulus words used in the adjective classification task. The mean (and SD) of the participants’ rating of the primes is listed.
List of Figures

5.1 An example of artefacted data. 69
5.2 Skin temperature as a function of phase 76
9.1 The difference between positive and negative adjectives. 140
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full term</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACID</td>
<td>Assessment Criteria Indicative of Deception</td>
</tr>
<tr>
<td>ANOVA</td>
<td>Analysis of Variance</td>
</tr>
<tr>
<td>AUC</td>
<td>Area Under the Curve</td>
</tr>
<tr>
<td>CBCA</td>
<td>Criteria Based Content Analysis</td>
</tr>
<tr>
<td>CIT</td>
<td>Concealed Information Test</td>
</tr>
<tr>
<td>EEG</td>
<td>Electroencephalograms</td>
</tr>
<tr>
<td>fMRI</td>
<td>functional Magnetic Resonance Imaging</td>
</tr>
<tr>
<td>IAT</td>
<td>Implicit Association Test</td>
</tr>
<tr>
<td>JMCQ</td>
<td>Judgement of Memory Characteristics Questionnaire</td>
</tr>
<tr>
<td>M</td>
<td>Mean</td>
</tr>
<tr>
<td>MANOVA</td>
<td>Multivariate analysis of variance</td>
</tr>
<tr>
<td>MCQ</td>
<td>Memory Characteristics Questionnaire</td>
</tr>
<tr>
<td>MEG</td>
<td>Magneto Encephalograms</td>
</tr>
<tr>
<td>RM</td>
<td>Reality Monitoring</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver Operating Characteristic</td>
</tr>
<tr>
<td>SCAN</td>
<td>Scientific Content Analysis</td>
</tr>
<tr>
<td>SD</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td>SUE</td>
<td>Strategic use of Evidence</td>
</tr>
<tr>
<td>SVA</td>
<td>Statement Validity Assessment</td>
</tr>
</tbody>
</table>
Acknowledgements.

Many people have helped me while working on this thesis and I am grateful to all of them. I would especially like to thank my supervisors. First and foremost, I want to acknowledge Professor Vrij, my first supervisor, for all his help and advice. Bedankt Aldert, voor alle hulp en tips van statistiek tot hoe om te gaan met Engelsen. I would like to thank Dr. Mann, my second supervisor, for all the practical help she gave me, ta Sam. Also, thanks to my third supervisor, Professor Granhag of the University of Gothenburg, for keeping a wide view and an open mind, tack för all din hjälp med min avhandling.

I’d like to thank Kalle Ask and Melanie Knieps, also from the University of Gothenburg, for allowing me to use their unpublished work. My gratitude extends to Sharon Leal, who gave me much sound advice.

I want to thank the technicians and administration staff at the department of psychology at the University of Portsmouth, particularly Dave Forrester, who assisted greatly in Experiment 1 and Tina Harding, who has helped me on numerous occasions. In addition, I’m grateful to my colleagues for all their help and support: Dr. Jackie Hillman, Dr. Gary Lancaster, Shyma Jundi, Dom Shaw, Zarah Vernham, Sarah Rohlfing, Dr. Sabine Quandte, Gemma Graham, Dr. Victoria Devonshire, Ailsa Millen, Vivian Lago, Jo Fraser, Ed Agobiani and Isolde Sommer.

Finally, I’d like to thank my family for their emotional support. Especially, my mother, Sophia Grovenstein, my father, Bert Warmelink, his wife, Pia Teeuw and my siblings, Rosa, Niklas and Vera.
Dissemination.

Articles.


Conference presentations.


Part I: Introduction and Literature Review

Chapter 1: Introduction to Thesis

Introduction.

The subject of this thesis is lying about intentions. The aim is to explore whether lies about intentions can be detected using the same techniques that are used to detect lies about past events or opinions, and how detection rates can be maximised. Previous research, which focuses on lying about past events and opinions, has found that people in general are not very good at detecting lies. Bond and DePaulo (2006) and Vrij (2008) found that laypeople could on average detect 54% of truths and lies. Professional lie catchers were not much better at 55%. Many tools and methods have been developed to improve these detection rates, including physiological methods, non-verbal behavior analysis, statement content analysis and reaction time analysis.

Physiological lie detection uses intricate tools and equipment to monitor the body’s activity for cues associated with deception. Polygraphs take several measures including heart rate, breathing rate and sweating (Ben-Shakhar & Elaad, 2003). The temperature often participant is measured with thermal camera’s (Pavlidis, Eberhardt, & Levine, 2002). Electroencephalograms (EEG) (Farwell & Donchin, 1991), functional Magnetic Resonance Imaging (fMRI) (Spence, Farrow, Herford, Wilkinson, Zheng & Woodruff, 2001) and Magneto Encephalograms (MEG) (Seth, Iversen & Edelman, 2006) are used to measure the activity of the brain. The non-verbal behaviour of the subject has also been analysed as a cue to deception. This is a technique used not only used by professionals, but also sometimes by laypeople (Akehurst, Köhnken, Vrij & Bull, 1996), and it has been studied extensively by scholars (Bond & DePaulo, 2006; DePaulo et al., 2003; Zuckerman, DePaulo and
Rosenthal, 1981). Other methods have focused on the content of the statement, such as Statement Validity Assessment (SVA) (Köhnken, 2004; Masip, Sporer, Garrido, & Herrero, 2005; Vrij, 2005) or Reality Monitoring (Sporer, 2004). Another technique is to measure the response time to truth- or lie-related stimuli. Truth related stimuli usually elicit a shorter response time, probably because humans show an orienting reflex to familiar stimuli and the truth is often more familiar than the lie (Johnson, Barnhardt & Zhu, 2003; Verschuere, Spruyt, Meijer & Otgaar, 2010). These different techniques all improve lie accuracy rates compared to not using them. These accuracy rates are also above chance level, ranging between 60 and 90%, the accuracy rates for the different techniques vary substantially (Vrij, 2008).

Each method has its own strengths and weaknesses. In this thesis, I will examine a range of these methods: nonverbal behaviour, the temperature of the speaker, the speech content, and reaction times, and will examine the utility of each when used to detect lies about intentions.

Lying about intentions is a relatively new subject in lie detection (Gran Hag, 2010). Detecting lies about intentions is at least as important as detecting lies about past activities. Detecting lies about past activities might help law enforcement officers solve crimes that have already taken place, but detecting lies about intentions might help intelligence and security services prevent crimes from ever happening. One area of law enforcement that would benefit from detecting lies about intentions is border control and transportation security. The main goal for these systems is to let people without criminal intent pass through as quickly as possible while detaining those with criminal intentions (Burgoon et al., 2009).

One of the main problems in studying lies about intentions is that the ground truth is not as clear as it is with lies about past events (Gran Hag, 2010). This is
because past events are definite and knowable, even if they are currently unknown. In contrast, intentions are in the future and therefore not definite or knowable. Even if someone intends to do something, it does not necessarily mean that they will execute their intentions. This means that even if someone tells the truth about their intentions they still know that their intended actions might not happen and those who lie about their intentions might feel that they are not truly lying, since even if they do not intend to execute their false intention, it is possible that in fact they will do so. This creates a grey area between truth and lie that may be detrimental to differentiating truth tellers from liars. This thesis is an attempt to study whether it is possible to differentiate between truth tellers and liars when they are discussing their intentions.

**Thesis outline.**

This thesis contains the following 5 parts.

*Part I: Introduction and Literature Review.*

Part I contains Chapters 1 and 2. In Chapter 1 the entire thesis is introduced, sets its aims and contains the thesis outline. Chapter 2 is a literature review on episodic future thought and mental time travel. It discusses previous research on how people imagine the future, and the similarities and differences between that and remembering the past. It also discusses intentions and how they differ from other types of future thought.

*Part II: Non-Verbal Behaviour and Thermal Imaging as Lie Detection Cues.*

Part II contains Chapters 3, 4 and 5. Experiment 1 is described, which investigated several different non verbal cues to lying about intentions in an airport.
In Chapter 3 the experimental method is described and the results from the
questionnaire used in Experiment 1. In Chapter 4 non-verbal behaviour is introduced
as a cue to lie detection. Those parts of the method of Experiment 1 that are specific
to the non-verbal behaviour cues are described, together with the results and a
discussion of those results. In Chapter 5 thermal imaging is introduced as a lie
detection cue. Those parts of the method of Experiment 1 that are specific to the
thermal imaging technique are described, and the results of the experiment relating
to the thermal imaging are reported and discussed along with a discussion of the
potential of thermal imaging as a lie detection tool in an airport setting. A modified
version of Chapter 5 was published as Warmelink, L., Vrij, A., Mann, S., Leal, S.,
*Law and Human Behaviour, 35, 1*, 40-48 (see Appendix 1).

**Part III: Verbal Cues. The Effect of Interviewees’ Previous Experience
and Time Prompting.**

In Part III the effectiveness of interviewing participants and using the content
of their answers as a cue to detect deceit, is investigated. Part III contains Chapters 6,
7 and 8 and describes two experiments. Chapter 6 contains a literature overview of
how interviewing can be used to detect lies about past events, opinions and
intentions. In Chapter 7 Experiment 2 is described. It tests whether truthful and lying
interviewees who have previously experienced executing the intention they are
describing, talk about the intention in more detail than truth tellers and liars who
have never executed the intention they are discussing. In Chapter 8 Experiment 3 is
described, which also focuses on detail. It is investigates whether the phrasing used
in interview questions influences how much temporal detail is given.
Part IV: The Effect of True and False Intentions on Reaction Times.

Part IV contains only Chapter 9 and Experiment 4. The fourth experiment tests the reaction times of participants with a truthful or a false intention. In the experiment participants did a priming task to elicit their intentions. The goal of the study is to investigate whether those with a true intention respond differently from those with false intentions.

Part V: General Discussion.

The fifth and final Part contains Chapter 10. It contains an overview of the main findings of the thesis. In it the practical and theoretical implications of the findings, the methodological and ethical limitations of the thesis and ideas on possible future research are discussed. It ends with a final conclusion.
Chapter 2: Episodic Future Thought Literature Review

Intentions and mental time travel.

In order to investigate and detect lies about intentions we need an understanding of what a true intention is like. Humans often think about the future (D’Argembeau, Renaud & Van der Linden, 2011). Sometimes people purposefully imagine the future, sometimes these thought are triggered by an environmental or internal cue (Berntsen & Jacobsen, 2008). Not all of these thoughts are intentions. Intentions are defined as an agent’s mental state that represents an action that is done on purpose (intentionally) (Malle, Moses & Baldwin, 2001). It is important to distinguish between desires and intentions. When someone has a desire, they have not decided that they will take action to make the desire become reality. Whereas when someone has an intention they have decided to perform an action (Malle & Knobe, 2001). For this thesis intentions are defined as future actions that someone is committed to executing and of which that person has an image of how and in which situation they will be executed (Granhag, 2010; Goschke & Kuhl, 1993).

Humans have the ability to imagine themselves in past or future situations. This ability is known as mental time travel (Suddendorf & Corballis, 1997). This mental time travelling involves imagining the self into the event and “mentally experiencing” it. Remembering a past event in this way is known as episodic memory and imagining oneself into a future event is known as episodic future thought (Atance & O’Neill, 2001, 2005; Szpunar, 2010). It has been proposed that mental time travel into the past and into the future relate to each other closely (D’Argembeau & van der Linden, 2006). This theory has been supported by findings that imagining the future and remembering the past rely, both neurophysiologically and cognitively, on the same processes (Schacter & Addis, 2007). Evidence for this
is provided by studies of patients with severe amnesia, who were found to have trouble engaging in episodic future thought, as well as having problems with episodic memory (Schacter, Addis & Buckner, 2008). More support is found in neuroimaging studies, which led Schacter and colleagues to propose a “core system” of brain regions (the medial prefrontal cortex, the medial temporal lobe, the precuneus/retrosplenial cortex, the lateral parietal cortex and the lateral temporal cortex), which mediate mental time travel into both the past and the future (Schacter, Addis & Buckner 2007). This core system would be responsible for taking elements of memories and using them as the building blocks to reconstruct memories of past events and to construct images of future events.

**The clarity and level of detail of the mental image.**

Johnson and her colleagues developed the Memory Characteristics Questionnaire, in which participants could rate their memories for clarity and level of detail (Johnson, 1988; Johnson et al., 1988). This questionnaire has been successfully adapted for participants to report on mental images of future events (D’Argembeau and Van der Linden, 2004; 2006). Research using the questionnaires has shown that differences in clarity and detail depend on the participant and on the content of the image. For example, D’Argembeau and Van der Linden (2006) found that people who score higher on the Vividness of Visual Imagery Questionnaire, which measures how vivid their visual images are in general, also report more clear and detailed future images. This indicates that the ability to imagine detailed future images is related to the ability to imagine other detailed images. The content of the image also affects the self-reported clarity, D’Argembeau and van der Linden (2004) found that positive events both in the past and in the future are more detailed than negative events, according to self-reports. Also events that are closer in time to the
moment of reporting are more detailed than events that are further ahead in the future or further back in the past. Arzy, Adi-japha and Blanke (2009) found evidence that mental time travel is not done instantaneously. Participants take longer to respond to questions about events that are longer before or after a specified time. In other words, events that require mentally time travelling further. This indicates that when they mentally time travelled between two points they mentally passed through the time in between.

Gamboz, Brandimonte and De Vito (2010) found that the familiarity of a future image affected the clarity of the image. The participants’ self reports indicated that images of unfamiliar events were less clear on spatial and temporal information than past events, but there was no difference between past events and future events that were familiar. Szpunar and McDermott (2008) found similar results. Future events that took place in surroundings the participants knew (e.g. their home) were rated by the participants as clearer and as more detailed than future events in novel surroundings (e.g. a jungle). They also found that when people based their future image on a familiar setting, they reported the image as more clear and detailed when they had experienced that setting recently compared to settings they had experienced long ago. Szpunar, Chan and McDermott (2009) also found differences on an fMRI between imagining a familiar future and an unfamiliar one. Certain brain areas (the superior occipital gyrus and the posterior cingulate and parahippocampal cortices) were only active during remembering the past and imagining a future that was set in a familiar context. When imagining a future event in an unfamiliar context these brain regions were no longer active. They suggest that this means that these brain regions are used to imagine the familiar surroundings in which events are set, whether these events are past or future.
These studies suggest that it matters little whether someone time travels to the past or to the future. What matters is whether someone has suitable memories to base an episodic future thought on and how to use that episodic memory to reconstruct new images. However, both D’Argembeau and Van der Linden (2004) and Gamboz, Brandimonte and De Vito (2010) found that images of past events were more detailed than images of future events, which suggests that episodic future thinking is not as easy as remembering. Overall, the results suggest that remembering past events and imagining the future are closely related, but not entirely the same.

**Intentions and non-intended future thought.**

Intentions are a type of episodic future thought, but are different from ordinary future images in the sense that intentions require the commitment to execute the action in order to create the desired end goal. The human mind seems to prioritise intentions and intention related stimuli over other future related stimuli. Goschke and Kuhl (1993) found that participants respond faster to stimuli that are related to an intention than to stimuli that are related to a future image that will not be executed. This effect is known as the intention superiority effect. Goschke and Kuhl (1993) suggested that the effect is caused by a higher sub-threshold level of activation of the intention in long-term memory. Long term memory can be envisioned as a network of nodes that can be active or not and can spread their activation to other nodes. The activation of a node can cause an action, or can make the content of the node reach consciousness (be remembered). Goschke and Kuhl (1993) suggest that when an intention is not present in consciousness, its node remains active, but not active enough to reach consciousness. This sub-threshold activation means that less extra activation is needed to make the activation of the intention high enough to be
remembered. This makes it more likely that an intention will be remembered if related nodes become active and the intention node will become active faster than other nodes in the presence of related activity. Marsh, Hicks and Bryan (1999) have suggested that this heightened activation helps with remembering the intention and executing it. In fact, Watanabe (2005) found that more intentions than unintended future actions are remembered.

Several researchers investigated the intentions superiority effect. Dockree and Ellis (2001) found that when intentions are cancelled, so the action no longer has to be performed, they no longer cause the intention superiority effect on a lexical decision task. They also found that the participants did not show an intention superiority effect when they expected that they would have written instructions for the task with them when they performed the task. Presumably the written instructions negated the need for memorising the task and with it the need to keep the intention at high activation disappeared. Marsh and colleagues found that completed and cancelled intentions lack an intention superiority effect, in fact stimuli related to cancelled or completed tasks elicited significantly slower reaction times than neutral stimuli (Marsh, Hicks & Bryan, 1999; Marsh, Hicks & Bink, 1998). They suggest that cancelled and completed intentions are actively inhibited, to prevent intentions that no longer need to be executed from interfering with the current task. The theory that intentions are held in higher sub-threshold activation is supported by the finding that intentions related words are harder to ignore then neutral words. Marsh, Cook, Meeks, Clark-Foos and Hicks (2007) asked participants to respond to visual cues, while listening to an auditory stream of neutral or intention–related words, which they were asked to ignore. The results showed that the
intention related words in the audio stream were remembered better than the neutral words by the participants.

When someone is lying about their intentions, instead of reporting their true intention, they discuss another future event which they do not intend to create or execute. This future event can be a detailed visual image or a simple thought with no imagery attached (Granhag & Knieps, 2011). This chapter has outlined what these concepts mean and how they differ from each other. Both intentions and unintended future thoughts are constructed from different elements of memories. But intentions have a special priority in the mind. Intentions affect certain experimental tasks, such as reaction times test or memory tests more than other future thoughts. The differences between them could be the beginning of a way to detect lies about intentions. Measuring reaction times, the verbal and non-verbal behaviour of participants discussing intentions and unintended future thoughts might yield cues to classifying statements as real intentions or false intentions. In the next chapter we will start by looking at verbal and non-verbal cues in behaviour.
Part II: Non-Verbal Behaviour and Thermal Imaging as Lie Detection

Cues

Chapter 3: the Method for Experiment 1 and the Questionnaire Results of Experiment 1

Part II of this thesis contains three chapters and the first experiment, which investigates several non-verbal cues and thermal imaging as indicators of deception. In Chapter Three the method of Experiment 1 is described and the results from the questionnaire part of the experiment are reported. In Chapter Four the non-verbal cues that were investigated are discussed. The non-verbal cues will be introduced, the parts of the method specific to these cues will be discussed and the results will be reported and discussed. In Chapter Five thermal imaging is discussed and it contains an introduction to thermal imaging, the specific thermal imaging method, the results and a discussion of the method and its practical use.

The cues to deceit that are the focus of this experiment are based on three different theoretical approaches. The first is that liars are more aroused or more nervous than truth tellers. The second is that lying is more cognitively demanding than telling the truth. The third approach is that liars attempt to control their behaviour more than truth tellers in order to appear truthful (Vrij, Fisher, Mann & Leal, 2006; Hartwig, Granhag and Strömwall, 2007; Vrij, 2008).

Experiment One was conducted at a large, international airport. This allowed us to interview people from all over the world. Psychology research often relies on participants from the same area as the researchers. These participants do not represent mankind as a whole very well, in fact they have been described as WEIRD (Western, Educated, Industrialised, Rich and Democratic, Henrich, Heine &
These participants often show traits or responses that might not extend to the general public. In order to get a wider spread of participants, we recruited participants from 12 different regions of the world: Northern European, Eastern European, South West European, South East European, East Asian, South East Asian, Central Asian, Iranian, Arabic, South American, West African and East African. In order to be considered part of a region a participant would have to be: a national of one of the countries in that region, have lived there for an extended period and not be a recent immigrant from a country from a different region. The participant could currently be living out of the region. No North American region was included, because its population consist in large parts of recent immigrants and many people identify both with their country of origin and their “new” country. We were worried this would complicate the classification. In this thesis the twelve regions are called ethnicities. We also scored each participant on how well they spoke English.

The interview had five phases: ‘general’, ‘destination plans and past’, ‘plans on plane’, ‘reassurance’ and ‘lie’. The questions were always asked in this order and concerned, respectively: factual information about the trip; detailed information about the destination and past actions concerning the trip; information on what the participant will do on the plane; any extra information the participant can give to reassure the interviewer they are telling the truth and in the lie phase all participants were asked to lie about their plans at their destination, regardless of whether they had been lying before. It was hypothesised that overall the liars would find the interview harder than truth tellers. When examining this per stage, it was hypothesised that liars would find the first four phases harder than truth tellers, but that truth tellers would find the final ‘lie’ phase harder than liars, because they had to switch from truth telling to lying, whereas liars could continue lying.
Method.

Participants.

Four hundred and eighty one participants were recruited at a major airport, 17 participants were excluded from the analysis, because they did not follow the instructions completely. There were 331 (68.8%) men, 142 (29.6%) women and 8 participants did not answer this question on the questionnaire. Their age ranged from 16 to 75 with an average of 35 years (SD = 12.5). Two hundred and seven participants were going home (44 %). One hundred and eight were visiting friends or family (23%). Ninety-two participants (20%) were going on work or study related trips, e.g. travelling to their workplace, on business trips or going to a conference. Fifty-four (12%) were going on holiday. Two participants were going for medical reasons (0.4%).

Design.

Independent variables.

The experiment has a 2 by 12 by 5 factorial design. The first factor, Veracity, is a between-subject factor and has two levels: named truth telling and lying (both groups did the other task on one question). The second factor, Ethnicity, is also a between-subjects factor and has 12 levels (N = 40 for all groups, except West African, N = 39) and East African, N = 41):

- Northern European, e.g. Swedish
- Eastern European, e.g. Russian
- South West European, e.g. Spanish
- South East European, e.g. Greek
- Central Asian, e.g. Indian
- East Asian, e.g. Chinese
- South East Asian, e.g. Malaysian
- Iranian
- Arabian, e.g. Kuwaiti
- South American, e.g. Brazilian
- West African, e.g. Nigerian
- East African, e.g. Somali, Kenyan

The third factor, Phase, is a within subjects factor and has five levels, representing the five different phases of the interview. The first phase, *general*, contains questions 2 and 3 of the interview, these questions are just general information questions (e.g. “What would you say is the main purpose of your trip?”). The second phase, *destination plans and past*, contains questions 4 to 9. This phase is mostly dedicated to getting a detailed description of what the participant is going to do (e.g. “What are you going to do when you get off the plane?”). The third phase, *plans on plane*, contains questions 10 and 11. These questions ask what the participant plans to do on the plane (e.g. “What are you going to do on the plane?”). The fourth phase, *reassurance*, contains questions 12 to 16 and asks the participants for different types of information that could confirm that they are telling the truth (e.g. “Is there any information you can give to reassure me you are telling me the truth about your trip?”). The fifth phase, *lie*, contains question 17. This question asks what the participants are going to do on the day they arrive at their destination, but this time all participants have to answer deceptively. As a consequence, participants who were telling the truth must make up a lie. Participants who were already lying should think of a second lie, because if they repeat their original lie it will become clear to the interviewer that it was a lie. Question 1, “Where are you flying to?”, does not belong to any of the phases and all participants were instructed to answer it
truthfully. Hence, it was used as a baseline in the thermal part of the experiment, in which the interview therefore consists of six phases. The complete list of questions is confidential for security reasons and we are not allowed to publish it in full.

*Dependent variables deriving from the questionnaire.*

There were five dependent variables based on the questionnaire: the difficulty score for each of the five phases. The participants received a list of all questions and to indicate whether each was difficult (1) or not (0). The difficulty score of each phase was calculated per participant by averaging the dichotomous scores for each question in that phase. This compensated for the fact that some phases contained more questions than others. The resulting score indicated the proportion of questions in that category that were rated as difficult.

*Procedure.*

The experiment took place airside in the departure hall of a large international airport. One of the experimenters approached passengers after they had come through security and said that he was looking for volunteers to participate in an experiment that involved being interviewed about their travel plans and that they could earn £10 by taking part. He told them that the experiment would take 20 minutes and that the interviews would be video recorded. An estimated 40% of the participants agreed to take part. Passengers who did not take part gave good reasons for not doing so: They were busy (had work or shopping to do), travelled with more people or had to catch their flight. The reasonably high response rate may have been a combination of an official looking experimenter combined with the opportunity to earn money whilst serving a good cause (scientific research that could be used to
enhance security at airports). The fact that most people in a departure hall have little else to do besides waiting for their flight might also have helped.

The participants were first asked the following two questions:

- (1) Where are you going to fly to today?
- (2) How would you describe the main purpose of your trip?

These questions were asked to establish the ground truth in the experiment. All participants (truth tellers and liars) were asked to reveal their true destination during the actual interview, and truth tellers were also asked to reveal the true purpose of their trip, while the liars were asked to lie about the purpose of their trip. These data show that, in their interview, all interviewees honestly reported their destination, and that all truth tellers honestly reported the true purpose of their trip. All liars truthfully reported their destination but lied about the purpose.

The experimenter continued: “My colleague will ask you a few questions about your forthcoming trip. Some people will be asked to tell the truth whereas others will be asked to lie during these interviews. My colleague, who does not know who is lying or telling the truth, will make a veracity judgement at the end of the interview.” The participants were allocated randomly to the truth/lie condition. Half of the passengers were requested to tell the truth while answering every single question during the interview, whereas the other half were asked to tell the truth about the destination they were flying to but lie about the purpose of the trip. The experimenter then asked the participants whether they needed preparation time. If the participant expressed a need to prepare (only two passengers, both liars, did), the experimenter gave them as much time as they wanted by asking them to return when they were ready to be interviewed, at which point the experimenter took the
participant to one of the two female interviewers. The interviewer introduced herself, invited the participant to take a seat, and conducted the interview.

The interview was recorded with a video camera and a thermal imaging camera; the distance between the cameras and participant was approximately three metres. After the interview, the participant completed a questionnaire. The questionnaire asked for demographic information and, for each question, whether the participant found the question easy or hard to answer. After that they received £10.

**Questionnaire results.**

A 2 (Veracity) x 5 (Phase) Repeated Measures ANOVA was conducted. The results showed a significant Mauchly’s test of Sphericity, therefore the Greenhouse-Geisser statistics are reported here. The Veracity*Phase interaction effect was significant (F (2.11, 943.3) = 8.57, p < 0.001, partial \( \eta^2 \) = 0.02). The main effects of Veracity and Phase were significant (Veracity: F (1, 447) = 9.62, p = 0.002, partial \( \eta^2 \) = 0.02; Phase: F (2.11, 943.3) = 138.45, p < 0.001, partial \( \eta^2 \) = 0.23). See Table 3.1 for the means and standard deviations.

The Veracity*Phase interaction effect is more informative than the Veracity and Phase main effects. The Veracity*Phase interaction effect is therefore the only significant effect discussed here. Liars found the overall interview, phases one, two and four significantly harder than truth tellers (overall: T (463.13) = -3.01, p = 0.002; phase 1: T (388.03) = -4.9, p < 0.001; phase 2: T (471) = -4.1, p < 0.001; phase 4: T (472) = -4.54, p < 0.001). There is no significant difference between liars and truth tellers in phases three and five (phase 3: T (469.84) = -1.06, p = 0.29; phase 5: T (473) = 1.63, p =1.03). Phase five is considered the hardest phase, followed by phase
four, phase two, phase three and phase one. All five phases differ significantly from all others (All T > 3.18, all \( p < 0.002 \)).

Table 3.1: The mean (SD) of difficulty per phase. Higher scores indicate more participants rated the questions in this category as difficult.

<table>
<thead>
<tr>
<th></th>
<th>General</th>
<th>Destination plans and past</th>
<th>Plans on plane</th>
<th>Reassurance</th>
<th>Lie</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truth tellers</td>
<td>.05 (.18)</td>
<td>.20 (.23)</td>
<td>.05 (.18)</td>
<td>.16 (.21)</td>
<td>.48 (.50)</td>
<td>.19 (.16)</td>
</tr>
<tr>
<td>Liars</td>
<td>.16 (.29)</td>
<td>.29 (.23)</td>
<td>.07 (.19)</td>
<td>.26 (.24)</td>
<td>.41 (.49)</td>
<td>.24 (.17)</td>
</tr>
</tbody>
</table>

**Questionnaire discussion.**

The findings show that the hypothesis was partially supported. Liars found, overall, the interview more difficult than truth tellers. This is an indication that lying may be more cognitively demanding than telling the truth. However, the Phase results show that this is dependent on the kind of question that is asked. Liars found phases one, two and four harder than truth tellers, as hypothesised, but there was no significant difference between truth tellers and liars in phase three and five, although the trends were in the hypothesised direction. Perhaps, if questions have a rather obvious answer, like those in the third phase (there is only so much one can plan to do on a plane), liars experience them as being as easy as truth tellers do. The “lie to me” question was found to be the hardest overall, and more truth tellers than liars found it hard. This may perhaps be because truth tellers had to change from truth telling to lying, which is associated with more cognitive load (Monsell, 2003). In contrast, liars had to do the same task again, although they had to do a new lie, they still lied.
Chapter four: Non-verbal Behaviour as a Lie Detection Tool

In Chapter 4 non-verbal cues to deception are introduced and the parts of Experiment 1 that concern non-verbal cues are reported. Non-verbal cues to deception have been studied extensively. For example, in their 2003 meta-analysis DePaulo, Lindsay, Malone, Muhlenbruck, Charlton and Cooper analysed 116 different articles, which together gave 1338 estimates of 158 different verbal and non-verbal cues. Non-verbal cues can be divided into visual and vocal cues (Vrij, 2008). Visual cues are in the body language of the participant: the way they move, the gestures they make, the body movements they make, their eye movements. Vocal cues are cues in the voice of the participant; the focus is on how they speak, such as speech rate and fluency of speech.

The present experiment focused on an array of visual and vocal cues, most have been used previously. The visual cues investigated were: hand movements (hand and finger movements, not associated with speech or arm movements), illustrators (hand movements that help illustrate speech), self-manipulations (touching parts of the body except the hands, the latter are classified as hand movements), leg and foot movements, blinking (how often the participant blinks), gaze aversion (the amount of time spent looking away from the interviewer) and laughing. The vocal cues were: latency (the amount of time before the answer), pausing (breaking speech by silence), uhmming (breaking speech with a meaningless vocalisation), stuttering (repeating words or parts of words), struggling (struggling to find a word or phrase) and speech rate (how fast the speech is). Three cues that concern the overall impression the interviewee gives were also investigated: whether the interviewee appears to be nervous, appears to be thinking hard and appears to be attempting to control his or her non-verbal behaviour. These cues are seen as
outward signs of the interviewee’s arousal, cognitive load or attempt at controlling their behaviour respectively. Because these cues concern the impression the participant gives they may be more subjective than the others. How someone is perceived depends not only on their behaviour but also on the perceiver. High scores for hand movements, self-manipulations, leg and foot movements, stuttering, speech rate and appearing nervous are usually seen as signs of nervousness and arousal (Vrij, 2008). Low scores on hand movements, self-manipulation, leg and foot movements, speech rate and controlling behaviour and high scores on gaze aversion, latency, pausing, uhmming, struggling and thinking hard are usually seen as signs of high cognitive load (Vrij et al. 2006). High scores on appearing to control behaviour and low scores on hand movements, illustrators and leg and foot movements and gaze aversion are usually seen as signs that the participant is trying to control his behaviour. This is because most people believe that liars fidget and look away, thus liars will try not to fidget and look away in order to appear convincing (Akehurst, Köhnken, Vrij & Bull 1996; Hartwig et al. 2007).

Some cues have been found to be somewhat diagnostic: liars make fewer hand movements and illustrators than truth tellers (De Paulo et al., 2003). For other cues the findings have been mixed. For example, liars made more self manipulations than truth tellers in Porter, Doucette, Woodworth, Earle, and MacNeill (2008), whereas liars made fewer self manipulations than truth tellers in Granhag and Strömwall, (2002). Leg and foot movements also show mixed results. For example, Vrij and colleagues (2008) found that liars make more leg and foot movements than truth tellers, while Levine Feeley, McCormack, Hughes, and Harms (2005) found that truth tellers make more leg and foot movements.
DePaulo et al.’s (2003) meta-analysis found no significant difference between liars and truth tellers in blinking, whereas individual studies have revealed mixed results. Leal and Vrij (2008) found a possible explanation for this discrepancy in the way that it is measured. They found that truth tellers and liars have different blinking patterns. They conducted a mock crime experiment in which all participants told the truth on baseline questions, whereas on the target questions half lied and half told the truth. They examined the blinks during the baseline and target questions and also directly after the target questions. They found that liars blinked less when answering target questions than when answering baseline questions but that directly after having answered those target questions liars blinked more than when answering baseline or target questions. Truth tellers blinked more when answering target questions than when answering baseline questions and directly after answering target questions they blinked at the same rate as when answering target questions. Leal and Vrij (2008) found no difference between truth tellers and liars in blink rate when they examined the blinks during answering the target questions and directly after the target questions combined. So, although liars and truth tellers showed different blinking patterns, the overall rate in blinking did not differ.

The results of previous studies on gaze aversion are inconclusive. DePaulo et al. (2003) found no significant difference between liars and truth tellers in gaze aversion. Some studies have found that truth tellers look at the interviewer more than liars (Miller, de Turck & Kalbfleish, 1983), whereas others have found that liars look at the interviewer more than truth tellers (e.g. Granhag & Strömwall, 2002). Laughing shows a mixed pattern again. Vrij, Edward and Bull (2001) found that truth tellers laugh more than liars, while others found no difference (e.g. Hocking & Leathers, 1980). Ekman, Friesen and O’Sullivan (1988) suggest that the discrepancy
between studies may be because there are two kinds of smiles: genuine smiles and
smiles to mask other emotions. In Ekman et al.s (1988) experiment, truth tellers
showed more genuine smiles than liars, whereas liars showed more masking smiles
than truth tellers. When they observed smiles overall there was no difference
between liars and truth tellers. Latency, the time between the end of the question and
the start of the answer, has been investigated in several studies. Some studies found
longer latencies for truth tellers (Parliament & Yarmey, 2007), other studies found
longer latencies for liars (Vendemia, Buzan and Green, 2005). Pausing did not
emerge as a diagnostic cue to deceit in DePaulo et al. (2003) and DePaulo et al.
(2003) where furthermore no difference was found in uhmming between truth tellers
and liars either. However, some studies have found that truth tellers used more uhms
than liars (e.g. Vrij et al. 2001b, 2007, 2008), whereas others found that truth tellers
used uhm more often (e.g. Arciul, Mallard & Villar 2010). Arciuli et al. (2010)
suggest that the prediction of whether truth tellers uhm more or less than liars
depends on how uhmming is examined. If uhmming is a filled pause then liars would
be expected to uhm more, whereas liars would be predicted to uhm less if “uhm” is
considered as a word, similar to other interjections.

DePaulo et al. (2003) found in their meta-analysis that stuttering occurs more
during lies than truths. For differences in struggling with speaking, DePaulo et al.
(2003) found no effect. For speech rate, there are conflicting results again. For
example Vrij et al. (2001b) found that truth tellers speak more slowly than liars,
while Vrij et al. (2008) found that liars speak slower. Regarding the overall
impression cues, DePaulo et al. (2003) found that liars appear more nervous than
truth tellers. The cognitive load approach predicts that liars should be thinking harder
than truth tellers and this may be visible to an observer. This visibility is not based
on a particular cue, but rather on the overall impression the observer has of the participant (Vrij et al., 2008). This Hartwig, Granhag and Strömwall (2007) found that liars are more likely to attempt to control their non-verbal behaviour. This attempted control may be visible to an observer.

The participants in our sample had many different ethnicities. Previous research has shown that non-verbal behaviour can differ between ethnicities, so it is important to test for any interaction effects between veracity and ethnicity on non-verbal behaviour. Some of the visual cues in this experiment have previously been investigated in people with different ethnic backgrounds. Johnson (2007) studied the non-verbal behaviour of guilty and innocent Americans of different ethnicities. He found that Caucasian Americans and Hispanic Americans who had committed a crime made more hand gestures than innocent Caucasians and Hispanics in interviews, while African Americans who had committed a crime made less hand gestures than those who had not. He also found that African Americans and Hispanics, both guilty and innocent made more hand gestures than guilty or innocent Caucasians. This means that Hispanics and African Americans have a higher risk of being falsely accused, on the basis of their hand movements, since observers typically associate more hand movements with deception (Akehurst, Köhnken, Vrij & Bull, 1996; Lakhani & Taylor, 2003). Johnson (2007) also found that guilty and innocent Caucasians made more eye contact with the interviewer than guilty or innocent Hispanics or African Americans whilst smiling was not affected by ethnicity. Blinking is unlikely to be influenced by ethnicity since studies have shown that features like eye colour, which are correlated with ethnicity, do not affect the blink rate (Bentivoglio et al. 1997).
The effect of ethnicity on the vocal cues is difficult to predict, because the level of English a participant speaks will be a confounding factor. Speaking in a second language has been shown to affect cues to deceit (Caldwell-Harris & Ayçiçeği-Dinn, 2009) and participants speaking in a second language are more likely to be judged as deceptive by observers (Da Silva & Leach, 2011). Presumably this is because speaking in a second language increases cognitive load. Since signs of cognitive load can be cues to deception, both truth tellers and liars speaking in a second language may show more cues to deceit than truth tellers and liars speaking in their native language. Participants of certain ethnicities are more likely to be native English speakers than people with other ethnicities. For example, the Northern European ethnicity includes native English speakers from the UK, while the Arabian ethnicity is less likely to contain native English speakers. By scoring each participant on how well they speak English we hope to be able to untangle the effects of veracity, ethnicity and level of English. The previous research on the effects of veracity on these vocal cues is quite inconclusive suggesting that if there is an effect, it will be small. The effects of ethnicity and level of English on vocal cues may be larger.

In this experiment, we investigated the presence of the 17 non-verbal cues described above in truth telling and lying participants. For most of these cues there are no specific hypotheses since the previous research is not conclusive. However some hypotheses can be formulated (I) truth tellers will make more hand movements and illustrators than liars and there will be an interaction effect between veracity and ethnicity on hand movements and illustrators. Lying Caucasians will make more hand movements and illustrators than Caucasians who are telling the truth, while for Africans this will be reversed (DePaulo et al., 2003; Johnson, 2007); (II) Ethnicity
will affect gaze aversion, with North-Europeans averting their gaze less (Johnson, 2007); (III) liars will appear more nervous, hard thinking and controlling independent of their ethnicity (Vrij et al., 2008) and (IV) ethnicity and level of English will affect vocal cues more than veracity will (Caldwell-Harris & Ayçiçeği-Dinn, 2009).

**Method.**

For the Participants, Design and Procedure sections of the method, please see Chapter three.

**Dependent variables.**

There are 17 dependent variables, the 16 nonverbal cues discussed above and interview length. The first variable is *hand movements*, the amount of hand movements that could not be classified as illustrators or self-manipulations. The variable *illustrators* indicates the amount of hand and arm movements the participant uses to modify and supplement their speech. The variable *self-manipulations* indicates the amount of time the participant touches a part of their body, clothing or hair, with their hands. Touching the hands was classified as hand movements rather than self-manipulations. The variable *leg/foot movements* indicates the amount of time moving the legs and feet. The variable *blinking* is the frequency with which the participant blinks. The variable *gaze aversion* indicates the amount of time the participant looks anywhere except at the interviewer. The variable *laughing* indicates the amount of time the participant laughs. The variable *latency* indicates how long it takes the participant to start answering each question. *Pausing* indicates how much the participant spent pausing in their answering. *Uhmming* indicates how often the participant makes meaningless sounds, like, uhm ah or oh. *Stuttering* indicates how much the participant stutters, i.e. repeats a word or a part of a word. *Struggling* indicates to what extent the participant seems to be struggling to speak, to find the
right word or expression. Speech rate indicates how fast the participant speaks. The variable nervous is an indication of how nervous the participant seems to be. Hard thinking indicates how hard the participant appears to be thinking about his or her answer and the variable control indicates how much the participant seems to be controlling his or her non-verbal behaviour, either by moving or by keeping still deliberately. Besides these variables the interview length was measured in seconds.

**Coding and reliability coding.**

The variable length was measured using the timer of the videos. The English ability variable was rated on a three point scale (poor, average, fluent) by the interviewer who conducted the interview. They based their rating on how well the interviewee understood their questions and how well they answered. The other 16 dependent variables were coded by two separate coders on a 1-7 scale with higher scores indicating the behaviour occurred more. The coder was experienced with coding work and had been trained in using this coding system. The coder first watched some videos in order to get familiar with how a typical behavioural cue looked or sounded and how often the average person showed each cue. An independent secondary coder coded 50 of the participants. Both coders were Northern European, female psychologists and blind to the veracity status of the participants. The second coder learned the definition of each cue from the first coder. The first and the second coder watched several clips together, while the first coder stated how much each cue was present, so the second coder could learn to recognise them and see how often the participants showed them. The second coder then rated a few clips by herself, these ratings were compared with the ratings from the first coder and, where necessary, the second coder received further instructions on how to code the clips. Then the second coder rated the other clips. The correlation between
the first and the second coder are shown in Table 4.1. A score of .50 or higher is usually considered sufficiently reliable, while over .60 is considered good and over .75 excellent (Fleiss, 1981 (as cited in Vrij, 2005). Correlations were preferred, over kappa, because kappa can be affected by the base rate of a cue (Spitznagel & Helzer, 1985). Since some cues may be quite rare, this may affect kappa in this experiment. Table 4.1 shows that, although all cues are correlated significantly (all $p < 0.01$), only illustrators are in the excellent range with a correlation of .82. Self-manipulation, leg and foot movements, laughing, pausing, uhmning, and stuttering have good correlations, while hand movements and gaze aversion are still reliable. Struggling, speech rate, nervousness, hard thinking and control are all under .50. In particular, nervousness is low with a correlation of .20. Since nervousness, thinking hard and control concern the overall impression the participant gives rather than a countable behaviour, it is perhaps not surprising that the correlation is lower for these variables. It does mean that the results for these variables should be viewed with great caution, but in order to be complete its results are reported further on in the thesis. The correlations for struggling and speech rate are also low, but higher than the impression variables. Speech rate can be measured objectively, by counting the number of words or syllables per minute, the fact that the two subjective coders did not correlate highly suggests that they were not very good at estimating how fast someone spoke. Perhaps certain accents give the impression that someone is speaking quickly or slowly and this impression may be different for different observers. The low score for struggle perhaps reflects that each of the two coders tried to compensate for how well a person spoke English in their own way.
Table 4.1: The correlation between the first and second coder $N = 480$.

<table>
<thead>
<tr>
<th>Cue</th>
<th>Hand</th>
<th>Illustrator</th>
<th>Self-manipulation</th>
<th>Leg/foot</th>
<th>Blink</th>
<th>Gaze aversion</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>.54**</td>
<td>.82**</td>
<td>.66**</td>
<td>.63**</td>
<td>.64**</td>
<td>.50**</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cue</th>
<th>Laughing</th>
<th>Latency</th>
<th>Pausing</th>
<th>Uhmming</th>
<th>Stuttering</th>
<th>Struggling</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>.69**</td>
<td>.41**</td>
<td>.71**</td>
<td>.64**</td>
<td>.72**</td>
<td>.42**</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cue</th>
<th>Speech rate</th>
<th>Nervous</th>
<th>Hard think</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>.47**</td>
<td>.20**</td>
<td>.36**</td>
<td>.35**</td>
</tr>
</tbody>
</table>

** indicates $p < 0.01$

Data analysis.

The 17 variables were analysed using a multivariate repeated measures analysis with the 5 phases as the repeating independent factor and Veracity, Ethnicity and level of English as the between subject factors. Post-hoc tests (multiple comparisons) with Bonferroni adjustments were also used. All variables showed a multivariate significant Mauchly’s tests of sphericity. In order to mitigate this, the Pillai’s Trace statistic is reported, for multivariate test, because this statistic is the most robust against abnormal distributions in cases where all groups are of equal size. Because the data is on a seven point scale it was decided to deal with the abnormality statistically rather than by excluding outliers. The coders based their scores on their experience of the entire range of behaviours and the scale, so that the most extreme behaviour occupies the top and bottom and of the scale, regardless of whether the behaviour is actually extreme. Removing the outliers would simply cut the bottom and top end of the scale off and this would result in loss of information. For the univariate tests the sphericity assumed effects were reported, if Mauchly’s test of sphericity was non-significant. Where Mauchly’s test was significant, the
Greenhouse-Geisser effects are reported, because these are most conservative. In order to see whether the 16 cues could be condensed into fewer factors, the scores on all questions were averaged together so that each participant has an overall score for each variable. A principle component factor analysis was then performed on these summed variables. A discriminant analysis was conducted to test the usefulness of these factors for lie detection.

Results.

Multivariate.

The multivariate repeated measure analysis revealed that all main effects were significant: Veracity: $F(17, 305) = 1.79, p = 0.03, partial \eta^2 = .09$; Ethnicity: $F(187, 3465) = 1.59, p < 0.001, partial \eta^2 = .08$; level of English: $F(34, 612) = 4.37, p < 0.001, partial \eta^2 = .2$; Phase: $F(68, 254) = 30.98, p < 0.001, partial \eta^2 = .89$.

Three interaction effects were significant: the three way Ethnicity*English*Phase effect ($F(1496, 6050) = 1.12, p = 0.003, partial \eta^2 = .22$) and two two-way interaction effects: the Ethnicity*Phase ($F(748, 2904) = 1.31, p < 0.001, partial \eta^2 = .25$) and English*Phase effect ($F(136, 510) = 1.38, p = 0.007, partial \eta^2 = .27$). The other interaction effects were all non significant (all $F < 1.10, all p > .12$). The univariate results for the significant main effects are described in the next paragraph. All interaction effect are either not significant or do not include veracity, the main interest of this thesis. Therefore they will not be discussed in further detail.

Univariate.

The univariate Veracity tests showed significant effects for struggling ($F(1, 321) = 11.10, p = 0.001, partial \eta^2 = .03$), hard thinking ($F(1, 321) = 6.52, p = 0.01$, ...
partial $\eta^2 = .02)$ and control ($F(1, 321) = 9.36, p = 0.002, partial \eta^2 = .03$), all others were not significant ($F < 2.69, p > .10, partial \eta^2 < 0.1$). The estimated marginal means show that truth tellers ($M = 2.19, SE = 0.08$) struggled more than liars ($M = 1.95, SE = 0.08$); that truth tellers ($M = 2.87, SE = 0.06$) appeared to be thinking harder than liars ($M = 2.85, SE = 0.07$) and that liars ($M = 2.55, SE = 0.07$) appeared more controlled than truth tellers ($M = 2.22, SE = 0.06$). However as reported earlier, the reliability scores for these three variables were quite low. The consequences of this are mentioned in the discussion.

The univariate Ethnicity tests showed significant effects for length ($F(11, 321) = 2.03, p = 0.03, partial \eta^2 = .07$), laughing ($F(11, 321) = 3.75, p < 0.001, partial \eta^2 = .11$), latency ($F(11, 321) = 1.93, p = 0.04, partial \eta^2 = .06$), pausing ($F(11, 321) = 2.16, p = 0.02, partial \eta^2 = .07$), stuttering ($F(11, 321) = 1.88, p = 0.04, partial \eta^2 = .06$), struggling ($F(11, 321) = 2.35, p = 0.01, partial \eta^2 = .07$), nervousness ($F(11, 321) = 1.9, p = 0.04, partial \eta^2 = .06$) and hard thinking ($F(11, 321) = 3.9, p < 0.001, partial \eta^2 = .12$), all others are not significant ($F < 1.64, p > .09, partial \eta^2 < 0.6$). The post hoc tests show (see Table 4.2 for means, standard errors and indications of significant differences) that there are significant differences between the different ethnicities. The ranking of the ethnicities is relatively stable over the different variables. Northern Europeans, Eastern Europeans and East Asians score significantly higher than other ethnicities on most variables, while South East Europeans and South East Asians score significantly lower than other ethnicities on most variables. South West Europeans, Central Asians, Iranians and South Americans tend to be between these two groups, while Arabs and West and East Africans show a less clear pattern scoring high on some variables and low on others.
In order to get a clearer view of the difference a factor analysis is conducted (see page 53).
Table 4.2: The means (and standard errors) for Ethnicity on those variables where Ethnicity had a significant effect.

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>Length in sec.</th>
<th>Laugh</th>
<th>Latency</th>
<th>Pause</th>
<th>Uhm</th>
<th>Stutter</th>
<th>Struggle</th>
<th>Nervous</th>
<th>Hard Think</th>
</tr>
</thead>
<tbody>
<tr>
<td>North Europe</td>
<td>99.19&lt;sup&gt;ab&lt;/sup&gt; (7.50)</td>
<td>2.39&lt;sup&gt;ab&lt;/sup&gt; (0.20)</td>
<td>4.10&lt;sup&gt;ab&lt;/sup&gt; (0.18)</td>
<td>3.20&lt;sup&gt;ab&lt;/sup&gt; (0.18)</td>
<td>3.90&lt;sup&gt;a&lt;/sup&gt; (0.22)</td>
<td>2.83&lt;sup&gt;a&lt;/sup&gt; (0.23)</td>
<td>2.54&lt;sup&gt;ab&lt;/sup&gt; (0.19)</td>
<td>3.04&lt;sup&gt;a&lt;/sup&gt; (0.16)</td>
<td>3.34&lt;sup&gt;ab&lt;/sup&gt; (0.16)</td>
</tr>
<tr>
<td>East Europe</td>
<td>108.41&lt;sup&gt;abc&lt;/sup&gt; (5.59)</td>
<td>2.18&lt;sup&gt;abc&lt;/sup&gt; (0.15)</td>
<td>4.33&lt;sup&gt;a&lt;/sup&gt; (0.13)</td>
<td>3.31&lt;sup&gt;b&lt;/sup&gt; (0.14)</td>
<td>3.80&lt;sup&gt;ab&lt;/sup&gt; (0.16)</td>
<td>2.59&lt;sup&gt;ab&lt;/sup&gt; (0.18)</td>
<td>2.37&lt;sup&gt;a&lt;/sup&gt; (0.14)</td>
<td>2.33&lt;sup&gt;bc&lt;/sup&gt; (0.12)</td>
<td>3.10&lt;sup&gt;a&lt;/sup&gt; (0.12)</td>
</tr>
<tr>
<td>South West Europe</td>
<td>96.84&lt;sup&gt;ab&lt;/sup&gt; (8.83)</td>
<td>1.44&lt;sup&gt;c&lt;/sup&gt; (0.23)</td>
<td>4.07&lt;sup&gt;ab&lt;/sup&gt; (0.21)</td>
<td>2.57&lt;sup&gt;c&lt;/sup&gt; (0.21)</td>
<td>3.57&lt;sup&gt;ab&lt;/sup&gt; (0.26)</td>
<td>2.16&lt;sup&gt;a&lt;/sup&gt; (0.26)</td>
<td>2.05&lt;sup&gt;abcd&lt;/sup&gt; (0.23)</td>
<td>2.24&lt;sup&gt;bc&lt;/sup&gt; (0.19)</td>
<td>3.03&lt;sup&gt;abc&lt;/sup&gt; (0.19)</td>
</tr>
<tr>
<td>South East Europe</td>
<td>85.76&lt;sup&gt;b&lt;/sup&gt; (8.39)</td>
<td>1.98&lt;sup&gt;c&lt;/sup&gt; (0.22)</td>
<td>4.10&lt;sup&gt;ab&lt;/sup&gt; (0.20)</td>
<td>2.63&lt;sup&gt;bc&lt;/sup&gt; (0.20)</td>
<td>3.84&lt;sup&gt;ab&lt;/sup&gt; (0.25)</td>
<td>2.18&lt;sup&gt;b&lt;/sup&gt; (0.25)</td>
<td>2.00&lt;sup&gt;d&lt;/sup&gt; (0.22)</td>
<td>2.32&lt;sup&gt;bc&lt;/sup&gt; (0.18)</td>
<td>2.63&lt;sup&gt;bc&lt;/sup&gt; (0.18)</td>
</tr>
<tr>
<td>Central Asia</td>
<td>91.57&lt;sup&gt;ab&lt;/sup&gt; (9.10)</td>
<td>2.30&lt;sup&gt;abc&lt;/sup&gt; (0.24)</td>
<td>3.55&lt;sup&gt;b&lt;/sup&gt; (0.22)</td>
<td>2.90&lt;sup&gt;abc&lt;/sup&gt; (0.22)</td>
<td>2.84&lt;sup&gt;b&lt;/sup&gt; (0.27)</td>
<td>2.32&lt;sup&gt;ab&lt;/sup&gt; (0.27)</td>
<td>1.88&lt;sup&gt;abcd&lt;/sup&gt; (0.23)</td>
<td>2.45&lt;sup&gt;bc&lt;/sup&gt; (0.19)</td>
<td>2.72&lt;sup&gt;abc&lt;/sup&gt; (0.2)</td>
</tr>
<tr>
<td>East Asia</td>
<td>118.72&lt;sup&gt;a&lt;/sup&gt; (6.06)</td>
<td>2.29&lt;sup&gt;ab&lt;/sup&gt; (0.16)</td>
<td>4.47&lt;sup&gt;a&lt;/sup&gt; (0.15)</td>
<td>3.35&lt;sup&gt;ab&lt;/sup&gt; (0.15)</td>
<td>3.95&lt;sup&gt;ab&lt;/sup&gt; (0.18)</td>
<td>2.86&lt;sup&gt;a&lt;/sup&gt; (0.18)</td>
<td>2.24&lt;sup&gt;abc&lt;/sup&gt; (0.16)</td>
<td>2.74&lt;sup&gt;ab&lt;/sup&gt; (0.13)</td>
<td>3.20&lt;sup&gt;a&lt;/sup&gt; (0.13)</td>
</tr>
<tr>
<td>South East Asia</td>
<td>87.01&lt;sup&gt;b&lt;/sup&gt; (5.84)</td>
<td>1.94&lt;sup&gt;c&lt;/sup&gt; (0.15)</td>
<td>4.40&lt;sup&gt;a&lt;/sup&gt; (0.14)</td>
<td>2.84&lt;sup&gt;abc&lt;/sup&gt; (0.14)</td>
<td>3.47&lt;sup&gt;ab&lt;/sup&gt; (0.17)</td>
<td>1.97&lt;sup&gt;b&lt;/sup&gt; (0.18)</td>
<td>1.70&lt;sup&gt;d&lt;/sup&gt; (0.15)</td>
<td>2.25&lt;sup&gt;bc&lt;/sup&gt; (0.13)</td>
<td>2.50&lt;sup&gt;e&lt;/sup&gt; (0.13)</td>
</tr>
<tr>
<td>Iran</td>
<td>102.90&lt;sup&gt;ab&lt;/sup&gt; (8.63)</td>
<td>2.18&lt;sup&gt;abc&lt;/sup&gt; (0.23)</td>
<td>3.85&lt;sup&gt;ab&lt;/sup&gt; (0.21)</td>
<td>2.82&lt;sup&gt;abc&lt;/sup&gt; (0.21)</td>
<td>3.35&lt;sup&gt;ab&lt;/sup&gt; (0.25)</td>
<td>2.81&lt;sup&gt;ab&lt;/sup&gt; (0.26)</td>
<td>2.20&lt;sup&gt;abcd&lt;/sup&gt; (0.22)</td>
<td>2.72&lt;sup&gt;abc&lt;/sup&gt; (0.18)</td>
<td>3.15&lt;sup&gt;ab&lt;/sup&gt; (0.19)</td>
</tr>
<tr>
<td>Arabia</td>
<td>84.95&lt;sup&gt;ab&lt;/sup&gt; (8.31)</td>
<td>3.16&lt;sup&gt;a&lt;/sup&gt; (0.23)</td>
<td>4.04&lt;sup&gt;ab&lt;/sup&gt; (0.21)</td>
<td>2.90&lt;sup&gt;abc&lt;/sup&gt; (0.21)</td>
<td>3.53&lt;sup&gt;ab&lt;/sup&gt; (0.25)</td>
<td>2.46&lt;sup&gt;ab&lt;/sup&gt; (0.26)</td>
<td>1.90&lt;sup&gt;abcd&lt;/sup&gt; (0.22)</td>
<td>2.51&lt;sup&gt;bc&lt;/sup&gt; (0.18)</td>
<td>2.52&lt;sup&gt;ab&lt;/sup&gt; (0.19)</td>
</tr>
<tr>
<td>South America</td>
<td>99.97&lt;sup&gt;ab&lt;/sup&gt; (6.68)</td>
<td>2.06&lt;sup&gt;bc&lt;/sup&gt; (0.16)</td>
<td>3.98&lt;sup&gt;ab&lt;/sup&gt; (0.16)</td>
<td>2.88&lt;sup&gt;abc&lt;/sup&gt; (0.16)</td>
<td>3.61&lt;sup&gt;ab&lt;/sup&gt; (0.20)</td>
<td>2.50&lt;sup&gt;a&lt;/sup&gt; (0.20)</td>
<td>2.36&lt;sup&gt;abcd&lt;/sup&gt; (0.17)</td>
<td>2.53&lt;sup&gt;bc&lt;/sup&gt; (0.14)</td>
<td>2.97&lt;sup&gt;abc&lt;/sup&gt; (0.15)</td>
</tr>
<tr>
<td>West Africa</td>
<td>102.86&lt;sup&gt;ab&lt;/sup&gt; (8.09)</td>
<td>1.89&lt;sup&gt;bc&lt;/sup&gt; (0.21)</td>
<td>4.46&lt;sup&gt;a&lt;/sup&gt; (0.19)</td>
<td>2.60&lt;sup&gt;c&lt;/sup&gt; (0.20)</td>
<td>3.64&lt;sup&gt;ab&lt;/sup&gt; (0.24)</td>
<td>2.63&lt;sup&gt;ab&lt;/sup&gt; (0.24)</td>
<td>1.77&lt;sup&gt;abcd&lt;/sup&gt; (0.21)</td>
<td>2.47&lt;sup&gt;bc&lt;/sup&gt; (0.17)</td>
<td>2.40&lt;sup&gt;e&lt;/sup&gt; (0.18)</td>
</tr>
<tr>
<td>East Africa</td>
<td>96.22&lt;sup&gt;ab&lt;/sup&gt; (7.59)</td>
<td>1.65&lt;sup&gt;c&lt;/sup&gt; (0.2)</td>
<td>4.21&lt;sup&gt;a&lt;/sup&gt; (0.18)</td>
<td>2.72&lt;sup&gt;bc&lt;/sup&gt; (0.18)</td>
<td>3.57&lt;sup&gt;ab&lt;/sup&gt; (0.22)</td>
<td>2.30&lt;sup&gt;b&lt;/sup&gt; (0.23)</td>
<td>1.67&lt;sup&gt;cd&lt;/sup&gt; (0.19)</td>
<td>2.36&lt;sup&gt;bc&lt;/sup&gt; (0.16)</td>
<td>2.63&lt;sup&gt;c&lt;/sup&gt; (0.17)</td>
</tr>
</tbody>
</table>

N.B. In columns, only cells that do not share any superscript letter differ significantly (at p < 0.05), Bonferroni corrections are used. Superscript letters earlier in the alphabet indicate higher scores.
The univariate tests for the level of English showed significant effects for length ($F(2, 321) = 7.19, p = 0.001, \text{partial } \eta^2 = .04$), illustrators ($F(2, 321) = 8.22, p < 0.001, \text{partial } \eta^2 = .05$), latency ($F(2, 321) = 13.32, p < 0.001, \text{partial } \eta^2 = .08$), pausing ($F(2, 321) = 8.73, p < 0.001, \text{partial } \eta^2 = .05$), uhmming ($F(2, 321) = 9.37, p < 0.001, \text{partial } \eta^2 = .06$), stuttering ($F(2, 321) = 8.27, p < 0.001, \text{partial } \eta^2 = .05$), struggling ($F(2, 321) = 46.53, p < 0.001, \text{partial } \eta^2 = .23$), speech rate ($F(2, 321) = 10.58, p < 0.001, \text{partial } \eta^2 = .06$), nervousness ($F(2, 321) = 4.08, p = 0.02, \text{partial } \eta^2 = .07$) and hard thinking ($F(2, 321) = 22.91, p =< 0.001, \text{partial } \eta^2 = .13$). The means, standard errors and indications of the post-hoc results are in Table 4.3. For all variables the results are as hypothesised. The participants who spoke English poorly obtained scores that indicate that they were less fluent than those who spoke English averagely and those who spoke English best appeared to be most fluent. Participants who spoke English poorly had longer interviews and illustrated more. These are presumably strategies to compensate for the lower level of English. Participants who spoke English poorly were also judged as thinking harder. Speaking a language you don’t know well is more difficult than speaking a language you know well. Although the correlation between the coders was low for this variable, these scores are in line with what could be expected.
Table 4.3: The means (and standard errors) for English on the variables where English had a significant effect.

<table>
<thead>
<tr>
<th>English</th>
<th>Length in sec</th>
<th>Illustrators</th>
<th>Latency</th>
<th>Pause</th>
<th>Uhm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poor</td>
<td>119.42 (6.15)^a</td>
<td>3.53 (0.21)^a</td>
<td>4.63 (0.15)^a</td>
<td>3.35 (0.15)^a</td>
<td>4.10 (0.18)^a</td>
</tr>
<tr>
<td>Average</td>
<td>95.65 (2.17)^b</td>
<td>3.02 (0.07)^b</td>
<td>4.17 (0.05)^b</td>
<td>2.92 (0.05)^b</td>
<td>3.64 (0.06)^b</td>
</tr>
<tr>
<td>Fluent</td>
<td>87.08 (3.57)^b</td>
<td>2.55 (0.12)^c</td>
<td>3.77 (0.09)^c</td>
<td>2.60 (0.09)^c</td>
<td>3.23 (0.11)^c</td>
</tr>
</tbody>
</table>

Speech

<table>
<thead>
<tr>
<th>Stutter</th>
<th>Struggle</th>
<th>rate</th>
<th>Nervousness</th>
<th>Hard think</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poor</td>
<td>2.79 (0.18)^a</td>
<td>3.35 (0.16)^a</td>
<td>3.19 (0.16)^c</td>
<td>2.62 (0.13)^a</td>
</tr>
<tr>
<td>Average</td>
<td>2.60 (0.07)^a</td>
<td>1.88 (0.06)^b</td>
<td>3.94 (0.06)^b</td>
<td>2.57 (0.05)^a</td>
</tr>
<tr>
<td>Fluent</td>
<td>2.13 (0.11)^b</td>
<td>1.40 (0.09)^c</td>
<td>4.14 (0.09)^a</td>
<td>2.33 (0.08)^a</td>
</tr>
</tbody>
</table>

N.B. In columns, only those that do not share any superscript letter differ significantly. Superscript letters earlier in the alphabet indicate scores associated with the lowest level of English.

The univariate Phase results showed that all variables were affected by Phase (all $F > 5.36$, all $p < 0.001$). The means, standard errors and significance indicators are in Table 4.4. For all variables, except nervousness and control, the general questions elicited the lowest scores, followed by the on plane questions. Reassurance questions and the lie question tended to elicit even higher scores and the destination questions tended to elicit the highest scores. For nervousness and control general questions elicited the highest scores with scores decreasing as the interview went on.
Table 4.4: The means (and standard errors) for Phase.

<table>
<thead>
<tr>
<th>Phase</th>
<th>Length</th>
<th>Hand</th>
<th>Illus.</th>
<th>Self</th>
<th>Leg</th>
<th>Blink</th>
<th>Gaze</th>
<th>Laugh</th>
<th>Laten.</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>21.11&lt;sup&gt;c&lt;/sup&gt;</td>
<td>3.17&lt;sup&gt;b&lt;/sup&gt;</td>
<td>2.04&lt;sup&gt;d&lt;/sup&gt;</td>
<td>1.52&lt;sup&gt;c&lt;/sup&gt;</td>
<td>2.41&lt;sup&gt;b&lt;/sup&gt;</td>
<td>3.52&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2.90&lt;sup&gt;b&lt;/sup&gt;</td>
<td>1.35&lt;sup&gt;b&lt;/sup&gt;</td>
<td>3.60&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>(0.82)</td>
<td>(0.12)</td>
<td>(0.09)</td>
<td>(0.07)</td>
<td>(0.12)</td>
<td>(0.08)</td>
<td>(0.08)</td>
<td>(0.05)</td>
<td>(0.09)</td>
<td></td>
</tr>
<tr>
<td>Destination</td>
<td>201.91&lt;sup&gt;a&lt;/sup&gt;</td>
<td>3.64&lt;sup&gt;a&lt;/sup&gt;</td>
<td>3.64&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2.55&lt;sup&gt;a&lt;/sup&gt;</td>
<td>3.06&lt;sup&gt;a&lt;/sup&gt;</td>
<td>3.78&lt;sup&gt;a&lt;/sup&gt;</td>
<td>3.57&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2.39&lt;sup&gt;a&lt;/sup&gt;</td>
<td>4.31&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>(3.83)</td>
<td>(0.09)</td>
<td>(0.09)</td>
<td>(0.12)</td>
<td>(0.10)</td>
<td>(0.08)</td>
<td>(0.07)</td>
<td>(0.09)</td>
<td>(0.07)</td>
<td></td>
</tr>
<tr>
<td>On plane</td>
<td>48.90&lt;sup&gt;d&lt;/sup&gt;</td>
<td>3.54&lt;sup&gt;ab&lt;/sup&gt;</td>
<td>3.04&lt;sup&gt;bc&lt;/sup&gt;</td>
<td>1.94&lt;sup&gt;b&lt;/sup&gt;</td>
<td>2.81&lt;sup&gt;ab&lt;/sup&gt;</td>
<td>3.68&lt;sup&gt;a&lt;/sup&gt;</td>
<td>3.51&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2.19&lt;sup&gt;a&lt;/sup&gt;</td>
<td>4.21&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>(2.12)</td>
<td>(0.11)</td>
<td>(.10)</td>
<td>(0.09)</td>
<td>(0.12)</td>
<td>(0.08)</td>
<td>(0.08)</td>
<td>(0.09)</td>
<td>(0.08)</td>
<td></td>
</tr>
<tr>
<td>Reassurance</td>
<td>152.81&lt;sup&gt;b&lt;/sup&gt;</td>
<td>3.38&lt;sup&gt;ab&lt;/sup&gt;</td>
<td>3.34&lt;sup&gt;ab&lt;/sup&gt;</td>
<td>2.42&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2.85&lt;sup&gt;ab&lt;/sup&gt;</td>
<td>3.75&lt;sup&gt;a&lt;/sup&gt;</td>
<td>3.66&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2.37&lt;sup&gt;a&lt;/sup&gt;</td>
<td>4.41&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>(4.16)</td>
<td>(0.09)</td>
<td>(0.10)</td>
<td>(0.11)</td>
<td>(0.10)</td>
<td>(0.08)</td>
<td>(0.07)</td>
<td>(0.09)</td>
<td>(0.07)</td>
<td></td>
</tr>
<tr>
<td>Lie</td>
<td>67.16&lt;sup&gt;c&lt;/sup&gt;</td>
<td>3.22&lt;sup&gt;c&lt;/sup&gt;</td>
<td>2.80&lt;sup&gt;c&lt;/sup&gt;</td>
<td>1.54&lt;sup&gt;c&lt;/sup&gt;</td>
<td>2.65&lt;sup&gt;a&lt;/sup&gt;</td>
<td>3.57&lt;sup&gt;a&lt;/sup&gt;</td>
<td>3.78&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2.27&lt;sup&gt;a&lt;/sup&gt;</td>
<td>4.16&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>(20.04)</td>
<td>(0.12)</td>
<td>(0.11)</td>
<td>(0.10)</td>
<td>(0.12)</td>
<td>(0.08)</td>
<td>(.09)</td>
<td>(0.09)</td>
<td>(0.09)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phase</th>
<th>Pause</th>
<th>Uhm</th>
<th>Stut</th>
<th>Strug</th>
<th>SR</th>
<th>Nerv</th>
<th>Hard think</th>
<th>Cont.</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>1.97&lt;sup&gt;c&lt;/sup&gt;</td>
<td>2.81&lt;sup&gt;c&lt;/sup&gt;</td>
<td>1.56&lt;sup&gt;c&lt;/sup&gt;</td>
<td>1.57&lt;sup&gt;c&lt;/sup&gt;</td>
<td>3.97&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2.69&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2.45&lt;sup&gt;c&lt;/sup&gt;</td>
<td>2.61&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>(0.06)</td>
<td>(0.10)</td>
<td>(0.07)</td>
<td>(0.07)</td>
<td>(0.06)</td>
<td>(0.06)</td>
<td>(0.05)</td>
<td>(0.06)</td>
<td></td>
</tr>
<tr>
<td>Destination</td>
<td>2.85&lt;sup&gt;b&lt;/sup&gt;</td>
<td>4.11&lt;sup&gt;a&lt;/sup&gt;</td>
<td>3.25&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2.51&lt;sup&gt;a&lt;/sup&gt;</td>
<td>3.84&lt;sup&gt;ab&lt;/sup&gt;</td>
<td>2.53&lt;sup&gt;ab&lt;/sup&gt;</td>
<td>2.72&lt;sup&gt;b&lt;/sup&gt;</td>
<td>2.40&lt;sup&gt;ab&lt;/sup&gt;</td>
</tr>
<tr>
<td>(0.07)</td>
<td>(0.07)</td>
<td>(0.10)</td>
<td>(0.08)</td>
<td>(0.07)</td>
<td>(0.06)</td>
<td>(0.06)</td>
<td>(0.06)</td>
<td></td>
</tr>
<tr>
<td>On plane</td>
<td>2.63&lt;sup&gt;b&lt;/sup&gt;</td>
<td>3.62&lt;sup&gt;b&lt;/sup&gt;</td>
<td>2.46&lt;sup&gt;b&lt;/sup&gt;</td>
<td>2.08&lt;sup&gt;b&lt;/sup&gt;</td>
<td>3.92&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2.38&lt;sup&gt;b&lt;/sup&gt;</td>
<td>2.60&lt;sup&gt;bc&lt;/sup&gt;</td>
<td>2.35&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>(0.07)</td>
<td>(0.08)</td>
<td>(0.09)</td>
<td>(0.07)</td>
<td>(0.06)</td>
<td>(0.05)</td>
<td>(0.05)</td>
<td>(0.05)</td>
<td></td>
</tr>
<tr>
<td>Reassurance</td>
<td>3.62&lt;sup&gt;a&lt;/sup&gt;</td>
<td>3.83&lt;sup&gt;ab&lt;/sup&gt;</td>
<td>2.90&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2.30&lt;sup&gt;ab&lt;/sup&gt;</td>
<td>3.81&lt;sup&gt;ab&lt;/sup&gt;</td>
<td>2.45&lt;sup&gt;b&lt;/sup&gt;</td>
<td>3.26&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2.30&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>(0.09)</td>
<td>(0.08)</td>
<td>(0.09)</td>
<td>(0.07)</td>
<td>(0.07)</td>
<td>(0.06)</td>
<td>(0.07)</td>
<td>(0.05)</td>
<td></td>
</tr>
<tr>
<td>Lie</td>
<td>3.47&lt;sup&gt;a&lt;/sup&gt;</td>
<td>3.62&lt;sup&gt;b&lt;/sup&gt;</td>
<td>2.18&lt;sup&gt;b&lt;/sup&gt;</td>
<td>1.88&lt;sup&gt;b&lt;/sup&gt;</td>
<td>3.61&lt;sup&gt;b&lt;/sup&gt;</td>
<td>2.42&lt;sup&gt;b&lt;/sup&gt;</td>
<td>3.29&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2.28&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>(0.11)</td>
<td>(0.10)</td>
<td>(0.09)</td>
<td>(0.07)</td>
<td>(0.07)</td>
<td>(0.05)</td>
<td>(0.08)</td>
<td>(0.05)</td>
<td></td>
</tr>
</tbody>
</table>

*N.B. In columns, only cells that do not share any superscript letter differ significantly. Superscript letters earlier in the alphabet indicate higher scores.*
**Factor analysis.**

The principal component factor analysis extracted 6 components with eigenvalues over 1 (see Table 4.5). The first component has an Eigenvalue of 3.91 and could be named “cognitive load”. The speech variables latency, pausing, uhmming, stuttering, struggling and hard thinking loaded highly on it, speech rate loaded highly, but negatively. Gaze aversion also loaded highly, suggesting that maintaining eye contact created extra cognitive load. Participants who scored highly on this component also appeared to be having difficulty answering the questions (the low interreliability score for this variable does mean the results should be viewed with caution).

The second component had an Eigenvalue of 1.64 and could be named “animation”. Illustrators, laughing, stuttering and speech rate loaded highly on this factor. Participants who scored high on this component appeared animated. The last four components had low Eigenvalues (1.47, 1.09, 1.08, 1.03) and the loadings are difficult to interpret. They are therefore not analysed further.
Table 4.5: The six components with eigenvalues over 1 extracted from the factor analysis. Bold cells indicate those cells with loading 0.45 or higher.

<table>
<thead>
<tr>
<th>Component</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>hand</td>
<td>.157</td>
<td>.130</td>
<td>.464</td>
<td>.651</td>
<td>-.296</td>
<td>-.096</td>
</tr>
<tr>
<td>illustrator</td>
<td>.357</td>
<td>.469</td>
<td>-.520</td>
<td>-.090</td>
<td>-.106</td>
<td>-.069</td>
</tr>
<tr>
<td>selfman</td>
<td>.372</td>
<td>.200</td>
<td>-.069</td>
<td>-.346</td>
<td>.497</td>
<td>-.089</td>
</tr>
<tr>
<td>legfoot</td>
<td>.095</td>
<td>.315</td>
<td>.334</td>
<td>.030</td>
<td>.350</td>
<td>-.569</td>
</tr>
<tr>
<td>blink</td>
<td>.055</td>
<td>-.302</td>
<td>-.207</td>
<td>.404</td>
<td>.614</td>
<td>.267</td>
</tr>
<tr>
<td>gazeaver</td>
<td>.507</td>
<td>-.220</td>
<td>-.006</td>
<td>.190</td>
<td>.285</td>
<td>.304</td>
</tr>
<tr>
<td>laugh</td>
<td>.167</td>
<td>.570</td>
<td>.178</td>
<td>.059</td>
<td>.207</td>
<td>.003</td>
</tr>
<tr>
<td>latency</td>
<td>.701</td>
<td>-.255</td>
<td>-.111</td>
<td>.010</td>
<td>-.172</td>
<td>-.009</td>
</tr>
<tr>
<td>pause</td>
<td>.743</td>
<td>-.245</td>
<td>.236</td>
<td>-.153</td>
<td>.022</td>
<td>-.157</td>
</tr>
<tr>
<td>uhm</td>
<td>.675</td>
<td>-.014</td>
<td>-.261</td>
<td>.061</td>
<td>-.186</td>
<td>.188</td>
</tr>
<tr>
<td>stutter</td>
<td>.542</td>
<td>.449</td>
<td>-.157</td>
<td>-.009</td>
<td>-.025</td>
<td>.231</td>
</tr>
<tr>
<td>struggle</td>
<td>.784</td>
<td>.135</td>
<td>-.074</td>
<td>-.022</td>
<td>-.141</td>
<td>-.141</td>
</tr>
<tr>
<td>speechrate</td>
<td>-.468</td>
<td>.502</td>
<td>-.161</td>
<td>.010</td>
<td>-.100</td>
<td>.284</td>
</tr>
<tr>
<td>nervous</td>
<td>.337</td>
<td>.398</td>
<td>.510</td>
<td>.102</td>
<td>.064</td>
<td>.394</td>
</tr>
<tr>
<td>hardthink</td>
<td>.761</td>
<td>-.183</td>
<td>.096</td>
<td>-.028</td>
<td>-.083</td>
<td>-.075</td>
</tr>
<tr>
<td>control</td>
<td>-.013</td>
<td>-.128</td>
<td>.563</td>
<td>-.544</td>
<td>-.064</td>
<td>.376</td>
</tr>
</tbody>
</table>

A MANOVA was conducted with the participants scores on the first two factors, as created by the factor analysis, cognitive load (score range = -2.06 to 4.15) and animation (score range = -2.14 to 2.88) as dependent variables and Veracity, Ethnicity and English as independent variables. The multivariate tests showed no significant effect for Veracity ($F (2, 321) = 2.47, p = 0.09, partial \eta^2 = .02$).

Ethnicity and English significantly affected these scores (Ethnicity: $F (22, 644) = 1.81, p = 0.01, partial \eta^2 = .07$; English: $F (4, 644) = 16.95, p < 0.001, partial \eta^2 = .10$). None of the two-way effects: Veracity*Ethnicity, Veracity*English,
Ethnicity*English were significant (All $F$’s < 1.15, all $p > 0.29$). The three way Veracity*Ethnicity*English effect was significant ($F (28, 644) = 1.84, p = 0.01$, 
\[
\text{partial } \eta^2 = .07
\]. The univariate tests showed that Ethnicity significantly affected Animation ($F (11, 385) = 1.86, p = 0.04$, 
\[
\text{partial } \eta^2 = .06
\]), but not cognitive load ($F (2, 321) = 1.64, p = 0.09$, 
\[
\text{partial } \eta^2 = .05
\]). The means, standard deviations and post hoc results are in Table 4.6.

Table 4.6: The means (and SD) of Ethnicity for the animation factor.

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>North Europe</td>
<td>0.32 (1.06)$_{ab}$</td>
</tr>
<tr>
<td>East Europe</td>
<td>0.08 (1.14)$_{abc}$</td>
</tr>
<tr>
<td>South West Europe</td>
<td>0.10 (1.03)$_{abc}$</td>
</tr>
<tr>
<td>South East Europe</td>
<td>-0.52 (0.81)$_c$</td>
</tr>
<tr>
<td>Central Asia</td>
<td>0.13 (0.97)$_{abc}$</td>
</tr>
<tr>
<td>East Asia</td>
<td>0.45 (1.23)$_a$</td>
</tr>
<tr>
<td>South East Asia</td>
<td>-0.36 (0.87)$_{bc}$</td>
</tr>
<tr>
<td>Iran</td>
<td>0.03 (0.93)$_{abc}$</td>
</tr>
<tr>
<td>Arabia</td>
<td>0.25 (0.97)$_{abc}$</td>
</tr>
<tr>
<td>South America</td>
<td>0.01 (0.95)$_{abc}$</td>
</tr>
<tr>
<td>West Africa</td>
<td>0.01 (0.82)$_{abc}$</td>
</tr>
<tr>
<td>East Africa</td>
<td>-0.29 (0.78)$_{abc}$</td>
</tr>
</tbody>
</table>

\textit{N.B. In columns, only cells that do not share any superscript letter differ significantly. Superscript letters earlier in the alphabet indicate higher scores.}

Level of English, rated by interviewers, significantly affected both cognitive load and animation (cognitive load: $F (2, 385) = 24.48, p < 0.001$, 
\[
\text{partial } \eta^2 = .13
\];
animation: $F(2, 385) = 9.57, p < 0.001, \text{partial } \eta^2 = .06$). The descriptive statistics showed that poor English speakers score highest on the cognitive load factor (M = 1.37 SD = 1.2) followed by the average English speakers (M = -0.06, SD = 0.87). The fluent English speakers had the lowest cognitive load scores (M = -0.31, SD = 0.83). The post hoc tests showed that these all differ significantly from each other (all $p < 0.001$) The poor English speakers also scored highest on the animation factor (M = 0.46 SD = 1.37), followed by the average English speakers (M = 0.07, SD = 0.97). The fluent English speakers had the lowest animation scores (M = -0.27, SD = 0.86). The post hoc tests showed that fluent speakers differed significantly from both poor speakers and average speakers (fluent-poor: $p < 0.001$; fluent-average: $p = 0.004$). Poor speakers and average speakers did not differ significantly from each other ($p = 0.06$). The three way Veracity*Ethnicity*English effect is significant for both the cognitive load and the animation factor (cognitive load: $F(14, 385) = 1.90, p = 0.03, \text{partial } \eta^2 = .08$; animation: $F(14, 385) = 1.84, p = 0.03, \text{partial } \eta^2 = .07$).

Discriminant analyses with the first two components as predictors showed that 52.3% of truth tellers and 62.6% of liars were classified correctly based on these two components ($\lambda = .98, \chi^2 = 7.93, p = 0.02$), where 50% would be classified by chance. For 17.4% of the participants their ethnicity could be classified correctly ($\lambda = .84, \chi^2 = 66.83, p < 0.001; \lambda = .94, \chi^2 = 22.64, p = 0.01$), where 8.3% could be classified by chance. For 49.1% of participants their level of English could be accurately predicted ($\lambda = .75, \chi^2 = 109.02, p < 0.001; \lambda = .99, \chi^2 = 3.94, p = 0.05$), where 33% could be classified correctly by chance.
Discussion.

This experiment tested the effect of Phase, Veracity, Ethnicity and level of English on 17 cues to deception. All cues had been examined before, but for most of them inconclusive results had been found in previous studies.

Phase.

The phase of the interview affected all cues. This is presumably because different types of question elicited different answers for all participants. The ‘destination’ questions asked for a lot of information, most of which was personal and, since most people like their trips, positive. This elicited long, animated answers, with high scores on both visual and vocal cues. The ‘general’ and ‘on the plane’ questions resulted in shorter answers, with less signs of animation or cognitive load, probably because there are only a few answers possible to these questions: there are only so many reasons for a trip and only so many things possible to do on a plane. The ‘reassurance’ and ‘lie’ categories can be answered in many different ways, but the questionnaire results showed that most people found these questions difficult to answer. Perhaps questions with many different answer possibilities increased the amount of cues the participants gave, while answering difficult questions decreased the amount of cues the participant gave. The ‘reassurance’ and ‘lie to me’ questions elicited a medium amount of cues while answering them. For nervousness and control the scores went down as the interview went on. Presumably, the participants relaxed as the interview progressed. This may be because of the friendly attitude of the interviewers or, alternatively, because the participants could simply not sustain the high level of arousal and control for more than a few minutes.
Veracity.

The veracity of the interview affected only the variables ‘struggling’ and how much the participant appeared to be thinking hard and controlling their behaviour. These three variables all had low inter-rater reliability, which means these results should be interpreted with caution. Hypothesis (I) was not supported. It stated that both hand movements and illustrators would be affected by Veracity and Ethnicity, but neither hand movements nor illustrators were affected by Veracity or Ethnicity. This is contrary to the findings of DePaulo et al.’s (2003) meta-analysis, which found that veracity affected hand movements and illustrators. It is possible that the greater power of a meta-analysis detected differences that could not be found in this single experiment. It is also possible that most of the studies in the meta-analysis used a less diverse set of participants in terms of age and ethnicity than we used, which may have affected the results. Hypothesis (III) was partly supported and partly not supported. It stated that liars would appear more nervous, to be thinking harder and more like they were controlling their behaviour. The finding that truth tellers and liars appeared equally nervous contradicted earlier studies (DePaulo et al. 2003). It is possible that the coders could not recognise nervousness very well: the inter-coder reliability was particularly low for this variable. It is also possible that the assumption that liars are more nervous than truth tellers is not true. Both liars and truth tellers may have reason to be nervous. Liars may be nervous because they may be afraid of getting caught. Truth tellers may be nervous because they may be concerned that they would fail to convince the interviewer that they are telling the truth. In an airport situation one would have to take into account that people may be nervous even without being stopped. Not so much because they are doing or
planning something illegal, but because they may be scared of flying or in a hurry to catch a plane.

Truth tellers gave the impression that they were thinking harder than liars, which is in contrast to what was predicted. The finding that truth tellers were thought to have to think harder contradicted Vrij et al. (2006; 2008), although the difference in mean scores in this experiment was small. Liars did experience the interview as harder (see chapter 3), according to their self-reports, so we can assume that they did in fact have to think harder, even if the coders did not rate them as such. It is possible that liars tried to be convincing by answering smoothly and quickly and hereby avoided appearing to think too hard. As predicted, liars seemed to control their behaviour more than truth tellers. The finding that liars appeared to control their behaviour more supports Hartwig et al.’s (2007) theory. It suggests that the strategies liars use to appear convincing generates cues to deception.

*Ethnicity.*

Ethnicity was found to affect length of interview, laughing, latency, pausing, uhmming, stuttering, struggling, nervousness and thinking hard. Because gaze aversion was not affected by ethnicity Hypothesis (II), which stated that Northern Europeans would avert their gaze least of all ethnicities, was not supported. Because hand movements and illustrators were also not affected by ethnicity Hypothesis (I), which stated that hand movements and illustrators would be affected by ethnicity was also not supported. Perhaps the difference with the observations of Johnson (2007), who found that Caucasians differ in their hand movements and gaze from Hispanic-Americans and African-Americans, was caused by a different definition of ethnicity. In that study the participants were all living the USA but with different
ethnic background, while in this experiment the different ethnicities came from different countries and continents. It may be that US ethnicities are not a reflection of world-wide ethnicities. So that what the effects Johnson (2007) found were effects of local culture. It did not appear to be the case that participants that originated from geographically close areas resembled each other more than participants that originated from areas further apart.

*English.*

The level of English affected the variables: length of interview, illustrators, latency, pausing, uhmming, stuttering, struggling, speech rate, nervousness and thinking hard. Hypothesis (IV) stated that the vocal cues would be affected more by Ethnicity and English than by Veracity. It was partly supported. Veracity only influenced one of the vocal cues, struggling, and this had a partial $\eta^2$ of .03. Ethnicity significantly influenced four vocal cues: latency, pausing, stuttering and struggling and all of these cues had higher partial $\eta^2$ than Veracity had for struggling. English affected all vocal cues: latency, pausing, uhmming, stuttering, struggling and speech rate and all these cues had higher partial $\eta^2$ than veracity. The fact that Ethnicity and level of English affected the vocal cues more than Veracity did, indicates that these vocal cues are not very useful for detecting deception in a sample which includes non-native speakers. Interestingly, we found no significant Ethnicity by English interaction effect, so the effect of the level of English remains the same for all ethnicities. Another possibility is that the vocal cues a speaker displays are affected by their mother tongue. For example, people whose mother tongue has more syllables per minute, may also speak more syllables per minute in English.
Factor analysis.

The factor analysis indicated that the cues investigated were not completely independent from each other. Instead, they appeared to be different operationalisations of underlying concepts. On the first component: latency, pausing, uhmming, stuttering, struggling, hard thinking and (negatively) speech rate loaded highly. These cues are all signs of cognitive load. The fact that the impression to think hard also loaded highly on this component indicates that the cues associated in the literature with cognitive load are related to the observers’ judgement of cognitive load. This cognitive load factor was influenced only by the level of English, not by Veracity or Ethnicity. Although the liars reported that they found the interview more difficult than the truth tellers did, this is not reflected in this factor. Perhaps the level of English affects the signs of cognitive load so much that the observers could no longer see the smaller differences between truth tellers and liars. On the second component illustrators, laughing, stuttering and speech rate loaded highly. This component seems to capture an animation or perhaps a social component. Participants loading high on this factor may come across as more animated or friendly. This factor was influenced by the participants’ ethnicity and level of English. That certain ethnicities show different non-verbal behaviours has been found before (Matsumoto, 2006). Participants with a lower level of English may speak more animatedly to compensate. In order to be understood well they illustrate and move around more. Hillman (2010) found that people sometimes use speech prompting gestures when they can’t think of the right word or phrase. These may also contribute to the high animation. The factors cognitive load and animation predicted whether a participant is telling the truth or lying in 57.3% (truth tellers 52.3%; liars 62.6%) of cases, which is similar to the level usually reached by lay
people without any training or using specific cues (Bond and DePaulo, 2006; Vrij, 2008).

**Limitations of using non-verbal cues to lying in an airport situation.**

A good lie detection tool at an airport should be both accurate and quick. Some airports handle hundreds of thousands of passengers a day (e.g. in 2010 London Heathrow handled 65,747,200 passengers, BAA airports ltd., 2011). Since very few passengers are actually intending to commit major crimes even a false positive rate of 1% would lead to hundreds of falsely accused passengers every day. This would irritate passengers and use up vast amounts of resources from security personnel and customs officers. Quickly and efficiently classifying innocent passengers as innocent, whilst successfully picking out the occasional wrongdoers, would be ideal in this situation. Using non-verbal cues as a lie detection method does not give high enough detection rates. In particular, the very high percentage of false positives (47.7%) is unacceptable. Also the method is probably not quick enough. The interviewer would not be able to keep track of all 17 cues in real life. The interviews would need to be filmed and analysed. This would take too much time to perform on even a very small sample of passengers.

Also this experiment found that the non-verbal cues are not just affected by whether the passenger is telling the truth or lying, but also by their ethnicity, how well they spoke English and what questions are being asked. The cues were more useful in predicting the participants’ level of English, their ethnicity and which type of questions were answered than veracity. These results mean that analysing someone’s non-verbal behaviour with these cues is not very effective at detecting
deception. Other methods may be more useful. The next chapter will study the use of thermal imaging as a possible technique for detecting deception.
Chapter five: Thermal Imaging as a Lie Detection Tool at Airports.


**Introduction.**

This chapter contains the thermal imaging element of Experiment 1. Thermal imaging is a technique with which changes in facial skin temperature patterns (and thus blood flow) are detected via special cameras. It has been suggested that the technique can be used as a lie detection tool. The assumption is that liars will show instantaneous warming as part of a fright-flight response. Pavlidis, Eberhardt, and Levine (2002a) published their thermal imaging lie detection experiment in *Nature* and claimed that the technique has “an accuracy comparable to that of polygraph examination…and has potential for application in remote and rapid screening, without the need for skilled staff or physical contact” and “it could be used for instantaneous lie detection without the subject even being aware of the test” (Pavladis et al., 2002a, p. 35).

Perhaps because of the boldness of the claims, the Pavladis et al. (2002a) article did not remain unnoticed, being cited repeatedly in scientific journals. Not surprisingly, it also attracted considerable attention from practitioners who sought to exploit thermal imaging’s lie detection capacity, a critical skill in today’s security-conscious world. An obvious application is airport screening, where the facial thermal patterns of every passenger could be measured non-intrusively thereby instantaneously revealing the mendacity of terrorists and other wrongdoers. Security companies working at international airports have approached experts in lie detection several times about the viability of thermal imaging at their airports. Such a complex, costly venture
should not be contemplated without proper research, especially as there is no direct evidence of its success.

The use of thermal imaging as a lie detection tool is controversial. The key issue is that the surge of blood flow can occur for reasons other than deception (National Research Council, 2003). Airports by their nature are places full of people who are anxious for all manner of reasons, and individuals who are anxious will show a surge of blood flow. They could be anxious because they are afraid of flying, are worried about missing the plane, are on their way to an important but problematic business meeting, and so on. In addition, excitement will also result in a surge of blood flow. Airport passengers could be excited about the act of flying, about going on holiday, about just spotting an attractive person, and so on. Measuring each passenger’s facial blood flow remotely and without them being aware of the test is thus not likely to be effective for mass screening at airports. Perhaps thermal imaging can be used to detect deception in an alternative way at airports, as part of an interview protocol? We examined this in the present experiment.

Pavlidis et al. (2002a) used an interview protocol that is typically used in polygraph examinations. In their experiment the relevant question was “Did you steal the $20?” In an airport setting an equivalent question could be “Are you involved in illegal activities?” The problem is that questions about wrongdoing may also be anxiety-provoking for innocent interviewees, making it difficult to discriminate between innocent and guilty interviewees (National Research Council, 2003).

Because the experiment was conducted airside in an airport, the researchers had to be escorted by airport security personnel whenever they were past the security
control. That means that time to carry out our experiment was limited. Ideally, different interview protocols would have been compared but, logistically, that would have been problematic. Given the unique opportunity to conduct the experiment at an international airport, it was decided to spend the allotted time using an interview protocol with which above chance lie detection accuracy rates of 72% - 74% had been obtained previously (Vrij, Granhag, Mann & Leal, 2011). This interview protocol has been shown to create more cognitive load in liars than in truth tellers. Vrij et al., (2011) found liars contradicted themselves more and their stories sounded less plausible. Contradictions and implausibility are signs of cognitive load (Köhnken, 2002, 2004). Mentally taxing interview protocols have been shown to increase arousal in liars (Vrij, Mann, Fisher, Leal, Milne, & Bull, 2008). It was expected that the information gathering style interview that was used would circumvent the anxiety inducement problem.

The interviewers in experiment one were asked to make instant veracity judgements. They had no access to the thermal imaging recordings and could therefore rely only on passengers’ speech and non-verbal behaviour. In previous research using this interview protocol, the interviewers correctly classified 72% of truth tellers and 74% of liars (Vrij et al., 2011). This is a high percentage as accuracy rates, based on judging speech and demeanour, typically fall in the 45% - 60% range (Bond & DePaulo, 2006; Vrij, 2008). Good performance could be attributed to several factors, including the interviewer’s skill, efficacy of the interview protocol, and type of lie (see Discussion). Alternatively, the good performance was just a matter of luck. The outcomes of this part of Experiment 1 one will shed light on this luck factor and examine whether thermal imaging might increase accuracy rates beyond those obtained with this interview protocol.
Method.

Participants.

The sample used here are a subsample of the participants described in Chapter Three. Only a subsample was used for technical reasons, explained in detail in the “technical limitations” section. A total of 51 participants took part, of whom 39 (77%) were male and 12 (23%) were female. Their age ranged from 18 to 60 with an average age of \( M = 32.83 \) years (\( SD = 11.4 \)). Most participants were going home \(( n = 19, 37\%)\) or visiting family \(( n = 11.21\%)\). The other five reasons for traveling were going on holiday \(( n = 10, 20\%)\), going to work \(( n = 4, 8\%)\), flying for business reasons \(( n = 3, 6\%)\), going to study \(( n = 3, 6\%)\) and visiting friends \(( n = 1, 2\%)\). The participants were of 22 different nationalities and came from Western Europe \(( n = 15\)\), Eastern Europe \(( n = 3\)\), Arabic countries \(( n = 8\)\), Iran \(( n = 12\)\), Pakistan/India \(( n = 9\)\) and China/Japan \(( n = 4\)\).

Thermal imaging coding.

We used a ThermoPro TP8 IR thermal camera. The data were analysed using Guide IrAnalyzer V1.7. The thermal recordings display three data points per second: (i) minimum temperature, the temperature of the coolest pixel in the image, which would be the wall in the background; (ii) average temperature, the average temperature of all pixels in the image; and (iii) maximum temperature, the temperature of the hottest pixel in the image, which would be the participants’ skin temperature, as the participant is warmer than the surrounding area. The dependent variable was created by averaging for each of the six phases the maximum temperatures of all seconds in that phase. It is possible to select certain areas of
interest using the technology, so that one can select only the face of the participant for analysis. We found that for maximum temperature there was a very high correlation ($r = .98$) between the maximum temperature of the face alone and the whole image, demonstrating that the hottest pixel is indeed in almost all cases in the participant’s face.

The baseline skin temperature was taken prior to the main interview while the participant sat in front of the clearly visible camera. In this baseline period the interviewer introduced herself to the participant. She also mentioned the participants’ participant number and asked the first question that was answered truthfully by all participants “Where are you going to fly to today?”

**Technical limitations.**

There were some major practical issues with the thermal imaging coding. Firstly, the thermal image has no sound and no reliable clock. This makes it impossible to judge how far along the file is. The only way to know what the interviewee was saying at a particular moment in the thermal video is by running a normal video alongside the image and comparing the interviewees’ movements, pausing one of video’s as soon as they start running out of synchrony. This is time consuming.

A second practical issue is that there seem to be artefacts in the output (see Figure 5.1). It is not uncommon for the output to state that an interviewee’s maximum temperature is 42 degrees Celsius, which is surely impossible. For many participants the maximum temperature rises steadily until impossible highs of 41 degrees Celsius or higher, then the program seems to recalibrate, the temperatures drops and starts drifting to higher temperatures again. Since it is highly unlikely that
participants actually experience these temperature differences it must be concluded that these are artefacts. These artefacts happen in many participants data. In order to obtain the 51 participants with unartefact data, 144 participants needed to be filmed and their data analysed. That means that only 35%, of the thermal videos were actually useful for analysis.

Figure 5.1: *An example of artefacted data.*

![Graph showing temperature changes over time](image)

**Dependent variables.**

There were six dependent variables: The maximum skin temperatures during (i) baseline; (ii) Phase 1, general; (iii) Phase 2, destination plans and past; (iv) Phase 3, plans on plane; (v) Phase 4 reassurance; and (vi) Phase 5, lie. The technique makes it possible to use the minimum, average and maximum temperature of the
area in the image. The maximum temperature was used, because the minimum and the average temperatures are affected by the parts of the image that are not the participants face, e.g. their hair, glasses and the wall behind them. In addition, we created the ‘largest difference score’. We calculated the difference scores between the baseline and each of the other five phases. For each participant, the largest of the five difference scores was selected and used as the largest differences score dependent variable. Finally, we calculated the interviewers’ accuracy in classifying truth tellers and liars.

**Results.**

**Mean skin temperatures as a function of veracity.**

A 2 (Veracity) x 6 (Phase) ANOVA was carried out with the skin temperature as dependent variable. The analysis revealed main effects for Veracity, $F(1, 48) = 6.73, p < .05, \eta^2 = .12, \text{d} = .30$, and Phase, $F(5, 48) = 12.93, p < .01, \eta^2 = .21$, and a Veracity x Phase interaction effect, $F(5, 48) = 3.74, p < .05, \eta^2 = .07$. The Veracity main effect revealed that liars’ skin temperature was higher ($M = 35.72$ degrees Celsius, $SE = .16$) than truth tellers’ skin temperature ($M = 35.15$ degrees Celsius, $SE = .16$). The Phase main effect revealed that the skin temperature rose with each phase, and was significantly higher in Phases 2, 3, 4 and 5 than in baseline and Phase 1 (Table 5.1).

The interaction effect revealed that the Phase main effect occurred only in liars. Their skin temperature rose with phase, $F(5, 20) = 5.33, p < .01, \eta^2 = .57$, and was significantly higher in Phase 1 than during the baseline and significantly higher in Phases 2, 3, 4 and 5 than in Phase 1. The temperature of truth tellers remained constant during the baseline and interview, $F(5, 20) = 2.03, p = .11, \eta^2 = .34$ (Table 5.1).
Further analyses revealed that truth tellers’ and liars’ skin temperature did not differ significantly from each other at baseline, $F(1, 49) = .82$, $ns, \eta^2 = .02$, at Phase 1, $F(1, 49) = 2.45$, $ns, \eta^2 = .05$ or at phase 4 $F(1, 49) = 3.93$, $ns, \eta^2 = .07$, but that liars had significantly higher skin temperatures than truth tellers at Phase 2, $F(1, 49) = 6.09, p < .05, \eta^2 = .11$, $d = .69$, Phase 3, $F(1, 49) = 8.43, p < .01, \eta^2 = .15$, $d = .82$ and Phase 5 $F(1, 49) = 6.92, p < .05, \eta a^2 = .13$ $d = .75$.

Table 5.1. Skin temperature as a function of phase

<table>
<thead>
<tr>
<th></th>
<th>Total M</th>
<th>SD</th>
<th>Truth tellers M</th>
<th>SD</th>
<th>Liars M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>35.14$^a$</td>
<td>.80</td>
<td>35.04$^a$</td>
<td>.83</td>
<td>35.24$^a$</td>
<td>.78</td>
</tr>
<tr>
<td>General</td>
<td>35.22$^a$</td>
<td>.88</td>
<td>35.03$^a$</td>
<td>.93</td>
<td>35.41$^b$</td>
<td>.79</td>
</tr>
<tr>
<td>Destination</td>
<td>35.41$^b$</td>
<td>.93</td>
<td>35.10$^a$</td>
<td>.90</td>
<td>35.71$^c$</td>
<td>.88</td>
</tr>
<tr>
<td>On plane</td>
<td>35.49$^b$</td>
<td>1.01</td>
<td>35.10$^a$</td>
<td>.89</td>
<td>35.87$^c$</td>
<td>.99</td>
</tr>
<tr>
<td>Reassurance</td>
<td>35.55$^b$</td>
<td>.94</td>
<td>35.29$^a$</td>
<td>.92</td>
<td>35.80$^c$</td>
<td>.92</td>
</tr>
<tr>
<td>Lie</td>
<td>35.36$^b$</td>
<td>.88</td>
<td>35.32$^a$</td>
<td>.86</td>
<td>35.94$^c$</td>
<td>.79</td>
</tr>
</tbody>
</table>

Truth – Lie Classifications, Between-Subjects Analyses.

ROC analyses were carried out for the between-subjects comparisons where truth tellers and liars were compared at baseline and the three interview phases (Table 5.2). The results revealed that detection accuracy for Phases 2, 3 and 5 were significantly superior to chance (note that the lower bounds of the areas computed for these phases fell above .5, the level of chance). The mean areas under the curve (AUC) were .71 for Phase 2; .74 for Phase 3 and .70 for Phase 5, the equivalent of Cohen’s $d’s$ of $d = .80; d = .83$ and $d = .75$ respectively (Rice & Harris, 2005). Effect sizes of $d = .80$ or higher are considered ‘large’ (Cohen, 1988). Discriminant
analyses revealed that in Phases 2, 68% of truth tellers and 62% of liars were correctly classified, that in Phase 3, 68% of truth tellers and 65% of liars were correctly classified and that in Phase 5, 64 % of truth tellers and 68% of liars were correctly classified. However, these results fall short of the performance of the interviewers: They correctly classified 72% of truth tellers and 77% of liars (AUC = .75, d = .96).

Regarding the thermal imaging data positive and negative predictive values were calculated for the baseline - Phase 3 comparison (the most successful comparison). They are dependent on the cut off score, which is arbitrary. The best results were obtained with a 35.29 degrees Celsius cut-off score. In that situation the positive predictive value was .71 and the negative predictive value was .74.

We further conducted a stepwise discriminant analysis with both Phase 3 and interviewers’ veracity judgements as predictors. Both variables contributed significantly to the final model, \(X^2(2, N = 51) = 18.07, p < .01\), with the interviewers’ veracity judgements being the strongest predictor (function coefficients were .79 and .56 respectively). The cross-validated accuracy rates revealed that on the basis of these two predictors the same number of truth tellers (72%) and liars (77%) were correctly classified as on the basis of the interviewers’ veracity judgements alone.

In order to see whether the interviewers’ judgement relates to the thermal imaging judgements, we correlated the thermal imaging data with the interviewers’ accuracy rates and veracity judgements. Spearman analyses revealed negative (the higher an interviewee’s skin temperature, the less accurate the interviewers) but not significant correlations between thermal imaging data and accuracy rates (at baseline \(r = -.18, p = .20\); in Phase 1 \(r = -.24, p = .10\); in Phase 2 \(r = -.08, p = .57\); in Phase 3 \(r = .10, p = .70\).
Pearson analyses revealed positive (the higher the interviewee’s skin temperature, the more likely it was that the interviewer made a lie judgement) but not significant (except in Phase 5) correlations between thermal imaging data and veracity judgements (at baseline $r = .16, p = .27$; in Phase 1 $r = .20, p = .16$; in Phase 2 $r = .24, p = .09$; in Phase 3 $r = .25, p = .08$; in Phase 4 $r = .20, p = .16$; in Phase 5 $r = .312, p = 0.03$).

**Truth – Lie Classifications, Within-Subjects Analyses.**

In the within-subjects ROC analysis the dependent variable was the largest difference score. The ROC analysis revealed that detection accuracy for this variable was significantly superior to chance (.71, $p = 0.01$), and that the lower bound of the areas for this variable fell above .5 (Table 5.2, Figure 5.2). A discriminant analysis revealed that on the basis of this within-subject variable 72% of truth tellers and 54% of liars were classified correctly. These accuracy rates are similar those obtained in the between-subjects analyses and again fall short of the performance of the interviewers. Again positive and negative predictive values were calculated. The best results were obtained with a cut-off score of .6 degrees Celsius difference between highest and lowest temperature. In that situation the positive predictive value was .70 and the negative predictive value was .67.

We further conducted a stepwise discriminant analysis with both the largest difference score and interviewers’ veracity judgements as predictors. Both variables contributed significantly to the final model, $X^2(2, N = 51) = 20.65, p < .01$, with the interviewers’ veracity judgements being the strongest predictor (function coefficients were .86 and .65 respectively). The cross-validated accuracy rates revealed that on the basis of these two predictors 68% of truth tellers and 85% of liars were correctly
classified. The interviewers’ accuracy rates did not correlate significantly with the largest difference score (Spearman’s $r = -0.05, p = .75$). The interviewers’ veracity judgements also did not correlate significantly with the largest difference score (Pearson’s $r = .07, p = .65$).
Table 5.2. Area under the ROC curves and related statistics and discriminant analyses regarding truth/lie classifications

<table>
<thead>
<tr>
<th></th>
<th>ROC</th>
<th>Discriminant</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC</td>
<td>SE</td>
<td>95% confidence</td>
<td>Eigen</td>
<td>Wilks’</td>
<td>X2</td>
<td>truth</td>
<td>lie</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>interval</td>
<td>value</td>
<td>lambda</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Between-Subjects Analyses</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>.60</td>
<td>.08</td>
<td>.44 - .75</td>
<td>.02</td>
<td>.98</td>
<td>.80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phase 1</td>
<td>.64</td>
<td>.08</td>
<td>.49 - .80</td>
<td>.05</td>
<td>.95</td>
<td>2.37</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phase 2</td>
<td>.71*</td>
<td>.07</td>
<td>.56 - .85</td>
<td>.12</td>
<td>.89</td>
<td>5.68*</td>
<td>68%</td>
<td>62%</td>
<td></td>
</tr>
<tr>
<td>Phase 3</td>
<td>.74**</td>
<td>.07</td>
<td>.60 - .88</td>
<td>.17</td>
<td>.85</td>
<td>7.70**</td>
<td>68%</td>
<td>65%</td>
<td></td>
</tr>
<tr>
<td>Phase 4</td>
<td>.66</td>
<td>.08</td>
<td>.50-.81</td>
<td>.08</td>
<td>.93</td>
<td>3.75</td>
<td>68%</td>
<td>62%</td>
<td></td>
</tr>
<tr>
<td>Phase 5</td>
<td>.70*</td>
<td>.08</td>
<td>.55-.85</td>
<td>.14</td>
<td>.87</td>
<td>6.40*</td>
<td>64%</td>
<td>68%</td>
<td></td>
</tr>
<tr>
<td>Interviewers’ judgements</td>
<td>.75**</td>
<td>.07</td>
<td>.61 - .88</td>
<td>.32</td>
<td>.76</td>
<td>13.32**</td>
<td>72%</td>
<td>77%</td>
<td></td>
</tr>
<tr>
<td>Phase 3 and interviewers’ judgements</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.46</td>
<td>.69</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Within-Subjects Analyses</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Largest difference score</td>
<td>.71**</td>
<td>.07</td>
<td>.56 - .85</td>
<td>.15</td>
<td>.87</td>
<td>6.82**</td>
<td>72%</td>
<td>54%</td>
<td></td>
</tr>
<tr>
<td>Largest difference score and interviewers’ judgements</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.45</td>
<td>.69</td>
<td></td>
<td></td>
<td></td>
<td>17.87**</td>
<td>68%</td>
<td>85%</td>
<td></td>
</tr>
</tbody>
</table>

* p < .05  ** p < .01
Discussion

*Thermal Imaging Results.*

Thermal imaging was initially proposed as a remote and rapid screening device for which no skilled staff are required and without subjects even being aware of the test taking place (Pavladis et al., 2002a). Using thermal imaging in this way is problematic because the surge of blood flow can occur for numerous reasons other than deception (National Research Council, 2003). The results provide evidence that thermal imaging cannot be used in this manner: the baseline measure, which was taken after passengers had been instructed to be truthful or deceptive, showed no differences in blood flow.
Thermal imaging may have potential as a lie detection tool if passengers are actually interviewed. Indeed, the experiment revealed a clear difference in skin temperature patterns for liars and truth tellers when being interviewed. Liars’ skin temperature rose during the interview whereas truth tellers’ skin temperature remained constant. At Phase 3, the difference in skin difference between truth tellers and liars was the largest, .77 degrees Celsius. Albeit not high, the effect size was considerable ($d = .80$) reflecting the low variability ($SD = .94$). Between-subjects analyses (comparisons of skin temperatures of liars and truth tellers during each phase) and within-subjects analyses (comparisons of difference in skin temperature between the baseline and subsequent phases during the interview for liars and truth tellers) revealed similar findings: between 64% and 68% of truth tellers and 62% and 68% of liars were classified correctly. These accuracy rates, however, are rather low for screening purposes at airports. Of particular concern are the relatively low accuracy rates for truth tellers. In airport settings the vast majority of passengers have no bad intentions and thus can be considered ‘truth tellers’. Classifying so many truthful people incorrectly is unacceptable as a public policy, particularly given the consequences it may have for truth tellers, such as further scrutiny or even a refusal to board their intended flight. A lie detection tool at airports should be virtually flawless in identifying those truth tellers to be of practical use.

**Interviewers’ Assessments.**

The interviewers correctly classified 72% of truth tellers and 77% of liars and outperformed both the thermal imaging results and the non-verbal cues results from Chapter four. In addition, accuracy rates based on interviewers’ judgements and thermal imaging recordings combined were the same as the accuracy rates obtained by the thermal imaging recordings. The interviewers had no access to thermal
imaging data. Furthermore, their accuracy rates (and veracity judgements) did not correlate significantly with the thermal imaging data. The only information available to them was the passengers’ answers and their demeanour. Speech and demeanour thus revealed more information about deceit than skin temperature. We did not systematically examine the cues the interviewers used. However, later, conversations with the interviewers revealed that they relied on contradictions in the story and the plausibility of the answer.

The accuracy rates were almost identical to those obtained by Vrij, Granhag et al. (2011) who used a similar interview protocol, and are considerably higher than typically achieved when people judge speech and demeanour (Bond & DePaulo, 2006; Vrij, 2008). There are several possible explanations for this. Accuracy is perhaps related to the type of lie. In most deception research truth tellers and liars discuss past activities, whereas in the current protocol they discuss their future actions, or, intentions. Perhaps lies about intentions are easier to detect than lies about past activities. This has been examined in only one experiment. Truth tellers and liars discussed their intentions and their past activities. Observers could distinguish truth tellers’ and liars’ intentions but not their past activities (Vrij, Leal, Mann, & Granhag, 2010). Perhaps people are more practiced in lying about what they have done than in lying about their intentions, and practice increases skill. Related to this, perhaps people have a better notion of what sounds convincing when they discuss their past activities than when they discuss their intentions. Knowing what sounds convincing is an essential element in trying to make a convincing impression.

Of course, there are more mundane explanations for why the interviewers’ accuracy rates were relatively high. The present experiment and Vrij et al.’s (2011) experiment used the same two interviewers and perhaps they are particularly skilful
lie detectors. They are experienced lie detection researchers and get feedback on how accurate they are after each experiment. Alternatively, our interview protocol is perhaps particularly effective in eliciting cues to deceit. Future research that manipulates the interviewers and interview protocols will provide information about these explanations.

**Issues to Consider.**

Six aspects of our experiment merit attention. First, the difference in skin temperature between truth tellers and liars was the highest in Phase 3 where the participants discussed their activities during the flight. One might expect the difference in temperature to be more substantial in Phase 2 where the participants told the truth or lied about the actual trip, or phase 4 where they are challenged to provide extra information. Our tentative (untested) explanation is that there is some delay in the thermal imaging responses, so that Phase 2 had a carry-over effect in Phase 3. The difference between truth tellers and liars gets smaller again in Phases 4 and 5, because truth tellers now become warmer. Perhaps because they feel more aroused by the challenge for extra information, which implies that the interviewer is not satisfied with the answers they gave previously.

Second, the fact that passengers need to be interviewed restricts the use of the thermal imaging test, because only a small percentage of passengers at airports can be interviewed. This is probably not how Pavlidis et al. (2002a, 2002b) had envisaged a thermal imaging test being used, but we believe that there is no way around this limitation.

Third, as mentioned earlier, a different interview protocol than typically used in anxiety-based lie detection interviews, was used. A typical anxiety-based question
would have been “Are you involved in illegal activities?” The problem is that questions about wrongdoing may also increase the anxiety level of innocent interviewees, making it difficult to discriminate between innocent and guilty interviewees (National Research Council, 2003). The higher temperature for truth tellers in Phase 4 may be caused by this. The difference between truth tellers and liars was smaller in Phase 4 than in Phase 3. An interview protocol in which passengers are asked about wrongdoings may also be undesirable for humanitarian reasons. There is no evidence that the passengers who are interviewed are actually involved in illegal activities, so why ask them such loaded questions?

A different type of protocol was used which asked passengers to discuss their forthcoming trip. By asking questions about their forthcoming trip rather than about wrongdoing we tried to avoid making innocent people feel anxious. We succeeded: Truth tellers did not become more anxious as a result of our questioning. A thermal imaging interview protocol, however, will discriminate between truth tellers and liars only if the questioning makes liars more anxious. The results revealed that in this case it did. We believe that the different pattern of arousal between truth tellers and liars was caused by differences in perceived difficulty of the interview protocol. Interviewees typically become anxious out of fear of not being believed. When interviewees struggle with their answers, which happened to liars in our protocol, such fear is likely to arise.

Fourth, the stakes are higher for real-life passengers who are subjected to a thermal imaging lie detection device than for our experimental participants. In real life negative stakes are involved. That is, passengers who fail the test could be excluded from flying. We believe that such a high stakes situation will not automatically improve the diagnostic value of a thermal imaging lie detection tool.
Liars may well become more aroused in such a high stakes situation when answering the questions, but truth tellers may also become aroused when they realise the consequences of failing to convince the interviewer. If both truth tellers and liars become more aroused, the lie detection tool will not become more diagnostic. Of course, whether increasing the stakes will change the diagnostic value of the instrument is an empirical question. This question cannot be tested easily. When stakes are introduced to participants, thus when they are threatened with being taken off the plane if they fail the test, passengers are unlikely to volunteer to participate in the experiment. It means that participation has to be made compulsory which goes against the ethical guidelines of psychological research.

Fifth, we do not argue that a variant of our cognitive load based interview protocol should replace the protocol currently used in polygraph examinations (Ben-Shakhar & Elaad, 2003). There is an important difference between our protocol and a polygraph protocol. In our protocol, we encourage examinees to give lengthy and detailed answers as liars will experience cognitive load only if they are asked to provide details. This makes our protocol unsuitable for polygraph examinations where the examinee is required to provide single-worded “yes” or “no” answers. One reason for asking examinees to give short answers in polygraph examinations is that vocal activity (e.g., speech) affects the measures recorded by the polygraph equipment.

Sixth, of interest to security personnel at airports are lies that people tell about their intended future activities. Such lies have rarely been the topic of investigation, and, apart from the present experiment, we are aware of only three studies examining deceptive intent (Granhag & Knieps, 2010; Vrij, Leal, Mann, & Granhag, 2010; Vrij et al., 2011). There are similarities between lying about intentions and about past
activities. For example, liars may have difficulty in fabricating on-the-spot answers in both instances, resulting in lack of plausibility in both lies about intentions and past activities (Vrij et al., 2010). However, there are also differences. For example, a direct comparison between lying about intentions and past activities revealed that liars were equally as detailed as truth tellers in describing their intentions but less detailed than truth tellers in describing their past activities (Vrij et al., 2010). One aspect that often makes truth tellers’ stories about past activities more detailed is that there is a wealth of perceptual details that truth tellers have experienced during these past activities that they can recall (if they still remember them). In contrast, when discussing their intentions about a forthcoming trip, truth tellers have not yet experienced anything that restricts the amount of detail they can recall.

Granhag and Knieps (2010) focused on a different aspect of intentions: The planning of intentions. This is relevant for airport research. When passengers arrive at the airport they have done some preparation and planning for the trip. They, or someone else on their behalf, have bought tickets and may have arranged hotels. They may have conducted some research into their destination, etc. Liars’ preparations of their false stories about their forthcoming trips differ from truth tellers’ preparations of their real forthcoming trip. In Granhag and Knieps’ (2010) experiment, mock-suspects who told the truth about their intentions agreed more frequently that planning their future actions evoked mental images than did mock-suspects who lied about their intentions. In addition, lying suspects who claimed to have activated a mental image during the planning phase provided verbal descriptions of the most dominant mental images that were less rich in content than truth tellers. In other words, asking questions about the past (planning) may reveal deception about the future (intent).
The Practical Application of Thermal Imaging.

We conclude with outlining the practical application of thermal imaging research. Based on our analysis of the NRC report and the present (baseline) data it is unlikely that thermal imaging can be used effectively at airports as a general screening device. Thermal imaging will classify too many non-deceptive passengers as liars because, as we outlined above, they may be anxious for all sorts of non-deceptive reasons. In addition, our data showed that, prior to being interviewed, thermal imaging did not register differences between truth tellers and liars. Truth tellers and liars displayed different blood flows during the interview. The technique does not give instantaneous results, the thermal images need to be analysed, so the classification is made at least several minutes after the interview. This means that the entire procedure would take at least 30 minutes. In airports with hundreds or thousands of passengers coming through, this would be unpractically long. Besides this, not enough truth tellers and liars could be correctly classified to be acceptable at airports. Of particular concern is the high percentage of incorrectly classified truth tellers, as most airport passengers are likely to be truthful. In addition, the thermal imaging accuracy rates were worse than the rates achieved by the two interviewers. This suggests that it would be more practical to invest resources to improve the interview approach rather than in improving the thermal imaging technique. The next part of this thesis will study the interview approach in more detail.
Part III: Verbal Cues. The Effect of Interviewees’ Previous Experience and Time Prompting

Chapter 6: Using Verbal Cues to Detect Deception Literature Review

In Part III of this thesis verbal cues to deception about intentions are examined. This chapter (Chapter 6) is a literature review of the previous research on verbal cues. In Chapter 7 Experiment 2 is described, how veracity and the interviewee’s previous experience of the discussed intention affect the content of their statement is examined. In Chapter 8 Experiment 3 is described, which tests the effect of asking specific time prompting questions in short interviews.

In this literature review, I discuss techniques to detect deception based on the content of the statement. These techniques can only classify statements that contain enough information. This means that the statement needs to be made via free recall or by answering open ended questions. Closed questions will elicit too little information to make reliable veracity judgements. The questions must be carefully phrased, because leading questions may affect the statement in a way that makes it harder to classify them correctly (Vrij, 2008). Research into investigative interviewing has developed techniques that help the interviewee to give sufficient and untainted information, e.g., the Cognitive Interview (Fisher & Geiselman, 1992; Milne & Bull, 1999), the PEACE model (Schollum, 2005) and the Strategic Use of Evidence (SUE) technique (Granhag, Strömwall & Hartwig, 2006). These interviews are designed to elicit statements that can be analysed. These statements can then be analysed in several different ways.

Statement Validity Assessment (SVA) (Köhnen, 2004) is a procedure that aims to verify the truthfulness of statements from witnesses or victims of crime. It
was developed by Udo Undeutsch who formulated the Undeutsch hypothesis, which states that statements which describe an experienced event will have characteristics that are different from statements that describe events that were not experienced (Steller, 1989). The SVA procedure has several phases. First, the case file needs to be studied for information about the interviewee, the event and the background. Then the interviewee is interviewed about the event of interest. The interviewee’s background, personality and cognitive abilities are also assessed, in order to establish what level of answers could be expected from this type of interviewee (e.g. a child’s answers will differ from those of an adult). The interview about the event is then transcribed and analysed using Criteria Based Content Analysis (CBCA). The CBCA contains 19 criteria, all of which are perceived as truth criteria, and it is assumed that they occur more frequently in truthful than in deceptive reports. The criteria are listed and described in Table 6.1.
Table 6.1: The CBCA criteria, continued on next page.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General characteristics</strong></td>
<td></td>
</tr>
<tr>
<td>Logical structure</td>
<td>Coherence and consistency</td>
</tr>
<tr>
<td>Unstructured production</td>
<td>Information presented non-chronologically</td>
</tr>
<tr>
<td>Quantity of details</td>
<td>Rich in detail</td>
</tr>
<tr>
<td><strong>Specific contents</strong></td>
<td></td>
</tr>
<tr>
<td>Contextual embedding</td>
<td>Events are placed in time and location</td>
</tr>
<tr>
<td>Descriptions of interactions</td>
<td>The perpetrator and interviewee react to each other</td>
</tr>
<tr>
<td>Reproduction of conversation</td>
<td>The interviewee quotes things said at the event</td>
</tr>
<tr>
<td>Unexpected complications</td>
<td>Unexpected things uncommon for the event are described (e.g. a car alarm went off)</td>
</tr>
<tr>
<td>Unusual details</td>
<td>Unique or surprising, but meaningful details (e.g. someone spoke with a stutter)</td>
</tr>
<tr>
<td>Superfluous details</td>
<td>Details not necessary to the event</td>
</tr>
<tr>
<td>Accurate, misunderstood details</td>
<td>Details mentioned but not understood by the interviewee</td>
</tr>
<tr>
<td>Related external associations</td>
<td>Events that are not part of the event are also mentioned</td>
</tr>
<tr>
<td>Accounts of subjective mental state</td>
<td>Description of what the interviewee felt or thought</td>
</tr>
<tr>
<td>Attribution of</td>
<td>Description of what the interviewee believed the perpetrator</td>
</tr>
</tbody>
</table>
perpetrator’s mental state felt or thought.

**Motivation related contents**

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spontaneous corrections</td>
<td>The interviewee corrects him/herself</td>
</tr>
<tr>
<td>Admitting lack of memory</td>
<td>The interviewee says he/she doesn’t remember something</td>
</tr>
<tr>
<td>Raising doubt about testimony</td>
<td>The interviewee indicates something sounds implausible or unlikely</td>
</tr>
<tr>
<td>Self-deprecation</td>
<td>Details unfavourable to the interviewee or self-incriminating</td>
</tr>
<tr>
<td>Pardoning perpetrator</td>
<td>Interviewee excuses perpetrator behaviour</td>
</tr>
</tbody>
</table>

**Offence specific elements**

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Details characteristic of the offence</td>
<td>Elements known by professionals to be typical for the event, but that are counterintuitive for non-professionals.</td>
</tr>
</tbody>
</table>

The statement is rated on each criterion, either by counting how often the criterion is present or more commonly by rating how commonly it occurs on a three or five point scale. These scores are then taken together to create a raw score. The statement is then evaluated on: the characteristics of the interviewee (e.g. susceptibility to suggestion); the characteristics of the interview (e.g. leading or coercive); the motivation of the interviewee, (e.g. are there outside pressures on the interviewee to make this statement) and the consistency (e.g. is the statement consistent with other statements by the interviewee or others, is the statement consistent with already known facts?). This evaluation is important, because these factors will affect the raw score of the statement and their effect needs to be separated from the effect the veracity of the interviewee will have on the criteria.
Vrij (2005) reviewed 37 studies to test the effectiveness of the CBCA part of SVA. It was found that CBCA classified 73% of truth tellers and 72% of liars correctly, which is not sufficient for SVA to be used as the crucial piece of evidence in a criminal court, where the requirement is that something must be proven beyond reasonable doubt.

The Reality Monitoring approach to lie detection also detects deception based on the content of the statement. Reality Monitoring (RM) was created by Johnson and her colleagues (Johnson, 1988; Johnosn & Raye, 1981; Johnson et al., 1988) to discriminate between perceived events and imagined events. Its creators suggest that statements based on perceived events include more contextual (spatial and time) information, more sensory information, and more semantic detail (e.g. physical, functional or emotional information about the topic), but fewer cognitive operations. The goal of Reality Monitoring was not to detect deception, but rather to investigate the mental processes behind source monitoring and for use in clinical settings.

Johnson et al. (1988) created the Memory Characteristics Questionnaire (MCQ), which consists of 39 items on which the memory is rated by the person who has the memory. The items refer to the clarity, vividness and context of the memory. The MCQ has been used for participants to rate their episodic future thoughts as well. Granhag & Knieps (2010) found that truth tellers rated their mental images of a future event more vividly and in more detail than liars.

RM was adapted as a verbal lie detection tool by Sporer and Kuepper (as cited in Sporer, 2004). They created the Judgement of Memory Characteristics Questionnaire (JMCQ). The JMCQ rates statements on 43 criteria, which are similar to the MCQ criteria. Unlike in the MCQ, which is used to rate one’s own memories, the JMCQ is used to rate statements made by someone else. For each criterion a
score on a 7 point scale, from barely present to strongly present, is given. The 43 items were factored into 8 scales listed in Table 6.2. Similar to CBCA, the cues, with the exception of cognitive operations, are associated with telling the truth. Some cues, like clarity/vividness, realism and time information are reliably associated with truth telling, while sensory information is sometimes found to be related to lying instead of truth telling, as would be hypothesised (Sporer & Kuepper, 2004).

Table 6.2: The scales of the JMCQ.

<table>
<thead>
<tr>
<th>Scale</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clarity and vividness</td>
<td>The statement is clear rather than vague</td>
</tr>
<tr>
<td>Sensory information</td>
<td>The statement contains information about what the interviewee saw, heard, smelt, tasted or felt physically.</td>
</tr>
<tr>
<td>Spatial information</td>
<td>Information on the arrangement of person and objects</td>
</tr>
<tr>
<td>Time information</td>
<td>Information on when the event occurred</td>
</tr>
<tr>
<td>Emotion and feelings</td>
<td>Information on how the interviewee felt psychologically</td>
</tr>
<tr>
<td>Reconstructability of the story</td>
<td>The statement gives enough information to reconstruct what happened</td>
</tr>
<tr>
<td>Realism</td>
<td>The statement is plausible, realistic</td>
</tr>
<tr>
<td>Cognitive operations</td>
<td>The statement contains inferences the interviewee made</td>
</tr>
</tbody>
</table>

Masip et al. (2005) reviewed 15 studies using the Reality Monitoring approach. They found that not all the studies had adapted RM in exactly the same way, but that the overall results were promising. All studies found significant effects; accuracy was between 64% and 85%. Some of the measures, e.g. spatial and temporal information and realism, were better predictors of truth than others, like emotions or cognitive operations. They concluded that the Reality Monitoring
The CBCA and RM approaches are not measuring completely different factors. Sporer (1997) tested CBCA and Reality Monitoring on 40 truthful and 40 deceptive statements. They found that certain CBCA criteria correlated significantly with some RM scales, e.g. CBCA’s logical consistency correlated highly with RM’s clarity/vividness, reconstructability and realism). They used factor analysis to combine both the CBCA criteria and the RM scales into five factors: logical consistency/realism; quantity of details and contextual embedding; feelings and thoughts; clarity; verbal and nonverbal interactions. Each factor loaded both CBCA criteria and RM scales, showing that CBCA and RM contain similar criteria.

CBCA and RM have served as a basis for newly designed interview protocols designed to detect deception. Assessment Criteria Indicative of Deception (ACID) is an example of this (Colwell, Hiscock & Memon, 2002; Colwell, Hiscock-Anisman, Memon, Taylor, & Prewett, 2007; Colwell et al., 2008). CBCA uses statements elicited via the request to freely recall the event. RM usually uses free recall as well, although it is also possible to code more scripted interviews. ACID uses a special interview to create the statements and is only used to analyse statements resulting from an ACID interview. This script contains several phases. In the free recall phase the interviewee is asked to describe the event in as much detail is possible. The interviewee is then asked to describe the event three more times, each time using a different mnemonic: recall from different perspective; recall in reverse order and retell the event. The mnemonics are meant to help truth tellers remember more about the event, while making the interview harder for liars. The different recalls are separated by series of yes/no questions. Only one of these blocks
is coded, the following question that appears in the third block: “Could you have made any mistakes in the information you have reported so far?” The other questions in the inferential blocks are not coded, but are included to increase cognitive load. The ACID technique codes for three types of detail in the statements: external, internal and contextual. External details are details providing sensory information. Internal details are information about the feelings and cognitive processes of the interviewee. Contextual details relate to information on where and when the event occurred. The response length is also measured. More detailed responses, longer responses and admitting to having possibly made a mistake are associated with telling the truth.

There are more lie detection techniques based on speech content and some of them are sold widely, such as the behavioural analysis interview (also known as the Reid technique) (Horvath, Blair & Buckley, 2007) and Scientific Content Analysis (SCAN). These tools do not seem to be able to discriminate truths and lies above the level of chance (Nahari, Vrij & Fisher, 2011) and are therefore not discussed further.

SVA, RM and those protocols based on them were all created to test the veracity of statements about past events. Previous research (e.g. D’Argembeau and Van der Linden, 2004; Gamboz, Brandimonte and De Vito, 2010) found that people’s mental images of past events are usually more detailed and vivid than images of future events. This means that we cannot simply assume that differences found between truth tellers and liars when they are interviewed on past events will also be found when they are talking about future events. The next two chapters report experiments that investigate the use of verbal cues to detect deception about intentions. The experiments use coding systems that are tailored to these experiments, but are grounded in and similar to SVA, RM and ACID.
Chapter 7: Experiment 2. The Effect of the Interviewees’ Previous Experience

In Experiment 2 we examine the verbal cues to deception that truth tellers and liars give in a statement about an intended trip. The verbal cues used in this study are mostly based on CBCA and RM. They are listed in the design section of this chapter and cover sensory information, contextual information, corrections, inconsistencies, cognitive operations and admitting lack of memory. The measure “Hope”, which indicates that the interviewee said they hope that something would occur, is not on the CBCA or RM list, but was added to examine the use of future related words. Liars are more ambivalent in their speech when lying (DePaulo et al. 2003). Using “hope” or “hopefully” could be a future specific way to express this ambivalence. The future is less clear than the past and liars may use this to their advantage by stressing that the future is uncertain. One way to do this is to use the word hope. This emphasises that they are not in control of the future and allows the liars to remain vague.

Experiment 2 tests the effect of having made the discussed trip before on the verbal cues in the statement. Research on thinking about the future suggests that an episodic future thought is based on memories of the past (D’Argembeau & Van der Linden 2004; Schacter, Addis & Buckner, 2008). Other researchers (Szpunar, Chan & McDermott, 2009) suggest that episodic future thoughts are influenced by the familiarity of the place where the future event is set.

A second reason that having experience with the discussed trip might affect the statement is that both truth tellers’ and liars’ goal is to be perceived as truthful. In order to do this truth tellers and liars may adopt different strategies and having experienced the event before may help them to develop and execute these strategies.
Hartwig, Granhag and Strömwall (2007) studied the strategies of truth tellers and liars prior to being interviewed. They found that liars were more likely to report a strategy than truth tellers. Truth tellers who did not have a strategy explained this by saying that they preferred to go into the interview without a plan, that they believed a strategy would not help, because they did not know what would be asked in the interview, or that they did not have a strategy because they were innocent. Regarding the latter, people often believe that their inner states are more visible to others than they really are (Gilovich, Stavitsky & Medvec, 1998). This leads some truth tellers to feel that their truthfulness is obvious to others and therefore they do not need to adopt strategies to convince others. In Hartwig et al. (2007), Liars who did not have a strategy, explained this by saying they did not know what the interview would be like and that they did not have time of think of a strategy. The ‘lack of time’ reason is possibly a consequence of the set up of Hartwig, et al.’s (2007) experiment (in which 10 minutes preparation time was included) rather than a reason that tells us something about the way liars plan their lies. Perhaps if they liars had had more than 10 minutes they would have used a strategy. Truth tellers who reported a strategy said they would tell the truth, remain calm, be cooperative and deny guilt firmly. Liars reported many different strategies. They said they would remain calm, be nice and pretend to be innocent. Regarding the statement content they reported that they would tell the truth as much as possible, firmly deny guilt, tell a detailed story, but avoid incriminating details, tell a consistent story, tell a story that seemed unrehearsed and say that others could vouch for their story.

Strömwall, Hartwig and Granhag (2006) found similar results with regard to verbal behaviour. They found that truth tellers simply try to tell the truth as the questions are asked, whereas liars try to keep the story simple, avoid hesitations and
give a lot of detail without incriminating themselves. Strömwall et al. (2006) also asked the participants to report their strategies regarding nonverbal behaviour. They found that both truth tellers and liars avoid making excess movements and attempted to maintain eye contact. Vrij, Mann, Leal and Granhag (2010) asked for the strategy of pairs of interviewees. They found that truth tellers reported using a strategy less frequently than liars, mostly because truth tellers believed that they would not need one. Pairs of truth tellers who did report a strategy said that they talked through everything that had happened to ensure they would remember and agree with each other. Liars with strategies reported that they would ensure that their stories matched. To achieve this they prepared specific details and in some cases agreed to control their nonverbal behaviour by keeping eye contact and remaining still.

Someone who lies about their intentions may be able to use previous experience to be more convincing. Leins, Fisher and Ross (2012) conducted a study in which participants were asked to give a lied alibi about where they were for a certain time frame the previous weekend. They found that 67% of the liars used the strategy of reporting an event they experienced on a different occasion as their lied alibi. Previous experience can help especially when liars use a ‘remain close to the truth’ or a ‘give a lot of detail’ strategy. Instead of talking about plans for what will happen, they can talk about things that have already happened. This will bolster the story and will decrease cognitive load, because less details have to be fabricated and remembered at the same time. This lower cognitive load will make the story smoother and help liars relax.

In order to examine how veracity and previous experience affect the content of a statement about intentions Experiment 2 was devised. Forty-two participants were asked to truthfully discuss the trip they were intending to make, and another 44
participants were asked to lie and pretend they intended to make a trip they did not intend to make. Four categories of questions were asked in the interview which combined address multiple facets of a trip: factual questions, questions about the core events of the trip, transportation questions, and questions about the planning of the trip. Questions about the facts of the trip (called ‘general questions’ hereafter) were hypothesised to be the most expected category, because most people expect factual questions (Vrij et. al. 2009; Granhag 2010). This expected category was compared to the three other categories which were different in focus, and all hypothesised to be unexpected by the participants: core event questions, transportation questions and planning questions. They ask respectively for details about the most important event of the trip; details about the method of travel to the destination; and details about how the trip was planned. Truth tellers and liars had time to prepare for the interview in this experiment. Liars tend to prepare answers to questions they expect to be asked (Vrij, Mann, Leal & Granhag, 2011) and hence, will often have a story ready when asked an expected question. Remembering and telling a rehearsed story is easier than thinking of a new one straight away, so preparing answers lowers cognitive load for liars. The task of remembering a story thought of in advance is rather similar to the task of a truth teller who is remembering the event. The liar can probably think of as much detail for a rehearsed story as a truth teller can remember from an event. If the liar is also using a strategy to give as much detail as possible, they may in fact create a story with more detail than a truth teller can remember or thinks is relevant to give in an interview. Therefore we hypothesised that liars would give more detail than truth tellers when answering the expected, general questions (Hypothesis 1) but fewer details than truth tellers about the unexpected core event, transportation and planning questions;
because the liars cannot prepare answers for questions that they did not expect to be asked (Hypothesis 2). Half of the truth tellers and half of the liars had previously been to the place they truthfully or deceptively claimed to be going to, while the other half had never been there before. It was hypothesised that participants who have made the trip before will give more details than those who have not made this trip before (Hypothesis 3). Since both truth tellers and liars should benefit from having previous experience, we expect this effect to occur regardless of whether someone is telling the truth or lying about their intention.

**Method.**

**Participants.**

Eighty-six participants were interviewed. Of these 65 were female, 20 male and one person did not state their gender on their questionnaire. They were on average 27.59 years old (SD = 12.34 years). Fifty-seven participants were undergraduate students (in psychology, science and humanities); the other 29 were university employees (in administration, management, support services etc.).

**Design.**

We used a 2 (Veracity) X 2 (Experience) X 4 (Question) mixed design, with the first two factors being between-subjects factors. Veracity consists of truth tellers and liars. Experience as a variable comprises those who had made the intended trip before versus those who had not. The number of participants with and without experience was equal across truth tellers and liars. Question is a within-subjects factor; it concerns the categories of question that were asked. There are four categories of question: general; core event; transportation and planning; they are always asked in that order. General questions are about the basics (e.g. “What is the
main purpose of your trip?”). Core event questions are about the details of the trip, the participants were asked to keep in mind one image of the most important thing they will do and were asked questions about that image, such as “Please tell me in as much detail as possible what you imagine you’re going to do in this picture?” The third question category, transportation, relates to the trip itself e.g. “How are you going to travel to your destination?” The fourth category, the planning, is about how they planned this trip, e.g.: “What part of the trip was easiest to plan?”

There are 17 dependent variables that were based on the interview. These include nine types of detail:

- **Visual details**  
  Information on what the interviewee sees

- **Auditory details**  
  Information on what the interviewee hears

- **Olfactory details**  
  Information on what the interviewee smells

- **Taste details**  
  Information on what the interviewee tastes

- **Tactile details**  
  Information on what the interviewee feels physically

- **Emotion details**  
  Information on what the interviewee feels psychologically

- **Action details**  
  Information on what the interviewee does

- **Spatial details**  
  Information on what is where

- **Temporal details**  
  Information on when the event and things within happen

There are seven other interview based dependent variables:

- **Spontaneous correction**  
  The interviewee correct him/herself

- **Hedging**  
  Interviewee indicates insecurity on the information e.g. “I’m not sure” or “probably”

- **Complication**  
  An unexpected complexity in the story, e.g. a trip has more than one purpose

- **Explanation**  
  The interview explains why certain things happened

- **Contradiction**  
  The interviewee contradicts something they said earlier in the statement

- **Quote**  
  The interviewee directly quotes another or him/herself
Hope

The interviewee states that they hope something will happen.

The 17th dependent variable is ‘total details’, it was created by adding these 16 variables together. Any detail given by the interviewee can only be placed into one specific category, but also counts towards the total details.

There were two dependent variables based on the post interview questionnaire: motivation and preparation. For motivation, the participants indicated how motivated they were to convince the interviewer on a seven point scale ranging from [1] not motivated at all to [7] extremely motivated. For preparation the participants indicated how much they prepared for the interview on a seven point scale ranging from [1] not prepared at all to [7] extremely prepared.

Procedure.

The participants were recruited via an online participant pool and advertisement. They were asked to participate in a study on lie detection in which they would be interviewed about a trip they planned to make. The advertisement said that they might be asked to tell the truth or to lie in the interview. After a participant signed up to the study a questionnaire was sent to them via e-mail. The questionnaire asked whether they had any travel plans in the near future and if so where, when and why they were going. It also asked them to list all countries and major cities they had been to in the past five years. This information was used to categorise the participants. Those who were planning to make a trip to a location they had not been to before were placed in the truth tellers without experience condition, whereas those who planned a trip to revisit a location they had been to previously and for the same reason, were placed in the truth tellers with experience condition. Those who had no
trip planned received instructions to lie. They were given a fake destination and a fake reason for their trip. When they were given their fake destination and reason, they were asked whether they had been there before to double-check. These fake destinations and reasons matched the destinations and reasons given by the truth tellers. Those who had been to the fake destination and were there for the same reason were placed in the ‘liars with experience’ condition and the others were placed in the ‘liars without experience’ condition. Participants who had been to a false destination but for a different reason were given a different false destination to avoid complications in the design. The matching was carried out to prevent a difference between the two groups in destinations and reasons for the trip and to control for any effect of the destination and the reason for the trip.

Those who were in the truthful conditions were given the following instructions: “In the interview I want you to answer the questions truthfully. Some people are asked to lie during the interview. The interviewer knows that some people may lie, but doesn’t know whether you are telling the truth or lying. Your goal is to convince the interviewer that you really are telling the truth.” The liars’ instructions were: “In the interview I want you to lie and pretend that you are travelling to [matched destination] and that you are going there for the purpose of [matched reason]. The interviewer knows that some people may lie, but does not know whether you are lying or telling the truth. Your goal is to convince the interviewer that you are telling the truth.” Both truth tellers and liars had at least three days in between receiving these instructions and the interview. They were told that they could prepare for the interview in any way they considered appropriate.

When the participants arrived for the interview they were met by the experimenter who asked them to sign an informed consent form. The experimenter
also verified whether they understood the instructions (all participants did). They were then brought to the interview room and interviewed by a female interviewer. The interview was audio and videotaped. The participants could see the camera and audio recorder; they were asked to try to ignore them and act as naturally as possible. The interviewer had the list of questions in front of her, and asked them all in the same order. The interviewer always started with the general questions, because in an interview general factual questions are typically asked first (Fisher, 2010). Second, it makes more sense to start an interview with expected questions followed by unexpected questions than vice versa. Starting an interview with unexpected questions is most likely to be perceived as odd. After the interview the participants filled out a questionnaire that asked for their demographics and motivation. They were also asked how much they had prepared and whether they had used strategies to appear convincing. Those who said they had prepared themselves were asked to outline how they had done so and what their strategy was. After completing the questionnaires, the participants were debriefed and given some sweets or a credit for the participant pool in exchange for their participation.

**Data analysis.**

The interviews were transcribed and coded. The coder, blind to the veracity status of the transcript, counted and classified each detail the participant mentioned within each question. They were then added up within each question category, creating variables that indicated how many of each type of detail were present in each questions category. The other interview based variables were counted and classified as well, per question category. A second rater coded ten percent of the sample. The second rater was also blind to the veracity status of the transcripts. The first coder showed the second coder examples of each type of detail and each other
cue. The second coder was given one interview to practice on, which was then compared with the first rater’s coding. Based on this comparison the second coder received some further explanation about how each detail should be classified. After this, the second coder rated the other interviews independently. The correlation between the two coders was high for most variables (see Table 7.1). Total details, Visual, Auditory, Olfactory, Action, Spatial, Temporal, Hedging, Explanation and Quote, all had correlations of over 0.85, a strong effect. Emotion and Hope had correlations of between 0.6 and 0.8, medium effects. Only Taste, Tactile and Correction had low correlations, under 0.5. For Complications and Contradictions the correlation could not be calculated, because they did not occur often enough.
Table 7.1: *The correlation between the first and the second coder.*

<table>
<thead>
<tr>
<th>Variable</th>
<th>$r$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total details</td>
<td>0.96</td>
<td>0.001</td>
</tr>
<tr>
<td>Visual details</td>
<td>0.98</td>
<td>$&lt; 0.001$</td>
</tr>
<tr>
<td>Auditory details</td>
<td>0.90</td>
<td>0.006</td>
</tr>
<tr>
<td>Olfactory details</td>
<td>0.90</td>
<td>0.002</td>
</tr>
<tr>
<td>Taste details</td>
<td>0.39</td>
<td>0.34</td>
</tr>
<tr>
<td>Tactile details</td>
<td>0.44</td>
<td>0.28</td>
</tr>
<tr>
<td>Emotion details</td>
<td>0.79</td>
<td>0.02</td>
</tr>
<tr>
<td>Action details</td>
<td>0.98</td>
<td>$&lt; 0.001$</td>
</tr>
<tr>
<td>Spatial details</td>
<td>0.88</td>
<td>0.01</td>
</tr>
<tr>
<td>Temporal details</td>
<td>0.99</td>
<td>$&lt; 0.001$</td>
</tr>
<tr>
<td>Spontaneous correction</td>
<td>0.37</td>
<td>0.48</td>
</tr>
<tr>
<td>Hedging</td>
<td>0.98</td>
<td>0.001</td>
</tr>
<tr>
<td>Complication</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Explanation</td>
<td>0.97</td>
<td>0.001</td>
</tr>
<tr>
<td>Contradiction</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Quote</td>
<td>0.92</td>
<td>0.01</td>
</tr>
<tr>
<td>Hope</td>
<td>0.78</td>
<td>0.67</td>
</tr>
</tbody>
</table>

**Results.**

*Expectedness of the questions.*

A pilot study was conducted to test which type of question was most expected. Forty-four participants who did not participate in the main experiment
were told about the interview that had been conducted in the main experiment, and were asked to list which questions they believed would be asked in the interview. For each question that was asked during the main experiment, it was noted how many participants wrote it on their list of expected questions. To be counted as the same, the actual question used in the interview, and the question that participants had written down, had to ask for exactly the same information although obviously the wording did not need to be exactly the same, e.g. “Where are you going?” would be seen as the same as “Where are you travelling to?” It was found that there was a significant difference between the general, core event, transportation and planning categories ($F(3, 22) = 4.78, p = 0.01$ $\eta^2 = .44$). Questions from the general questions category were listed on average by the highest number of participants: 17.5 (SD = 3.97), indicating that on average a question in the general question category was listed by 17.50 of the 44 participants. Questions from the other categories were expected less (core event: $M = .33$, SD = 3.24; transportation: $M = 4.50$, SD = 3.24; planning: $M < 0.001$ SD = 3.24), as was hypothesised.

**Questionnaire results.**

The participants rated themselves as highly motivated. On a seven point scale, from [1] not motivated at all to [7] extremely motivated, they gave themselves a mean score of 5.83 (SD = 1.17) for motivation, with no difference between truth tellers and liars, $t(84) = 1.36, p = 0.18$.

An ANOVA with Veracity and Experience as independent variables was conducted on the preparation variable. Liars prepared more than truth tellers (truth tellers: $M = 1.98$, SD = 1.00; liars: $M = 3.07$, SD = 0.95; $F(1, 86) = 27.63, p = 0.00001$ $\eta^2 = .25$). Participants without experience prepared more than
participants with experience (with experience: \( M = 2.27, SD = 1.04 \); without experience: \( M = 2.81, SD = 1.13 \); \( F (1, 86) = 6.28, p = 0.01 \), \( \text{Partial } \eta^2 = 0.07 \)). There was no significant Veracity by Experience interaction effect \( (F (1, 86) = 0.26, p = 0.61 \)).

**Interview content.**

A 4 (Question) X 2 (Veracity) X 2 (Experience) mixed repeated measures MANOVA was conducted including all 17 dependent variables. The multivariate analysis showed a significant Question * Veracity * Experience effect: \( F (51, 696) = 1.50, p = .02 \), \( \text{Partial } \eta^2 = .09 \). The Question * Veracity effect was also significant \( (F (51, 696) = 1.42, p = .03 \), \( \text{Partial } \eta^2 = .09 \)), as was the Question main effect \( (F (51, 696) = 21.89, p < .001 \), \( \text{Partial } \eta^2 = .62 \)). However, the Question * Experience effect \( (F (51, 696) = .63, p = .17 \), \( \text{Partial } \eta^2 = .08 \)), the Veracity* Experience effect \( (F (17, 66) = 1.68, p = .07 \), \( \text{Partial } \eta^2 = .30 \)) and the Veracity and Experience main effects were non-significant (Veracity: \( F (17, 66) = 0.99, p = .48 \), \( \text{Partial } \eta^2 = .20 \); Experience \( F (17, 66) = 1.62, p = .08 \), \( \text{Partial } \eta^2 = .30 \)).
Table 7.2: The significant multivariate effects for those variables with a significant three way interaction effect.

<table>
<thead>
<tr>
<th></th>
<th>F</th>
<th>df</th>
<th>p</th>
<th>Partial $\eta^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total details</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question * Veracity *</td>
<td>3.58</td>
<td>3, 246</td>
<td>0.02</td>
<td>0.04</td>
</tr>
<tr>
<td>Experience</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question*Experience</td>
<td>40.06</td>
<td>3, 246</td>
<td>0.01</td>
<td>0.05</td>
</tr>
<tr>
<td>Question</td>
<td>111.43</td>
<td>3, 246</td>
<td>&lt;0.001</td>
<td>0.57</td>
</tr>
<tr>
<td><strong>Visual details</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question * Veracity *</td>
<td>3.77</td>
<td>3, 246</td>
<td>0.01</td>
<td>0.04</td>
</tr>
<tr>
<td>Experience</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question*Experience</td>
<td>3.28</td>
<td>3, 246</td>
<td>0.03</td>
<td>0.04</td>
</tr>
<tr>
<td>Question</td>
<td>92.41</td>
<td>3, 246</td>
<td>&lt;0.001</td>
<td>0.5</td>
</tr>
<tr>
<td><strong>Spatial details</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question * Veracity *</td>
<td>3.74</td>
<td>3, 246</td>
<td>0.03</td>
<td>0.04</td>
</tr>
<tr>
<td>Experience</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question*Experience</td>
<td>4.03</td>
<td>3, 246</td>
<td>0.02</td>
<td>0.05</td>
</tr>
<tr>
<td>Question</td>
<td>41.03</td>
<td>3, 246</td>
<td>&lt;0.001</td>
<td>0.33</td>
</tr>
<tr>
<td><strong>Temporal details</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question * Veracity *</td>
<td>2.85</td>
<td>3, 246</td>
<td>0.047</td>
<td>0.03</td>
</tr>
<tr>
<td>Experience</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question*Experience</td>
<td>5.45</td>
<td>3, 246</td>
<td>0.001</td>
<td>0.06</td>
</tr>
<tr>
<td>Question</td>
<td>60.71</td>
<td>3, 246</td>
<td>&lt;0.001</td>
<td>0.43</td>
</tr>
<tr>
<td><strong>Explanations</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question * Veracity *</td>
<td>4.57</td>
<td>3, 246</td>
<td>0.004</td>
<td>0.05</td>
</tr>
<tr>
<td>Experience</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question*Experience</td>
<td>1.12</td>
<td>3, 246</td>
<td>0.34</td>
<td>0.01</td>
</tr>
<tr>
<td>Question</td>
<td>32.09</td>
<td>3, 246</td>
<td>&lt;0.001</td>
<td>0.28</td>
</tr>
</tbody>
</table>

This section will only describe the univariate effects that are significant at multivariate level: the Question * Veracity * Experience effect, the Question * Veracity effect and the Question main effect. First, for total details, the three way Question * Veracity * Experience interaction effect was significant. The Question *Veracity effect and the Question main effect were also significant (see Table 7.2). The analysis reveals that liars mentioned significantly more details in the general section than truth tellers ($t(84) = 2.49, p = 0.02$; Truth $M = 15.9$, SD = 11.44; Lie $M$...
= 26.02, SD = 23.9), as was predicted in Hypothesis 1. In addition, liars reported fewer details in the other three sections than truth tellers, as was predicted in Hypothesis 2. The difference between liars and truth tellers was significant for the transportation section (one-sided \( t (84) = 1.94, p = 0.03 \); Truth M = 100.86, SD = 72.85; Lie M = 74.16, SD = 53.57), but not significant for the core event and planning sections \( (t < 1.39, \text{all } p > 0.17) \). This Question * Veracity effect was stronger for those with experience than for those without experience. For those without experience there was no significant difference between truth tellers and liars in any of the question categories \( (\text{all } t < 1.46, p > 0.15) \). For those with experience, liars gave significantly more details than truth tellers in the general questions \( (t (84) = 2.04, p = 0.05) \), but gave fewer details in the three unexpected questions categories. This was only significant in the transport and planning sections \( (\text{core event: } t (84) = 1.44, p = 0.16; \text{transport: } t (84) = 2.32, p = 0.03; \text{planning: } t (84) = 2.26, p = 0.03) \), see Table 7.3 for the means and standard deviations.

Table 7.3: The mean and (SD) of the amount of total details per questions category, experience level and veracity

<table>
<thead>
<tr>
<th>Experience level and veracity</th>
<th>Truth, not experienced</th>
<th>Truth, experienced</th>
<th>Lie, not experienced</th>
<th>Lie, experienced</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>16.1 (14.91)\textsuperscript{a}</td>
<td>15.73 (7.39)\textsuperscript{a}</td>
<td>22.46 (13.35)\textsuperscript{a}</td>
<td>29.59 (31.06)\textsuperscript{a}</td>
</tr>
<tr>
<td>Core event</td>
<td>96.2 (51.13)\textsuperscript{b}</td>
<td>128.32 (78.92)\textsuperscript{b}</td>
<td>117.82 (73.66)\textsuperscript{b}</td>
<td>98.64 (56.13)\textsuperscript{b}</td>
</tr>
<tr>
<td>Travel</td>
<td>78.15 (37.49)\textsuperscript{c}</td>
<td>121.5 (90.29)\textsuperscript{b}</td>
<td>78.59 (54.92)\textsuperscript{c}</td>
<td>69.73 (53.09)\textsuperscript{c}</td>
</tr>
<tr>
<td>Planning</td>
<td>92.75 (35.16)\textsuperscript{b}</td>
<td>122.77 (70.01)\textsuperscript{b}</td>
<td>100.59 (60.68)\textsuperscript{b}</td>
<td>83.46 (41.79)\textsuperscript{bc}</td>
</tr>
</tbody>
</table>

\textit{N.B. Within columns, cells that do not share any superscript letter differ significantly.}

The results for visual details show the same pattern. The Question * Veracity * Experience effect, the Question * Veracity effect and the Question main effect
were all significant (see Table 7.2). Similar to the total details, liars gave significantly more visual details than truth tellers in answer to the general questions \((t (84) = 2.54, p = 0.01; \text{Truth M} = 8.14, \text{SD} = 6.12; \text{Lie M} = 13.09, \text{SD} = 11.28)\). In the core event, transport and planning categories, there were no significant differences between truth tellers and liars \((t < 1.60, \text{all } p > 0.11)\). Similar to the results for the total details, for visual detail, this Question * Veracity interaction effect was stronger in those with experience than in those without experience. For those without experience there was no significant difference between truth tellers and liars in any of the question categories \((t < 1.41, \text{all } p > 0.16)\). For those with experience the liars gave more details in the general questions category \((t (84) = 2.20, p = 0.03)\), and fewer details in the other three categories. This difference was significant in the transport category \((t (84) = 2.09, p = 0.04)\), marginally significant in the planning category \((t (84) = 2.00, p = 0.05)\) and not significant for the core events category \((t (84) < 2.00, p > 0.05)\), see Table 7.4 for the means and standard deviations.

Table 7.4: The mean and (SD) of the amount of visual details per questions category, experience level and veracity.

<table>
<thead>
<tr>
<th>Category</th>
<th>Truth, not experienced</th>
<th>Truth, experienced</th>
<th>Lie, not experienced</th>
<th>Lie, experienced</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>8.25 (7.91)^a</td>
<td>8.05 (4.07)^a</td>
<td>11.14 (6.72)^a</td>
<td>15.05 (14.40)^a</td>
</tr>
<tr>
<td>Core event</td>
<td>44.95 (25.28)^b</td>
<td>60.68 (38.41)^b</td>
<td>59.09 (37.93)^b</td>
<td>47.73 (31.26)^b</td>
</tr>
<tr>
<td>Travel</td>
<td>31.45 (17.14)^c</td>
<td>54.82 (49.14)^b</td>
<td>34.82 (26.86)^c</td>
<td>29.56 (28.62)^c</td>
</tr>
<tr>
<td>Planning</td>
<td>43.6 (18.25)^b</td>
<td>60.27 (35.89)^b</td>
<td>48.32 (29.18)^b</td>
<td>41.96 (23.55)^b</td>
</tr>
</tbody>
</table>

\(N.B.\) Within columns, only cells that do not share any superscript letter differ significantly.

The results for spatial details showed a similar pattern to the patterns for total and visual details. The Question * Veracity * Experience effect, the Question *
Veracity effect and the Question main effect were significant (see Table 7.2). Liars mentioned significantly more details in the general section than truth tellers ($t$ (84) = 2.06, $p = 0.04$; Truth M = 2.26, SD = 1.99; Lie M = 3.39, SD = 3.01), as was predicted in Hypothesis 1. In addition, liars reported less detail in the other three sections than truth tellers, as was predicted in Hypothesis 2. The difference between liars and truth tellers was significant for the transportation section ($t$ (84) = 2.55, $p = 0.01$; Truth M = 13.55, SD = 13.18; Lie M = 7.84, SD = 6.65), but not significant for the core event and planning sections ($t < 0.62$, all $p > 0.54$). Similar to the results for the total and visual details, this Question * Veracity interaction effect was stronger in those with experience than those without experience. For those without experience there was no significant difference between truth tellers and liars in any of the question categories (all $t < 1.48$, all $p > 0.15$). For those with experience the liars gave more details in the general questions category (spatial: $t$ (84) = 2.12, $p = 0.04$), and fewer details in the other three categories. This difference was significant in the transport category (spatial: $t$ (84) = 2.53, $p = 0.02$), and not significant for the core events category and the planning category (both $t < 1.61$, all $p > 0.11$), see Table 7.5 for the means and standard deviations.
Table 7.5: The mean and (SD) of the amount of spatial details per questions category, experience level and veracity.

<table>
<thead>
<tr>
<th>Category</th>
<th>Truth, not experienced</th>
<th>Truth, experienced</th>
<th>Lie, not experienced</th>
<th>Lie, experienced</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>2.60 (2.39)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1.96 (1.53)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>3.18 (2.75)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>3.58 (3.29)&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Core event</td>
<td>10.60 (6.54)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>17.64 (19.84)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>14.77 (10.93)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>10.46 (6.47)&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Travel</td>
<td>10.25 (7.27)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>16.55 (16.47)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>8.68 (6.83)&lt;sup&gt;c&lt;/sup&gt;</td>
<td>7.00 (6.51)&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>Planning</td>
<td>7.45 (3.72)&lt;sup&gt;c&lt;/sup&gt;</td>
<td>9.27 (8.14)&lt;sup&gt;c&lt;/sup&gt;</td>
<td>8.36 (7.01)&lt;sup&gt;c&lt;/sup&gt;</td>
<td>6.86 (4.73)&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

N.B. Within columns, only cells that do not share any superscript letter differ significantly.

The results for temporal details showed significant Question * Veracity * Experience, Question * Veracity and Question effects (see Table 7.2). The estimated marginal means for the entire interview showed that truth tellers with experience (M = 9.35; SE = 0.88) gave the most temporal details overall, followed by liars without experience (M = 6.67; SE = 0.88), who interestingly gave more temporal details than truth tellers without experience (M = 5.69; SE = 0.92) and liars with experience (M = 5.67; SE = 0.88). Post hoc tests show that over the entire interview truth tellers with experience gave significantly more temporal details than truth tellers without experience and liars with experience (all p < 0.03), but not significantly more than liars without experience (p = 0.20). Truth tellers without experience, liars without experience and liars with experience did not differ significantly from each other (all p > 0.20). The Question * Veracity effect is similar to the effect found with total details, visual details and spatial details. Liars gave significantly more temporal details than truth tellers in answer to the general questions (t (84) = 3.30, p = 0.001). In the core event, transport and planning categories, the truth tellers gave more details than liars. The difference was significant for the transport and planning...
categories, but not for the core event questions (core event: $t = 0.363$, all $p = 0.72$; transport: $t = 2.12$, all $p = 0.03$; planning: $t = 2.20$, all $p = 0.03$). The mean scores and standard deviations per questions category are listed in Table 7.6. Again, this Question * Veracity effect was stronger for those with experience then for those without experience. For those without experience, liars gave significantly more temporal detail then truth tellers ($t (84) = 2.91, p = 0.01$), there was no significant difference between truth tellers and liars in the other three questions categories (all $t < 1.01, p > 0.28$). For those with experience, liars gave marginally significantly more detail than truth tellers in the general questions ($t (84) = 1.89, p = 0.07$), but fewer details in the three unexpected questions categories. This was only significant in the transport and planning sections (core event: $t (84) = 1.61, p = 0.12$; transport: $t (84) = 2.54, p = 0.02$; planning: $t (84) = 3.06, p = 0.004$)

Table 7.6: The mean and (SD) of the amount of temporal details per questions category, experience level and veracity.

<table>
<thead>
<tr>
<th>Category</th>
<th>Truth, not experienced</th>
<th>Truth, experienced</th>
<th>Lie, not experienced</th>
<th>Lie, experienced</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>1.10 (1.37)$^a$</td>
<td>1.46 (1.36)$^a$</td>
<td>2.86 (2.38)$^ab$</td>
<td>2.86 (3.23)$^a$</td>
</tr>
<tr>
<td>Core event</td>
<td>3.80 (4.56)$^b$</td>
<td>7.18 (5.56)$^b$</td>
<td>5.59 (5.91)$^{ab}$</td>
<td>4.73 (4.51)$^{ab}$</td>
</tr>
<tr>
<td>Travel</td>
<td>7.60 (4.96)$^c$</td>
<td>12.46 (8.88)$^c$</td>
<td>7.09 (6.20)$^b$</td>
<td>6.96 (5.61)$^{bc}$</td>
</tr>
<tr>
<td>Planning</td>
<td>10.25 (4.70)$^c$</td>
<td>16.14 (11.58)$^c$</td>
<td>11.14 (8.27)$^c$</td>
<td>8.14 (4.07)$^c$</td>
</tr>
</tbody>
</table>

N.B. Within columns, only cells that do not share any superscript letter differ significantly.

The results for ‘explanations’ showed significant Question * Veracity * Experience and Question effects. The Question * Veracity effect was not significant (see Table 7.2). The mean scores and standard deviations are listed in Table 7.7. Explanations were most common in the planning category ($M = 3.61$, $SD = 0.29$),
second most common in the core events category (M = 2.75, SD = 0.35), followed by the transport category (M = 1.75, SD = 0.2) and were least common in the general questions category (M = 0.72, SD = 0.19). There were no significant differences between truth tellers and liars in any of the question categories, whether experienced or not (all $t < 1.92, p > 0.06$).

Table 7.7: The mean and (SD) of the amount of explanations per questions category, veracity and experience level.

<table>
<thead>
<tr>
<th></th>
<th>Truth, not experienced</th>
<th>Truth, experienced</th>
<th>Lie, not experienced</th>
<th>Lie, experienced</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>.55 (1)$^a$</td>
<td>.32 (.57)$^a$</td>
<td>.5 (.74)$^a$</td>
<td>1.5 (3.23)$^a$</td>
</tr>
<tr>
<td>Core event</td>
<td>1.95 (3.19)$^{ab}$</td>
<td>3.05 (3.03)$^{ab}$</td>
<td>4 (4.21)$^b$</td>
<td>2 (2.39)$^a$</td>
</tr>
<tr>
<td>Travel</td>
<td>1.5 (1.47)$^{ab}$</td>
<td>2.23 (2.11)$^{ab}$</td>
<td>1.36 (1.62)$^a$</td>
<td>1.91 (2.09)$^a$</td>
</tr>
<tr>
<td>Planning</td>
<td>3.5 (1.85)$^b$</td>
<td>4.05 (3.05)$^b$</td>
<td>4.09 (3.31)$^b$</td>
<td>2.82 (2.32)$^a$</td>
</tr>
</tbody>
</table>

N.B. Within columns, only cells that do not share any superscript letter differ significantly.

The remaining 12 dependent variables did not have significant univariate three way interaction effects (see Table 7.8). Since non-significance of an interaction effect usually precludes analysis of the lower order effects, they are not further discussed.
Table 7.8: The three way interaction effects for the 12 variables where this effect was non-significant.

<table>
<thead>
<tr>
<th>Variable</th>
<th>F</th>
<th>DF</th>
<th>P</th>
<th>Partial $\eta^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auditory details</td>
<td>2.26</td>
<td>3, 246</td>
<td>.08</td>
<td>.03</td>
</tr>
<tr>
<td>Olfactory details</td>
<td>1.83</td>
<td>3, 246</td>
<td>.14</td>
<td>.02</td>
</tr>
<tr>
<td>Taste details</td>
<td>.14</td>
<td>3, 246</td>
<td>.94</td>
<td>.002</td>
</tr>
<tr>
<td>Tactile details</td>
<td>.12</td>
<td>3, 246</td>
<td>.95</td>
<td>.001</td>
</tr>
<tr>
<td>Emotion details</td>
<td>.41</td>
<td>3, 246</td>
<td>.75</td>
<td>.01</td>
</tr>
<tr>
<td>Action details</td>
<td>2.05</td>
<td>3, 246</td>
<td>.11</td>
<td>.02</td>
</tr>
<tr>
<td>Spontaneous corrections</td>
<td>.88</td>
<td>3, 246</td>
<td>.45</td>
<td>.01</td>
</tr>
<tr>
<td>Hedging</td>
<td>1.21</td>
<td>3, 246</td>
<td>.31</td>
<td>.02</td>
</tr>
<tr>
<td>Complication</td>
<td>.8</td>
<td>3, 246</td>
<td>.5</td>
<td>.01</td>
</tr>
<tr>
<td>Contradiction</td>
<td>.87</td>
<td>3, 246</td>
<td>.46</td>
<td>.01</td>
</tr>
<tr>
<td>Quote</td>
<td>1.97</td>
<td>3, 246</td>
<td>.12</td>
<td>.02</td>
</tr>
<tr>
<td>Hope</td>
<td>.79</td>
<td>3, 246</td>
<td>.5</td>
<td>.01</td>
</tr>
</tbody>
</table>

As reported, it was found that liars gave more detail (total detail) than truth tellers in the general questions category (which, for both truth tellers and liars, was the question category in which the least amount of detail was given, see Tables 7.3 to 7.7), whereas liars gave fewer details than truth tellers in the other categories. This implies that the ratio between amount of total detail in the general category and the other categories was larger for liars than for truth tellers. To assess whether this difference could be used as a cue to deception a ratio variable was created. First the amount of total detail given in the core event, transportation and planning categories was averaged for each participant. This number was divided by the amount of total details the participant gave in the general questions category. Truth tellers gave more details in the unexpected categories for each detail they gave in the expected category than liars (truth tellers ratio: $M = 8.89$, $SD = .89$; liars ratio: $M = 5.13$, $SD = .87$; $t (84) = 3.08$, $p = 0.003$). A Roc analysis was carried out for this ratio variable. It showed that the analysis could classify truth tellers and liars significantly above the
level of chance (AUC = .75, SE = 0.05, p = 0.01, 95% confidence interval .64-.85).
The best cut off value was at a ratio of 5.73 details in unexpected categories per one
detail in the expected, at that level the positive predictive value was .69, and the
negative predictive value was .75. Therefore, 69% of truth tellers and 75% of liars
could be classified correctly.

Discussion.

The results showed a significant three way interaction effect for five of the
dependent variables: total details, visual details, spatial details, temporal details and
explanations. For the other 12 variables no significant effects were found. For most
of these cues, it is possible that no effect was found because the cue was so
uncommon that there was not enough variance in the data to find differences.
Olfactory, taste, tactile and emotion details, as well as corrections, complications,
contradictions, quotes and hope have estimated marginal means of three even in the
condition where they are most common. Auditory details and hedges had estimated
marginal means of around five. The only non-significant variable which was as
common as the significant variables was action details. Perhaps these details are
easier to make up for a liar, because they do not require that much imagination.
When travelling, certain actions always have to be taken, to leave, take transport,
arrive etc. The somewhat scripted nature of these actions may help set up a false
narrative with similar amounts of action details as truth tellers, even when less
contextual details about the trip are given.

For total details, visual details, spatial details and temporal details, a
significant Veracity X Questions interaction effect was also found, which showed
that liars and truth tellers responded differently to the different question categories.
For total details, visual details and spatial details liars mentioned significantly more
details in total than truth tellers in the general section, supporting Hypothesis 1. This opposes previous research which has found that liars mention fewer details than truth tellers (DePaulo et.al., 2003; Vrij, 2008). A difference between that previous research and our present experiment is that we encouraged the participants to prepare themselves and gave them a few days to do so. Our results showed that liars took the opportunity to prepare more than truth tellers. Possibly, in an effort to be convincing, liars may have been eager to share that prepared story and did so in response to the first question that could possibly allow it. That first opportunity was likely to arise in the general questions sections, because these are the first questions in the interview and interviewees expect to be asked these questions, as the pilot study revealed. Of course, it makes sense to give the prepared response at the first opportunity. Interviewees cannot know whether the later questions would ask for the details that they had prepared. They may have felt that if they did not give their prepared story at the first opportunity, they may not have the opportunity to do so later. Therefore an empirical question worth examination is what would happen if unexpected questions are asked first? Liars would then face the dilemma of having a prepared story but no appropriate questions (yet) in order to give their prepared answers. They have two options. They may wait until an expected question is asked; thereby risking the possibility that such opportunity will not arise. Alternatively, they may give their prepared answer to an unexpected question that does not really ask for the information they have prepared, risking the possibility that their answer will come across as odd or irrelevant. This could be tested by replicating this experiment but counterbalancing the order in which the different categories of question are asked.

More in line with previous research was that liars mentioned fewer details than truth tellers in the other three sections of the interview (DePaulo et.al., 2003;
Vrij, 2008). However, these differences were only significant in the transportation category and, for some variables, in the planning category, whereas previous research has often found significant differences also with questions that resemble the core event questions. One possible explanation is that the substantial preparation time we gave the liars helped them to include enough detail in their answers in the core event section to negate the differences with truth tellers. However, the problem with this explanation is that questions about the core event were not really expected, making it unlikely that the liars had prepared answers for them. Alternatively it may be because the participants told the truth or lied about intentions rather than past activities. Previous experiments regarding lying about intentions did not find differences in detail between truth tellers and liars either (Vrij, Granhag, Mann, & Leal, 2011; Vrij, Leal, Mann, & Granhag, 2010). Perhaps descriptions of truthful intentions are generally not as detailed as truthful stories about past activities. This is supported by research by D’Argembeau and Van der Linden (2004) and Gamboz, Brandimonte and De Vito (2010), who found that images of past events were more detailed than images of future events. When truth tellers discuss their past activities, there is a wealth of perceptual details that they have experienced, that they can recall, if they still remember them. In contrast, when discussing their intentions about a forthcoming trip, truth tellers have not yet experienced enough which restricts the amount of detail they can report about their intentions. As a result, truth tellers are not so detailed when discussing their intentions, which would make it easier for liars to provide the same amount of detail as truth tellers when discussing them.

The results show that the differences between truth tellers and liars on the different questions only occurred in the condition where the participants have previous experience. This implies that having experience helps liars to generate more
details to the expected questions, which suggests that they use their previous experience in their preparation for the interview. Having experience helps truth tellers to give more detail on the unexpected questions, specifically the travel and planning sections. This suggests that their previous experience gives them a more vivid and detailed image of the future. This is in line with the findings of Szpunar and McDermott (2008) who found that future events in familiar settings are clearer and more detailed than future events in unfamiliar settings. This result also suggests that knowing whether someone who is reporting on their intentions has experienced the future event they are describing previously can be beneficial for lie detection. Having experience affects how liars and truth tellers respond on expected and unexpected questions and this can be used as a cue to deception.

Hypothesis 3, which stated that regardless of veracity condition those with experience would give more detail than those without, was not supported. An explanation for this perhaps lies in the finding that those with experience prepared themselves less than those without experience. Probably, those with experience believed that using their previous experience would suffice and hence felt that no other preparation was needed. It may be that this lack of preparation by those with experience caused the absence of a main experience effect. For temporal details there was an interaction between experience and veracity. Accurate timing of travelling is often important, after all, planes and trains leave on time and can be missed. It is possible that truth tellers with experience have deliberately remembered and memorised the times from their last trip, in the knowledge that they would have to memorise them anyway for their upcoming trip. Truth tellers without experience may wait to memorise timings until they their trip is more imminent, fearing they may forget these times otherwise in the meantime. Liars may avoid mentioning
times, because they are usually checkable and they do not know how much the interviewer knows about the time journeys may take.

Whether someone can use their previous experience may depend on how often they have experienced events that are relevant, how long ago they experienced these events and how well they remember them. In other words, there are many possible factors that could influence whether or not an interviewee can benefit from their previous experience that have not yet been explored. This may be worth examining in future research. It may also be that interviewees used their previous experience to their benefit in a way that the present experiment did not examine. For example, perhaps those with experience told a more plausible story or controlled their non-verbal behaviour more, two aspects not examined in the present experiment.

The ratio between the amount of details mentioned in the expected general questions and the unexpected question categories has been shown to be a cue to deception. A Roc analysis showed that the ideal cut off point was 5.73 details in unexpected categories to 1 detail in the expected category. This cut off had a positive predictive value of .69 and a negative predictive value of .75. These predictive values are significantly better than chance. It is also higher than the 52.3% of truth tellers and 62.6% of liars who were correctly classified based on nonverbal behaviour or the 64% to 68% of truth tellers and 62% to 68% of liars who were correctly classified based on thermal imaging. The accuracy rates are, however, lower than the rates that would be necessary to make the technique useful in practical situations. Also since the interview is still quite long it would be impossible to use it in transport settings where high numbers of innocent people need to be passed through quickly.
In conclusion, this experiment shows that there are differences between truth tellers and liars in the amount of detail they give in an interview. Liars give more details than truth tellers to expected questions, but truth tellers give more details to unexpected questions. This effect is stronger in those participants who have experienced the trip they are describing. This suggests that although liars can be experienced and well prepared, they can still not give the same amount of detail a truth tellers can to unexpected questions and this is a cue to their mendacity. The next chapter will look at the verbal cues to deception in a much shorter interview.
Chapter 8: Experiment 3. The Effect of Time Prompting

In Experiment 3 we investigated verbal cues truth tellers and liars give in a very short, one question interview. Such short interviews may be very helpful in practical situations where clearing large numbers of innocent people in as little time as possible is of great importance, e.g. in airports (Burgoon et al., 2009). Very short answers are unlikely to elicit a large amount of cues and many cues that discriminate between liars and truth tellers in longer interviews will probably be absent completely. The best way to analyse the data may be to focus on a single cue that is common enough to help classify truth tellers and liars on its own. Focussing on one type of detail should also make coding easier and this would further reduce the amount of time necessary for the interview and the classification. Experiment 2 revealed that the questions category by veracity effect was largest for temporal details, indicating that the difference in the amount of temporal details given by truth tellers and liars was larger than for other details. This suggests that differences in temporal details might be more likely to be found even in very short interviews. Therefore, Experiment 3 measures the level of detail of the statement by looking at temporal information given. Experiment 2 showed that truth tellers give more temporal information overall than liars, although the difference was not significant. Other studies have found that truth tellers give more temporal information than liars (e.g. Vrij et al., 2009), although in that case the participants were interviewed on past events rather than intentions. Johnson et al. (1988) found that people mentioned more temporal details when relating perceived events compared to imagined events.

In this experiment we try to manipulate the amount of temporal information that is given in such a way that the difference between truth tellers and liars gets larger, making it easier to classify them. The way to do this would be to ask
questions that encourage truth tellers more than liars to provide temporal information. Questions that explicitly ask participants to give as much temporal information as possible are unlikely to achieve this as they will encourage liars as well as truth tellers to give temporal information. Asking such a question would be a large prompt to liars that a ‘give a lot of detail’ strategy would be successful. However, questions that refer to time in a subtle manner may work. Since truth tellers probably already have a more temporally detailed image of their intention, they may be more sensitive to a subtle time prompt. The liars might respond also to the subtle time prompt, but with less temporal information in the image of their intention, they may give less extra information than truth tellers would.

Experiment 2 was conducted in the lab. Although participants were discussing their own intentions rather than intentions induced by the experimenter, this is still a rather artificial set up. The participants were in a strange place, that was unrelated to their intention and the intentions they spoke about could be weeks or months later. They also knew about the interview several days beforehand. In order to make the experimental situation resemble the real life situation of interviews at border control or for transport security more, Experiment 3 was conducted on the Portsmouth-Caen ferry. The participants were interviewed about their intentions for the day after they disembarked. Half were asked to tell the truth, the other half to lie. Half of each group was asked a subtle time prompt question; the other half of each group was asked a control question.

Since Experiment 2 and previous research (e.g. Vrij et al., 2009) suggest that truth tellers give more temporal information than liars, we hypothesised that those who describe a truthful intention would include temporal detail more often in their statement than those who discuss an intention they are not planning to execute
(Hypothesis 1). It was also hypothesised that truth tellers would respond more strongly to a subtle time prompt than liars, so truth tellers will give more detail when prompted, while liars will give no more or only very little more detail when prompted. This will make the classification of truth tellers and liars better in the time prompt condition, because truth tellers have more temporal information in their memory to use than liars, which they are triggered to give by the time prompt (Hypothesis 2).

Method.

Participants.

A total of 84 participants (36 males and 48 females) took part. Their average age was \( M = 58 \) years old (\( SD = 12.6 \)).

Design.

The design is 2 by 2 between-subjects factorial. The first factor Veracity consists of liars versus truth tellers. The second factor, Question, consists of time prompt versus control.

Procedure.

The experiment took place on the Portsmouth - Caen ferry. The experimenter approached passengers on the ferry and asked whether they would agree to be interviewed and audio taped for a study about lie detection. An estimated 75 % of the passengers who were approached agreed to take part. This percentage of volunteers is relatively high, possibly because most passengers on the ferry were not busy. If they agreed to participate they were asked to sign an informed consent form. The experimenter asked for their sex, age and what they were doing on the trip. The last
question was used to establish ground truth. The participants were randomly assigned to be a truth teller or a liar. The truth tellers ($N = 42$) were then given the following instructions: “An interviewer will come and ask you for details about your trip. Please answer truthfully. Try to convince the interviewer that you are telling the truth”. The liars ($N = 42$) were told: “An interviewer will come and ask you for details about your trip. Please lie to them and pretend you are making another trip than you in fact are going to make. Try to convince the interviewer that you are telling the truth.” Liars were not given any instruction what their lie should be. The participant was given time to prepare if they wished, none of them asked for more time than it took for the experimenter to bring the interviewer over and the interviewer to get set up, which was usually around a minute. The interviewer, who was blind to veracity condition, introduced him/herself, gave the participant the Dictaphone to speak into like a microphone and asked the passengers for their plans at their destination via a single question, using either the control question or time prompt question. The control question was: “Please describe in as much detail as possible what you are going to do today at your destination?” The time prompt question was: “Please describe in as much detail as possible what your time-table is for today at your destination?” The interviewer would listen attentively and allow the passenger to say as much as they wanted but was not allowed to ask follow-up questions. When the passenger had finished answering, the interviewer thanked the participant for taking part in the experiment and gave him/her a debriefing form and a bag of sweets. The entire procedure (from being approached by the interviewer to receiving the debriefing form and candy) lasted only a few minutes.
**Data analysis.**

The interviews were transcribed and coded by two raters. They read the transcript and marked the number of temporal details, the correlation between them was 0.94 ($p < 0.001$). Temporal information was rare, perhaps because the interviewees’ answers were very short in this experiment. In order to investigate whether more precise temporal information, (e.g. “quarter to three” or “two o’clock”) was more diagnostic, two coders also marked whether or not the participant used such a specific phrase. The raters agreed on 83 out of 84 transcripts. The transcript on which the raters disagreed was discussed and agreed upon.

**Results.**

The average duration of the entire interview was short, 46.77 seconds (SD = 17.93). This includes the question asked by the interviewer. The average number of words given by the participants was 33.69 (SD = 20.71). A MANOVA shows that there was no significant effect of Question or veracity on the temporal information ($F(3, 80) = 0.66, p = 0.58$). A binary logistic regression was conducted with Veracity and Questions as the covariates (using enter method) and Mentioning time as the dependent variable. The model had a $\chi^2(2)$ of 18.17, $p < 0.001$. The -2 log likelihood of the model was 63.63, $R^2$(Cox & Snell) = .2, $R^2$(Nagelkerke) = .31. The constant, Veracity and Question significantly affected whether or not a time was mentioned (Constant: $B = -4.02$, S.E. = .91, Wald = 19.46, $p < 0.001$, Exp(B) = 0.02; Veracity: $B = 1.54$, S.E. = .67, Wald = 5.24, $p = 0.02$, Exp(B) = 4.67; Question: $B = 2.45$, S.E. = .82, Wald = 8.97, $p = 0.003$, Exp(B) = 11.61). This means that when the participant was telling the truth or answering the time prompt question, the chances
that they mentioned a specific time were significantly higher (see Table 8.1 for percentages per condition).

Table 8.1: The percentage of participants from each condition that did or did not mention a time in their answers.

<table>
<thead>
<tr>
<th></th>
<th>Truth, time prompt</th>
<th>Truth, control</th>
<th>Lie, time prompt</th>
<th>Lie, control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time mentioned</td>
<td>11.9</td>
<td>2.38</td>
<td>4.76</td>
<td>0.00</td>
</tr>
<tr>
<td>No time mentioned</td>
<td>13.10</td>
<td>22.62</td>
<td>20.24</td>
<td>25.00</td>
</tr>
<tr>
<td>Overall</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
</tr>
</tbody>
</table>

Discriminant analyses were conducted to test how well mentioning a time could classify the participants as truth telling or lying in both Question conditions. The analysis showed that with the time prompt question 64.3% of the participants could be correctly classified ($\lambda = .91$, $\chi^2(1) = 3.81$, $p = 0.05$). The model did have a lie bias. Of the truth tellers only 47.6% were classified correctly, while 81% of liars were. Without the time prompt, 54.8% of participants were correctly classified, which was not significantly above chance level ($\lambda = .95$, $\chi^2(1) = 2.03$, $p = 0.16$). The model is now almost completely lie biased, 9.5% of truth tellers and 100% of liars are correctly classified.

Discussion.

The results show that the interview was indeed very short, three quarters of a minute on average. This time is not spent solely by the interviewee speaking, it includes the introduction and closure of the interview and the time the interviewer used to ask the question. This length of interview is much more practical in real life.
The results also showed that truth tellers mentioned times more than liars when describing their intentions, supporting Hypothesis 1. This is in line with previous research showing that truth tellers tend to include more temporal detail than liars when discussing their past activities (e.g. Vrij et al., 2009). Temporal information is therefore a cue to truthfulness in discussing both past and future activities. Participants answering the time prompt question were also more likely to mention times. The discriminant analyses showed that classification of truth tellers and liars was more accurate when they were asked the time prompting questions, supporting Hypothesis 2. The higher accuracy in the time prompting condition was caused by the higher occurrence of time mentioning. Providing temporal information spontaneously appears to be a rare event in this interview, with only 10 % of truth tellers and none of the liars doing this. This low spontaneous occurrence of temporal information gives investigators who are interested in temporal information as a cue to deceit little option other than to prompt for this information. The efficiency of the prompt may be higher if the prompt is made even more subtle, so that liars respond to it less than they did in this experiment. On the other hand a stronger time prompt may increase the classification rate because truth tellers are better at giving temporal information and can therefore respond more strongly to the time prompt. What phrasing of the question is most efficient from a lie detection view should be studied in further experiments. Further research might also look into whether these time prompts can be integrated in longer interviews.

The overall accuracy rate in the time prompt condition was 64%, with 47.6% of truth tellers and 81% of liars correctly classified and significantly above chance level. Without the time prompt the classification was not significantly above chance.
The accuracy rates with the time prompt were comparable to the accuracy rates found in the thermal imaging experiment, described in Chapter 5 and higher than those found when using non-verbal cues, described in Chapter 4. They are lower than the 69% and 75% found in Experiment 2. These rates are too low for use in practical situations but it does show that even one question interviews yield cues that can help classify truth tellers and liars at above significance rate, if that question is a suitable question.

In Part III of this thesis, verbal cues to deception have been examined. The accuracy rates in classifying truth tellers and liars who discuss their travel plans ranged from 72% in the long interviews (Experiment 2) to 64% with the short interviews (Experiment 3). These accuracy rates compared favourably to the 57% and 66% accuracy found by using non-verbal cues and thermal imaging, presented in Chapters Four and Five. However the accuracy rates were still lower than would be necessary for use in court, for which the accuracy rate would need to approach 100%. Both Parts II and III of this thesis have focussed on cues to deception that are elicited in interviews. Part IV will focus on truth tellers’ and liars’ reaction times to intention related stimuli to examine whether these might be a cue to deception.
Part IV: The Effect of True and False Intentions on Reaction Times

Chapter 9: Detecting Lies about Intentions Using a Lexical Decision Reaction Time Task

Part IV consists only of Chapter Nine. Chapter Nine describes an experiment that uses participants’ reaction times to judge whether they are telling the truth or lying. Using reaction times to detect deception has a long tradition (Marston, 1920). Several different paradigms have been developed to elicit the reaction times that are used as a cue to lie detection.

The “oddball” paradigm is one such paradigm. It is most suitable for use in a Concealed Information Test (CIT). CITs are used to assess whether the participants know something particular (e.g. details about a crime) although they claim not to. Rather than focussing on the story and classifying it as truthful or deceptive, CITs focus on discovering the truth (Verschuere & De Houwer 2011). In an oddball task the participant is asked to respond to particular target stimuli, which they have learned by heart beforehand, by pressing one key whilst pressing another for irrelevant stimuli, those that were not learned. The targets are shown on about 15% of the trials, which means that the participant gets more used to responding to irrelevant stimuli than to targets. This leads to longer response times for targets than for irrelevant stimuli. About 15% of the non-target trials are probe trials. These are stimuli that are related to the information that the participant is believed to be concealing. Since probes are different from targets, the participants should respond to them as an irrelevant item. However, research has found that people who are concealing information related to the probes respond more slowly to probes than those who are not concealing information related to them (Farwell & Donchin, 1991; Allen, Iacono & Danielson, 1992). It is believed that this is caused by stimulus
response compatibility. For truth tellers an oddball task is a simple task of distinguishing familiar from unfamiliar stimuli. Liars must also classify stimuli as familiar-unfamiliar, but they must also classify some words that are familiar and meaningful to them as unfamiliar. Therefore they have three different types of stimuli but must classify them in two categories. This is a harder task, demanding more cognitive load which leads to a greater latency when answering probes (Verschuere & De Houwer, 2011).

Another paradigm to elicit reaction times for lie detection is the Implicit Association Test (IAT). This test was originally developed to measure the association between a target and an attribute. The test has five phases. First, the participant must associate stimuli as being part of one of two target groups (e.g. black or white) by pressing different keys. In the second phase, the participant classifies stimuli according to an attribute (e.g. pleasant or unpleasant). In the third phase, each response option is associated with both one of the target groups and one of the attributes. One key must be pressed when a stimulus belongs to one target group or if it has one attribute (e.g. black or pleasant) and the other key if the stimulus belongs to the other target group or has the other attribute (e.g. white or unpleasant). In the fourth phase, phase one is repeated but the response keys are reversed. In phase five, phase three is repeated, but with the target-attribute combination reversed (e.g. black-pleasant becomes white- pleasant) (Greenwald, McGhee & Schwartz, 1998). Sartori and colleagues (2008) adapted this IAT for lie detection. In the first block participants had to classify true and false statements (e.g. true: I’m in the psychology department, false: I’m at the beach). In the second they classified guilty and innocent sentences. These were related to mock crimes the participants committed or read about before the IAT or to real drug use crimes they
had committed before in their lives. In the third phase, the participants classified true
and guilty sentences with one key and false and innocent sentences with another. In
the fourth phase, they classified guilty and innocent sentences with counterbalanced
key presses. In the fifth phase they classified true and innocent sentences with one
key press and false and guilty sentences with another. Sartori et al (2008) found that
in phase three (guilty/true vs. innocent/false), guilty participants respond faster than
innocent participants, while in phase five (guilty/false vs. innocent/true) innocent
participants responded faster than guilty ones. Over several experiments with
different types of mock and real crimes, they could classify 91% of participants
correctly as innocent or guilty.

A third paradigm used for eliciting reaction times, is the Sheffield Lie test
(Spence et al. 2001). In this paradigm the participant reads or hears sentences, which
have to be answered with yes or no by button press. The possible answers appear in
colour on a computer screen, the participant is instructed to answer truthfully if the
answer options are in one colour and to answer deceptively if the answer options are
in a different colour. Spence et al. (2001) used this paradigm and found that
deceptive responses are associated with longer reaction times.

A fourth paradigm, the one used for this experiment is based on Ferguson
and Bargh’s (2004) research into intentions. It is used in this experiment because it is
derived from intentions related research and that makes it easier to adapt to a
situation in which the participants are lying about their intentions. They used a
reaction time task in which participants were asked to classify adjectives as positive
or negative. Before the adjective was shown, participants were primed by being
shown a word for 150ms. This word would be related to an unrelated task the
participants would either have to continue performing after the adjective
classification task or which they had already finished. Ferguson and Bargh’s (2004) results showed that those who had completed the task classified positive and negative adjectives at the same speed, while those who were going to continue the task after the adjective classification responded slower to negative adjectives than to positive adjectives. Ferguson and Bargh’s (2004) suggest that this effect indicates that the participants have automatically evaluated the primes that are related to the future task as positive. They suggest that this positive evaluation might be caused by the intention creating an associative network in memory which facilitates links to ideas related to the intention. They also suggest that the effect might be beneficial to completing the goal, since it makes things that are probably helpful in attaining the goal more attractive.

Ask, Granhag, Juhlin and Vrij (2011) adapted this paradigm for lie detection purposes. They asked some participants to plan to go on a shopping task and complete it later on. They asked other participants to plan to plant a USB-stick in a store. They instructed the shopping group to tell the truth about their intended task if they were asked. They instructed the planting group to lie and say that they were going on the shopping task if they were asked questions. The participants then completed an adjective classification task in which they were primed with words relating to the shopping task (e.g. till, see Table 9.1), which was the true intention for one group of participants (truth tellers) but a false intention for the other group (liars). They found that liars responded with equal speed to positive and negative words (see Table 9.1), while truth tellers responded more slowly to negative words.

The experiment described in this chapter is based on the same paradigm as the experiment by Ask et al (2011). Forty participants, truth tellers, were asked to go to a nearby shopping centre and buy two presents. They were told to tell the truth if
they were asked questions by other experimenters en route. Half of the participants were told they would complete the task immediately after the planning, the other half were told they would complete it two weeks later. Another forty participants, liars, were asked to go to the mall and plant a USB-stick in one of the stores. They were instructed to lie and say their task was to buy presents if they were questioned en route. Again half were told to go immediately, the other half to go two weeks later. After planning their task and, for liars, their cover story, the participants were asked to complete the adjective classification task. The primes were related to the shopping task, which was the real intention for the truth telling group and the false intention for the lying group. Because these primes are related to the truth tellers real intention they should be a positive words for them and this should make them faster on positive adjectives then on negatives. Because the primes are not related to the liars’ intention, although they do know the words, they should not show this effect in their reaction times. We therefore hypothesised that lying participants would classify positive and negative words equally fast, while truth tellers would be slower on negative words than on positive words (Hypothesis 1). Since events further in the future seem to be less vivid in someone’s mind (D’Argembeau and van der Linden, 2004; Arzy, Adi-japha and Blanke, 2009) than events that will occur sooner, we hypothesised that this effect would be larger for those in the immediate group than for those in the delay group (Hypothesis 2).

Method.

Participants.

Eighty-eight participants took part. Of these, 72 were female and 16 were male. They were on average 21.68 years old (SD = 6.83). Sixty-six participants were
psychology students, ten were other university students and three were A-level students. The other participants were mostly university employees in support roles, such as administrators and technicians. There was one farmer.

**Design.**

A 2 (Veracity) by 2 (Delay) design was used. Veracity consists of truth tellers and liars. Delay consists of those who were instructed to complete their mission immediately (immediate) and those who were instructed to return in two weeks and complete their mission (delayed). Both Veracity and Delay were between subject variables.

There were three dependent variables based on the reaction times: the average response time to positive words, the average response time to negative words and the difference between these two variables. This final variable was calculated by subtracting the average for negative words from the average for positive words.

There were four dependent variables based on the questionnaire: ‘motivation’, ‘convinced’, ‘intend’ and ‘preparation’. ‘Motivation’ asks the participants how motivated they were to complete the mission. ‘Convinced’ asks to what extent they were convinced they were really going on the mission. ‘Intend’ asks to what extent they intended to carry out the mission. ‘Preparation’ asks how much they prepared for the mission. Each was measured by the participants indicating their answer on a seven point scale ranging from [1] not at all to [7] extremely.
Procedure.

Participants were recruited via an online participant pool and advertisement. They were told that they would be going on a mission to a nearby shopping centre and that they may be asked to complete the mission on the same day as receiving the instruction or to return two weeks after receiving the instructions. They were also told that they may be asked to lie about the mission. When the participants arrived for the study they were randomly allocated to the truth telling or lying condition and to the delay or no delay condition.

Those in the truth telling condition were given the following instructions. “When you go on your mission [later today or two weeks from now, depending on the delay condition], I want you to go to Gunwharf [a nearby shopping centre]. I want you to buy two presents from two different stores. I will give you £20 to do this. Try to spend exactly £20. You may not spend more money and try to have as little change left as possible. Come back to the university and give me: the presents, the receipts and any change you may have left. You won’t be allowed to keep the presents or the change. While you are doing this you will be intercepted by people from the experiment. They will ask you questions about your mission. Please tell them the truth about you mission and your actions. They may seem suspicious of you because they know some participants are lying to them. Your goal is to convince them you are telling the truth”

Those in the lying condition were given the following instruction. “When you go on your mission [later today or two weeks from now, depending on the delay condition], I want you to go to Gunwharf. I will give you a USB-stick. Please plant this USB stick in the Past Times store, on the middle shelf of the bookcase in the
back of the store. Once you have done this come back to the university. While you are doing this you will be intercepted by people from the experiment. They will ask you questions about your mission. Please lie to these people. Do not tell them about the USB-stick or that you are planting anything. Instead use the following cover story: your mission is to buy two presents from two different stores. Say that I’ve given you £20 to do this and have asked you to try and spend exactly £20. After buying the presents you’re going to come back to the university and give me the presents, the receipts and any change you may have left. The people who are asking you questions are trying to find out who is really going on this mission and who is lying. It is your goal to convince them that you are telling the truth.”

After the instructions both truth tellers and liars were given the chance to ask questions about the mission and they had eight minutes to prepare themselves. They could use the internet to prepare. The Gunwharf shopping centre has a website with maps and lists of stores and most stores have websites with lists of products. The participants were instructed not to make notes of their prepared plans. When their preparation time was up they were sent to a different room to obtain their £20 or the USB-stick, which they were told was in a big, padded envelope so that the confederate who was handing it out did not know what it contained or what their mission was. The confederate then asked: “What is your mission?” Both truth tellers and liars should now say that they were going to buy presents. Lying participants who answered that they were going to plant a USB-stick were excluded from the study. Two participants did this; both later said that it was not that they had not understood the instructions. Rather they had not anticipated that they would be asked to lie so soon and in their confusion stated the truth. If the participant answered correctly, the confederate told the participant that she did not want to give them the
envelope until the participant had been interviewed and judged to be telling the truth by the interviewer. They were told to return to the experimenter and tell them to arrange the interview. After the participant had done this, the experimenter told the participant that she would arrange for the interview and asked them to complete the computerised lexical decision task, described under Materials. They were not told that this task was meant to detect their intentions. Instead they were told that it was a filler task that prevented them from mentally preparing for the interview, while the experimenter was away arranging the interview.

After finishing the computer task, the experimenter informed the participant that they would not have to complete the mission. The experimenter apologised for the deception and explained that this had been necessary, because the experiment required the participant to have an intention, but it did not require them to execute that intention. The participant was asked to fill in a questionnaire and sign a second informed consent form, which included all the information regarding the deception. No participant refused to sign the second informed consent form. They were then debriefed and given their reward, some candy or credit for the university participant pool.

Materials.

The computer task was created using e-prime. The computer program contained two tasks. First was the primed lexical decision task that acted as the lie detection tool. Table 9.1 lists the stimuli used in the task. The program gave the participants instructions on the task. They were told that they would see 72 trials, each consisting of a prime, a word that would appear and disappear quickly, and then an adjective. They were told to only respond to the adjective by classifying it as
positive or negative. This was done by pressing one key if the adjective was positive and another if it was negative. The keys to be pressed were counterbalanced. They could practice eight trials first. The task started when the participant pressed a key. The 72 trials were shown with one break after 36 trials. The prime lasted 150 ms, a blank screen was shown for one second, then a fixation mark for one second, then a blank screen for 150 ms and then the adjective was shown until the participant responded. This triggered a new trial. Each prime appeared eight times and each adjective twice. Each prime was presented twice with a positive adjective and twice with a negative adjective. After 72 trials, the participants received instruction for the second task. They were shown the primes one by one and were asked to rate these as positive or negative on a seven point scale from [1] very negative via [4] neutral to [7] very positive. The participants rated ‘cashier’ and ‘receipt’ as neutral, not significantly different from [4] (all $t < 1.14$, all $p > 0.26$). They rated ‘twenty’, ‘buy’, ‘purchase’ and ‘shopping’ as positive, significantly higher than [4] (all $t > 2.93$, all $p < 0.004$). They rated ‘pay’, ‘price’ and ‘till’ as negative, significantly lower than [4] (all $t < -1.92$, all $p < 0.04$). The mean and standard deviation of the ratings are listed in Table 9.1. There was no significant difference between positive and negative words regarding their length ($t (34) = 0.39$, $p = 0.70$) or how common they are in the English language, measured by their rank in a list of most frequently used words in English ($t (34) = 0.65$, $p = 0.52$) (Harris, 2003).
Table 9.1: The stimulus words used in the adjective classification task. The mean (and SD) of the participants’ rating of the primes is listed.

<table>
<thead>
<tr>
<th>Prime</th>
<th>Mean (SD)</th>
<th>Positive words</th>
<th>Negative words</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>4.33 (1.02)</td>
<td>Happy</td>
<td>Disgusting</td>
</tr>
<tr>
<td>Buy</td>
<td>4.72 (1)</td>
<td>Sympathetic</td>
<td>Detestable</td>
</tr>
<tr>
<td>Cashier</td>
<td>4.04 (0.83)</td>
<td>Attractive</td>
<td>Terrible</td>
</tr>
<tr>
<td>Pay</td>
<td>3.70 (1.39)</td>
<td>Exquisite</td>
<td>Foul</td>
</tr>
<tr>
<td>Price</td>
<td>3.44 (0.88)</td>
<td>Funny</td>
<td>Painful</td>
</tr>
<tr>
<td>Purchase</td>
<td>4.78 (1.07)</td>
<td>Cozy</td>
<td>Wretched</td>
</tr>
<tr>
<td>Receipt</td>
<td>4.11 (0.89)</td>
<td>Cute</td>
<td>Miserable</td>
</tr>
<tr>
<td>Shopping</td>
<td>5.07 (1.4)</td>
<td>Beautiful</td>
<td>Gruesome</td>
</tr>
<tr>
<td>Till</td>
<td>3.84 (0.7)</td>
<td>Cheerful</td>
<td>Nasty</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Excellent</td>
<td>Selfish</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lovely</td>
<td>Tragic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pleasant</td>
<td>Gross</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gorgeous</td>
<td>Horrible</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Awesome</td>
<td>Dreadful</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Delicious</td>
<td>Deplorable</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Brilliant</td>
<td>Ugly</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Superb</td>
<td>Useless</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Delightful</td>
<td>Awful</td>
</tr>
</tbody>
</table>

Data analysis.

The data were taken from e-data and exported to SPSS. For each participant their mean reaction time and the standard deviation were calculated. The following responses were then excluded: wrong responses, responses where the reaction time was below 250 ms and responses where the reaction time was higher than the participant’s own mean plus three times their own standard deviation. These were excluded because these trials might be an indication that the participant did not know the word they were asked to classify or may have been disturbed in some way during the trial. In this way 176 out of 6336 trials were excluded.

Descriptive analyses showed that the data were skewed, as is common with reaction time data (Whelan, 2008). Positive adjectives had a skew of 7.76 (SE = .26)
and a kurtosis of 67.57 (SE = 0.51). Negative adjectives had a skew of 8.18 (SE = .26) and a kurtosis of 72.56 (SE = 0.51). In order to combat this, the data was log transformed. After the transformation, the positive adjectives had a skew of 2.10 (SE = .26) and a kurtosis of 11.34 (SE = 0.51). Negative adjectives had a skew of 2.37 (SE = .26) and a kurtosis of 12.99 (SE = 0.51).

All trials completed by participants who answered more than 12 trials incorrectly were excluded as were participants whose average response time was over three standard deviations higher than the average of all participants. This left 83 out of 88 participants in the analysis. Now, positive adjectives had a skew of 0.28 (SE = .26) and a kurtosis of -0.20 (SE = 0.52). Negative adjectives had a skew of .31 (SE = .26) and a kurtosis of -0.43 (SE = 0.52). In total 10.4% of datapoints were excluded. Then an average for all, not excluded, positive adjectives (range: 6.15-7.00) and all not excluded negative adjectives (range: 6.17-6.98) were computed. Finally, a difference variable of the average of positive words minus the average of negative words was computed (range: -0.18-0.20).

**Results.**

The mean score for ‘motivation’ was 5.42 (SD = 0.94), with 85.5% indicating 5 or higher; for ‘convinced’ the mean score was 5.69 (SD = 1.37) with 81.9% indicating 5 or higher; for ‘intend’ the mean was 6.08 (SD = 0.97), with 91.6% indicating 5 or higher and for ‘preparation’ it was 4.56 (SD = 1.21) with 50% indicating 5 or higher. The high means for convinced and intend suggest that the participants did believe that they were going to go on their mission to Gunwharf. Preparation scores were lower, suggesting that participants did not prepare much. Perhaps they felt that they would be able to complete the task without much
preparation, since they all knew the mall. A MANOVA with Veracity and Delay as the independent variables and motivation, convinced, intend and preparation as dependent variables revealed no significant interaction effect between Veracity and Delay on any variables (all $F < 3.35$, all $p > 0.07$). T-tests showed that liars were more motivated than truth tellers ($t (81) = -2.33, p = 0.02$). Neither Veracity nor Delay affected any of the other variables (all $t < -1.65$, all $p > 1.03$).

An ANOVA was conducted with the difference between positive and negative words as the dependent variable and Veracity and Delay as the independent variables. A significant interaction between Veracity and Delay was found ($F (1) = 7.65, p = 0.01, \eta^2 = 0.09$). No significant main effects were found (Veracity: $F (1) = 0.67, p = 0.42, \eta^2 = 0.01$; Delay: $F (1) = 0.06, p = 0.8, \eta^2 < 0.00$). Figure 9.1 shows the pattern related to the significant interaction effect. The mean scores are not easy to interpret due to the log transformation of the y-axis. The estimated marginal means show that for all cells the mean is negative, indicating that participants were faster on positive adjectives than on negative adjectives. Immediate truth tellers obtained a significantly smaller difference between positive and negative adjectives than immediate liars ($t (42) = 2.55, p = 0.02$). Delayed liars obtained a smaller difference than delayed truth tellers -the opposite pattern compared to the immediate participants but the difference between truth tellers and liars was not significant ($t (37) = -1.38, p = 0.17$).

In further analyses we compared whether the difference obtained between positive and negative adjectives differed from zero. For immediate truth tellers and delayed liars, the difference between positive and negative adjectives was not significant (immediate truth tellers: $t (19) = -0.85, p = 0.41$; delayed liars: $t (21) = -1.35, p = 0.19$). In contrast, delayed truth tellers and immediate liars were
significantly faster on positive adjectives than on negative adjectives (delayed truth
tellers: \( t (21) = -3.93, p = 0.001 \); immediate liars: \( t (18) = -5.15, p < 0.001 \)).

Figure 9.1: *The difference between positive and negative adjectives.*

N.B. the lower the number on the y-axis the greater the difference in reaction
time between positive and negative adjectives, with positive adjectives being
responded to faster.

A ROC curve analysis showed that the participants could not be classified as
thruth tellers or liars at above chance rates (AUC = 0.57, SE = 0.06, \( p = 0.3 \)). The
discriminant analysis showed that 57.8\% could be classified correctly, which did not
differ significantly from chance (\( \lambda = 1, \chi^2 = 0.39, p = 0.53 \)).
Discussion.

The results revealed that an interaction between veracity and delay affected the participants’ reaction times. Immediate truth tellers responded equally fast to both types of adjectives, but immediate liars responded faster to positive adjectives than to negative adjectives, an outcome quite the opposite of what was predicted in Hypothesis 1. Delayed liars responded equally fast to positive and negative adjectives, while delayed truth tellers responded significantly faster to positive adjectives than to negative adjectives. This was in line with Hypothesis 1, however there was no significant difference between delayed truth tellers and liars. In Hypothesis 2 it was predicted that any difference between truth tellers and liars would be larger in the immediate condition than the delayed condition. This was not found and Hypothesis 2 was therefore not supported. Although we did find significant differences in the immediate condition and not in the delayed condition, the direction of the effects was opposite.

The finding that in the immediate condition truth tellers responded equally fast to positive and negative adjectives while liars responded faster to positive than to negative adjectives was opposite to Ask et al.’s (2011) results. There are several possible explanations for this. It may be an effect of language. Ask et al. (2011) conducted their experiment in Swedish, while this experiment was conducted in English. It may be that the English adjectives or primes elicited different responses than the Swedish ones. One possible reason for a difference in response between languages could be that English is more orthographically irregular than Swedish, which means that Swedish may be easier to read (Aro & Wimmer, 2003). Another difference between the studies was the time lapse between the prime and the adjective shown, it was 2150 ms in the present experiment and 150 ms in Ask et al.
It is also possible that there were small differences in the instructions or the behaviour of the experimenter or confederate. Any of these explanations suggests that the effect is not robust and small changes in the materials and set up of the experiment result in large outcome differences. Future research should aim to find out why these small differences in method create such large differences in respond. It may be that different methods tap into different cognitive processes. Discovering the mechanisms behind these findings may yield important information regarding these cognitive processes.

The finding that differences between truth tellers or liars was greater in the immediate condition than in the delayed condition supports the idea that intentions that are to be executed further ahead in the future are less salient than intentions that are to be executed sooner. This is consistent with D’Argembeau and van der Linden (2004) and Arzy, Adi-japha and Blanke (2009) who argued that mental time travel is more salient when there is less time between the present and the mental event. Lies about intentions that are to be executed closer to the present may be easier to detect than lies about intentions that are to be executed further from the present, because those who genuinely hold a later intention, hold it with less salience and vividness than sooner intentions, decreasing the difference between truth tellers and liars. The finding has practical implications for detecting lying about intentions. It suggests that it may be important in practical situations to know when someone plans to execute their intention. This should then be taken into account when analysing their response.

The non significant classification rate obtained in this experiment, as well as the contradictory results compared to the Ask et al. (2011) experiment, suggests that the current paradigm is at the moment of no practical use. In order to clear up the
contradiction with Ask et al. (2011) experiment, a study could be run, using the same paradigm, in which the lag between prime and target is systematically varied to see the effects this has on the reaction times of truth tellers and liars. It is also possible that this entire prime adjective classification paradigm is not optimal and that using more traditional lie detection tasks, (e.g. the oddball or IAT paradigm) may be more successful in detecting lies about intentions if they were suitably adapted.

The classification rate for this reaction time experiment was, together with non-verbal analysis, the lowest scoring of the methods studied in this thesis. Further research may improve on the hit rates of this method, but such a small and apparently non robust effect must be considered a disadvantage for the practical use of this method. The method also has other disadvantages. It requires a working computer with the necessary software and someone to instruct the participant on how to carry out the task. The task also requires the participant to be motivated and to truly respond as fast as they can. A lying participant could easily sabotage the task by answering slowly on random trials. Also participants for whom the language of the task is not their first language may find it difficult to complete the task, which may affect the reliability of the results. Overall, the reaction time method used here is not ready to be used in practical situations.
Part V: General Discussion

Chapter 10: General Discussion

Summary of findings.

The aim of this thesis was to explore whether lying about intentions could be detected using the same methods that are used to detect other types of lie. In Part I, the thesis was introduced and Part I contained a literature review of research into future thinking and intentions. In Part II non-verbal cues and thermal interviewing were investigated as cues to deception. In Part III verbal cues in longer and shorter interviews were examined. In Part IV reaction times in a lexical decision task were investigated.

In Part II (Chapters 3 to 5) Experiment 1 was described in which participants of twelve different ethnicities were interviewed at an airport about their forthcoming trip. Half of the participants were instructed to answer truthfully, the other half to answer deceptively. In Chapter 3 Experiment 1 was introduced and the methods used were described. In Chapter 4 the differences between truth tellers and liars on 17 non-verbal cues were described. The participants’ veracity affected only three of the cues: struggling, thinking hard and controlling behaviour. The participants’ ethnicity and level of English affected more cues, especially paraverbal cues. The 17 cues were reduced to two factors, cognitive load and animation. These two factors could correctly classify 52.3% of truth tellers and 62.6% of liars.

In Chapter 5 the results of the thermal imaging footage taken in Experiment 1 were described. It was found that liars tended to get warmer during the interview, while the truth tellers’ temperatures remained constant. When comparing truth tellers and liars (between subjects) 68% of truth tellers and 65% of liars could be classified
correctly. When participants were classified according to how much their temperature increased (within subjects) during the interview, 72% of truth tellers and 54% of liars could be classified correctly. These percentages were above chance, but lower than the accuracy rates achieved by the interviewers in the study: they classified 72% of truth tellers and 77% of liars correctly. When both the interviewers’ judgement and the rise in the participants’ temperature were taken into account 68% of truth tellers and 85% of liars could be classified correctly.

In Part III (Chapters 6 to 8) two experiments were described. In Chapter 6 verbal cues used in lie detection were introduced. In Chapter 7 an experiment in which participants were interviewed, answering truthfully or deceptively, about a forthcoming trip was described. The interview contained general questions, which were expected by the participants and three types of unexpected questions: questions about the core event, travel and planning. Half of the participants had been to the travel destination before, while the other half had never been there. The results showed that the participants’ veracity, their previous experience and the type of question asked affected the amount of total details, visual details, spatial details and temporal details, as well as the amount of explanations they gave. Truth tellers generated more details on the unexpected questions than liars, while liars provided more details on expected questions than truth tellers. This pattern was more pronounced in participants who had been to their travel destination before than for those who hadn’t. When we classified participants as truth tellers or liars, based on the difference in detail they mentioned in response to the expected questions and the unexpected questions, a positive predictive value of .69 and a negative predictive value of .75 were obtained. These figures were significantly higher than chance.
In Chapter 8 it was investigated whether verbal cues to deception could be obtained in very short, one question interviews about a forthcoming trip. The study focussed on temporal details, since these had appeared the most promising in Experiment 2. The interview question was asked in two forms, one with a time prompt and one without. The results showed that truth tellers and those answering the time prompting question were more likely to give temporal information. The results showed that 47.6% of truth tellers and 81% of liars could be classified correctly.

In Part IV (Chapter 9) reaction times were investigated as a cue to deception. Participants were asked to perform a mission, either immediately or two weeks later. Half of the participants was instructed to answer any questions asked by experimenters en route truthfully, the other half was asked to answer deceptively. The study used a task in which participants classified adjectives as positive or negative and were primed by words related to the real intention of truth tellers, which was also the cover story for liars. Contrary to the hypothesis, it was found that in the immediate condition, liars answered more quickly on positive than negative adjectives, while the truth tellers responded with equal speed to both adjectives. In the delayed condition no difference was found between truth tellers and liars.

Theoretical implications.

The accuracy rates based on interviewers’ judgments (Experiment 1) or verbal cues (Experiment 2) were in line with the results found in previous work on lying about intentions (Vrij, Granhag, Mann and Leal, 2001; Vrij, Leal, Mann and Granhag, 2010). The finding that liars gave more detail than truth tellers (Experiment 2) to expected questions is contrary to DePaulo and colleagues’ (2003) meta-
analysis. However, since most studies probably contained at least some unexpected questions, to which liars give less detail, this finding is not a contradiction, but rather shows that those findings might not be generalisable to all situations. The lack of significant differences between truth tellers and liars regarding the non-verbal cues (Experiment 1) was contrary to DePaulo and colleagues’ (2003) analysis. That meta-analysis revealed differences between truth tellers and liars for hand movements, illustrators and appearing nervous. This difference may be caused by the difference between lying about intentions and lying about other issues, as examined in the studies used in the meta-analysis. It is also possible that the difference is caused by the wide range of ethnicities included in this study, while most studies in the meta-analysis would have relied on Western participants. Participants from different ethnicities and backgrounds may exhibit different verbal and non-verbal behaviour (Johnson, 2007; Caldwell-Harris & Ayçiçeği-Dinn, 2009). Again this suggests that some of the findings in the lie detection literature are not necessarily generalisable to other situations than those used in the studies. The finding that liars appear to control their behaviour more (Experiment 1) supported Hartwig and colleagues’ (2007) finding that liars use more strategies to appear convincing. The results of Experiment 4 for reaction times were contrary to previous research (Ask et al. 2011). This may be due to differences in methodology, such as the language in which it was conducted and the timing of the stimuli. This does suggest the effect obtained by Ask et al. (2011) may not be very robust.

The results generally support the cognitive load theory of lie detection, which states that lying is more cognitively demanding than telling the truth and that this can yield cues to whether someone is lying (Vrij et al. 2006). In Experiment 1 it was found that lying participants experienced the interview as harder and were
judged by observers as appearing to be thinking harder. Experiment 2 revealed that liars prepared more, suggesting that they believed the interview would be harder than if they were to tell the truth, and that preparation would benefit them. The liars gave less detail than truth tellers on questions they had not expected and could not prepare for, suggesting that they found answering these questions harder. Also in Experiment 1, the 17 non-verbal cues could be reduced to two factors, the most prominent of which contained cues related to cognitive load. Overall, the results suggest that lying about intentions is more cognitively demanding than telling the truth about intentions.

The results of Experiment 2 and Experiment 4 support previous research about future thinking and mental time travel, in which it was found that images of the future are usually less clear and detailed than images of the past (D’Argembeau and Van der Linden, 2004; Gamboz, Brandimonte and De Vito, 2010). Experiment 2 showed that there was no overall difference in number of details between truth tellers and liars who were discussing the core event of their intention, while previous research into lies about the past has found that truth tellers mentioned more details than liars. This difference may be because images of the future are vaguer than images of the past, making the truth tellers’ accounts less detailed when they discuss their intentions than when they are discussing past events. Experiment 4 revealed that differences in reaction times between truth tellers and liars were larger in the immediate condition than in the delayed condition. This supports the idea that intentions that are to be executed further in the future are less salient than intentions that are to be executed sooner. This is consistent with D’Argembeau and van der Linden (2004) and Arzy, Adi-japha and Blanke (2009) who argued that mental time travel is more salient when there is less time between the present and the mental
event. Overall, the experiments have generally supported previous research into mental time travelling, episodic future thought and verbal cues to lying about intentions. Previous deception research regarding non-verbal cues and reaction times has not been consistently supported. This suggests that the cues to deception found in studies on lying about things other than intentions cannot be generalised to lying about intentions. Research into mental time travel and episodic future needs to be taken into account in order to detect lies about the future more efficiently.

**Practical implications.**

The goal of detecting lies about intentions is to prevent crime, rather than apprehend criminals. Police and other law enforcement agencies on occasion have to detect lies about intentions (e.g. when deciding whether to let someone go or to detain them longer). The police also help prevent crime by helping the public to protect themselves and their property (Mayor’s Office for Policing and Crime, 2012) or by focussing on specific programs (e.g. Boston’s Operation Ceasefire, a project aimed to lower the homicide rate in Boston) by focussing on particular contributory crimes such as gun trafficking (Kennedy, Graga, Piehl & Waring, 2001). The goal in these cases is usually not to detect intentions, but to change people’s intentions (e.g. to take more security measures or to discourage people from committing crime because the risk is too high). Detecting lies about intentions is more common for other agencies. Security and intelligence agencies investigate people who are suspected of intending or planning to commit crimes, so they have a vested interest in discovering ways to detect intentions. Also transport and border control agencies aim to detain those with criminal or harmful intent whilst allowing those with innocent intentions to pass through as quickly as possible (Burgoon, 2009).
Border control and transport security agencies have finite resources and often have to monitor many people. Therefore they should spend as few resources as possible on each innocent person that passes through. The finding of Experiment 3, that even interviews lasting less than a minute yielded a cue to deception, is therefore encouraging. However none of the experiments discussed in this thesis have accuracy rates that are high enough for use in these security systems. The experiments would leave between 15% and 37.4% of liars undetected. These percentages are not particularly good, but the lie detection methods may be beneficial when used in addition to other methods already in use for detecting potential criminals. Another problem is the number of truth tellers (between 31% and 52.4%) that are classified as liars. Classifying an innocent person as a liar would presumably mean that they would be investigated further. Considering that transport hubs, such as airports or train stations, and borders handle so many people (e.g. Heathrow’s 65,747,200 passengers in 2010), it would impossible to investigate between a third and half of them further.

In this thesis there are several suggestions that may prove useful in practical situations. In Experiment 2 it was shown that liars gave more detailed stories in response expected questions than truth tellers, when they had experience with their intention. Truth tellers gave more detail to unexpected questions than liars, when they had experienced the intention. Firstly this suggests that asking unexpected questions may be beneficial in interviews where detecting lies is one of the goals. Secondly it suggests that an interviewer must take into account how much experience the interviewee has. The difference between truth tellers and liars is larger if they have experience of their intentions. This means that liars should be
easier to detect if they have such experience and conversely if someone has never been to their intended destination, detection may be harder.

In Experiment 4 it was found that differences between truth tellers and liars were larger when the intention was to be executed immediately compared to when the intention was to be executed two weeks later. This suggests that it may be harder to detect people with criminal intentions that they plan to execute a long time after the moment they are investigated. A system designed to detect those with criminal intent may be able to take this into account. If someone is suspected of having a criminal intent, investigating them multiple times over a longer time frame may be beneficial. Also additional security checks might be placed as closely as possible, spatially and temporally, to the event that is to be secured. This might make those with criminal intent easier to detect and, if there is a long gap between security checks and the event, people might develop criminal intents in the mean time. Overall, the present thesis on detecting lies about intentions suggests that the accuracy rates are as yet not high enough to implement any of these lie detection methods in real-life situations. However, the research does suggest that some techniques may be beneficial in combination with others, e.g. combining verbal coding and thermal imaging may be useful. It also suggests that having certain pieces of background information on those being interviewed may also be beneficial.

Limitations.

Methodological considerations.

Psychology research often relies on participants and studies tend to use the easiest available participants: university students. These participants do not represent humanity as a whole very well, in fact they have been described as WEIRD
(Western, Educated, Industrialised, Rich and Democratic; Henrich, Heine & Norenzayan, 2010). These participants often show traits or responses that might not extend to the general public. Experiments 2 and 4 used psychology students from the University of Portsmouth, participants that mostly fit the WEIRD category. This means that the results from these studies might not generalise very well to other people. Experiments 1 and 3 tried to address this problem by using people from the general public as participants. Experiment 1 used participants from all over the world, which means they were not all Western, or from industrialised or democratic countries. They did all have to speak English, which suggests that most of them were educated and since they were able to afford this education and the flight most would probably qualify as rich compared to the world-wide average and perhaps the average of their home country. Experiment 3 used participants travelling between England to France. This increased the age range and the range of cultures of the participants compared to using psychology students, but the participants would still classify as WEIRD, again making generalisation to non-WEIRD people difficult.

Another methodological issue with these experiments is the motivation of the participants and the stakes involved for them. The “motivational impairment” effect is relevant in this context. This effect suggests that people who are highly motivated to be believed are, in fact, less likely to be believed. This may be because they try to control their behaviour more, which gives an inhibited and overcontrolled impression (DePaulo & Kirkendol, 1988). If there is a lot at stake for the liar they will likely be more motivated and that should make it easier to detect their lies (Vrij, 2008). However, at the same time higher motivation may lead to more preparation, which may make liars more likely to be believed. In all experiments, the experimenter attempted to motivate the participants by giving them a monetary,
nutritional or university credit reward. However, it would be unethical to threaten the participants with serious punishment or promise them large rewards. In all experiments, when the participants were asked how motivated they were to convince the interviewer, the participants rated themselves as highly motivated. Also, the experimenter often noticed that participants were nervous. Participants often stated they wanted to convince the interviewer and were proud when they did convince and disappointed when they did not. All this suggests that the participants were quite motivated. However, there is a limit to how motivated someone can be for something that can have no serious consequences. Therefore, the stakes in all the experiments were relatively low. This would suggest that our liars were not under a lot of pressure, which would be beneficial to them and make them harder to detect.

Overall, the results found in this thesis are based on low stake situations and may not generalise to high stake situations. The motivational impairment effect suggests that the effects found may be stronger in high stake/high motivation situations, so that truth tellers would give more details in high stake interviews, while liars may not give more details. Also, liars may appear even more controlled than in the low stake situations used in these experiments. This needs to be studied empirically, perhaps by using case studies or using data available in the public domain, since experiments would be unethical.

**Ethical considerations.**

As was made clear in the practical implications section, none of the methods to detect lies about intentions are accurate enough to be used in practical situations. However, if we had a method to detect lies about intentions in 100% of the cases would it be ethical to use it? Freedom of thought is an ancient right (Bury, 1913) and is listed in article 18 of the Universal Declaration of Human Rights (UDHR, 1948).
It can be argued that anyone has the right to hold any intention and no-one has the right to try and discover or reveal this intention. Some human rights can be ignored under certain circumstances. For example, the human right declaration lists the right to life, liberty and personal security, but governments can legally execute or imprison people if they are convicted of certain crimes. Would such an exception be made for criminal intentions? If there is evidence that someone holds a criminal intention, could their right to freedom of thought be violated in order to reveal this intent? And if this is the case, could someone be punished for holding a criminal intention, even if they committed no illegal act and did not conspire with anyone else to commit an illegal act? These are serious questions that should be debated, not only in a scientific context but also in the public domain. A world in which people can be punished for their intentions has already been imagined and is mostly viewed as a dystopia (Dick, 1956). It would be beneficial for these ethical questions to be debated before a perfect or near perfect lie detection system is developed, however unlikely it might seem now that such a system could ever be created.

Aside from the ethical problems associated with a perfect detector of lies about intentions, there are ethical problems associated with a less than perfect detector. Polygraphs have been used by the police for years, although they are usually not admissible in court (Ben-Shakhar, Bar-Hillel, & Kremnitzer, 2002). When innocent people fail polygraph tests they may come under increased scrutiny by the police and this may lead to convictions (see e.g. the innocence project, 2012). However, when the topic of the polygraph test was a past event, there are other ways to prove their innocence, e.g. via witnesses or physical evidence. When someone fails a polygraph on their intentions there is no other way to prove the test was wrong. The lack of ground truth for intentions makes it harder to test the validity of
the lie detector and makes the consequences of a false classification as a liar more serious. This reiterates the need for a clear and open debate on whether the positive uses of detecting lies about intentions weigh up against the possible negatives for society.

**Future research.**

The results from the experiments suggest that using verbal cues is the most efficient to detect lies about intentions. Therefore, future research should focus on improving this method even further. Experiment 2 asked the expected, general questions before the unexpected questions about the core event, travel and planning, in accordance with standard practice. It may be interesting to investigate whether asking unexpected questions first would affect participants’ answers. It may be that liars prefer to get their prepared story out and answer with more details to the first questions asked, regardless of what is asked. Liars were found to give more details to expected questions than truth tellers. It would be interesting to test whether liars do this with all prepared stories (e.g. alibis). If the liar gets a simple question that can be answered in very few words, do they answer with very few words or do they give their entire prepared story? Also more research is needed into what makes a question unexpected and whether certain types of unexpected questions are more beneficial to detecting deception than others. Experiment 3 showed promising results for temporal detail when using very short interviews. More research should be conducted examining whether short interviews can be used to discriminate between truth tellers and liars on details other than temporal details, and whether the phrasing of the questions can be altered to increase the difference between liars and truth tellers on these other details.
Regarding thermal imaging, further research into the cause of the liars’ warming up, whether it is a sign of nervousness or cognitive load, would be useful. Also, the thermal imaging technique used in the experiment was relatively difficult to use, analysing the data was a slow and error-prone process, and further development of the recording and analysing elements of the thermal imaging technique may be beneficial. A way to analyse the data in real time would be a great improvement in the practicality of using it. Experiment 1 showed that participants need to be interviewed for temperature differences to occur. Further research should focus on improving the interview question in such a way that the liars warm up most, while the truth tellers remain at a constant temperature. Experiment 1 suggests that information gathering questions, rather than accusing questions are required to achieve this. Perhaps more unexpected questions or more difficult questions would be beneficial.

Experiment 4 revealed some results that contradict theories and empirical findings on true and false intentions (e.g. Ask et al. 2011). More research into the effects of changing these reaction times task slightly is needed to establish whether this experiment was an outlier or whether this indicates that the theory behind earlier research is not always valid. It would also be interesting to investigate other reaction time paradigms that have been used for detecting lies about the past and see whether they are suitable for lies about intentions. Both the Implicit Association Test, which relies on association between a statement related to the truth or lie and statements that are known to be true or false and the oddball paradigm, which relies on using familiar, unfamiliar and crime-related words in the correct frequency, should be relatively easy to adapt to lies about intentions and may give more robust results.
Future research may also investigate other techniques that have been used to detect lies about the past. Polygraphs are the most well-known lie detection tools. The Control Question technique could be adapted to lying about intentions (Kircher, Horowith & Raskin, 1988). The Concealed Information Test might be harder to adapt, since it relies on precise knowledge of the event discussed (Ben-Shakhar & Elaad, 2003). It may be difficult to get the correct items to conduct a lie detection test about intentions. Since both techniques assume the participants have remembered details about the event in question, it would be interesting to see whether their intentions are clear enough to make a difference between true and false intentions in the polygraph measures.

EEG had been used in lie detection to detect lies about the past (Farwell & Donchin, 1991). It may be possible to adapt the paradigm to investigate whether lies about intentions can also be detected. Also, EEG research has been carried out outside the field of deception to investigate cerebral activity when someone is intending to perform an action (e.g., Libet, Gleason, Wright & Pearl, 1983). It was found that shortly before an action is performed there is a discernible pattern in brain activity and that this activity may be present before the participant is aware of having made the decision to execute the action. It may be possible to use the theories and techniques from this field of research to detect lies about the future.

FMRI (Spence, Farrow, Herford, Wilkinson, Zheng & Woodruff, 2010) and Magneto Encephalograms (MEG) (Seth, Iversen & Edelman, 2006) are among the latest methods to enter the arsenal of the lie detector. FMRI has also been used to study how humans mentally time travel (Addis et al., 2009; Szpunar, Chan & McDermott, 2009). Knowledge from these two fields of research may be combined to design fMRI experiments in which people think about true or false intentions.
Perhaps differences in their brain activity can be found. Although there is evidence that finding such as method may be hard to use in real life (e.g. Ganis et al., 2012)

Besides investigating different methods to detect lies about intentions future research should also focus on situational differences. Does a liar respond differently in high stakes compared to low stakes situations? It will be difficult to investigate this, since high stakes cannot be induced in the lab. It may be difficult to find real life examples of someone lying about an intention, since it is difficult to know whether they were lying at the time the statement was made or whether they changed their minds about their intention between the statement and the moment the intention was executed. Ground truth will be even harder to establish than in cases of high stakes lies about past events. Another possible area of research is whether participants with different backgrounds or personalities discuss their true or false intentions differently. Besides research into the affects of the “big five” personality traits, one might examine if people who are more fatalistic, who believe in fate or destiny, discuss their intentions differently from those who believe they have control over the future. Those with a more fatalistic worldview may have a less detailed image of the future, because someone who believes they control their future would need a clear image of the future to work towards. Another point of interest is whether all intentions are discussed in the same way. Some intentions are more likely to be executed successfully than others, so how does this affect the way the intentions are discussed? More likely futures may be more detailed than less likely ones.

**Conclusion.**

This thesis has demonstrated that five different techniques that have been used successfully to detect lies about past activities or opinions can also be used to
detect lies about intentions at an above chance rate. However these techniques are not accurate enough to be used in practical situations, such as airports. This thesis has also shown that certain techniques (e.g. verbal cues and thermal imaging) are more accurate at classifying liars and truth tellers than others (e.g. non-verbal cues and reaction time analysis). This thesis suggests that the participants’ previous experience of the intention and when they plan to execute it might affect the techniques ability to detect lies. Overall, the conclusion of the thesis is that true and false intentions differ from each other in the way they are discussed and responded to, which means that it should possible to discriminate between them.
References


http://esotericonline.net/docs/library/Philosophy/History%20of%20Philosophy/0.%20General/General%20Histories/Bury%20-%20A%20History%20Of%20Freedom%20Of%20Thought.pdf


speakers. *Legal and Criminological Psychology*, advance online publication.


way to interview to detect deception. Forensic Update, 88, 25-29.


Appendix I