.

Verbal Lie-Detection using the Reality Monitoring Approach: An Analysis of its Effectiveness and Moderating factors

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor of Philosophy by Stamatis Elntib

September, 2015

TABLE OF CONTENTS

Page number

PART 2: Empirical work

PART 3: Discussion and Conclusions

LIST OF TABLES

LIST OF APPENDICES

Abstract

A large body of research findings suggests that verbal cues to deception can boost deception-detection accuracy rates to levels significantly above chance. This thesis examines the effectiveness of, and influences on, one of the most popular and widely used verbal lie-detection approaches, Reality Monitoring (RM). The RM approach has advantages not only in terms of its underpinnings in memory research and theory, but also its ease with regard to practical application. The RM approach assumes that deceptive verbal accounts, because they are artificial and not based directly on actual experience, differ from truthful accounts according to a variety of criteria (truthful accounts contain more vivid, spatial, temporal and affective information, etc.). However**,** so far, as in many other areas of lie-detection research, research in the area of RM has lacked methodological standardisation; consequently, we know little of the potential effects of contextual and other moderating variables on RM measures. In view of this, the primary aims of the present thesis were a) to assess whether, if conveyed in a standard format, the RM approach has any value overall in distinguishing between truthful and deceptive accounts, and b) if it does, to investigate the circumstances under which it might give optimal results; i.e. to assess some of the main factors which may moderate its efficacy in this respect. To these ends, six experimental studies were conducted which looked at truthful and deceptive accounts (generated by participants in the laboratory using video and autobiographical sources) and considered the effects of a number of different moderators; these included, first and second-language effects, modality (i.e. written vs. spoken discourse), absence/presence of others, demand characteristics effects, scoring systems (rating scales vs. raw frequencies) and standardisation for account

length. Overall, results indicated that in most studies there was evidence that total RM scores, as determined by the procedures applied here, successfully discriminated between truthful and deceptive accounts. However, results varied when RM criteria were considered individually, and when the influence of various moderators was assessed. For example, frequency measures of spatial and temporal information were found to be two of the most consistently effective diagnostic RM criteria. However, overall, RM was a more effective diagnostic tool before accounts were standardised for length; indeed, total RM scores failed to distinguish between truthful and deceptive accounts after accounts had been standardised for length. Also, the presence of others and modality (written or spoken) were two key moderator variables whose impact on total RM scores varied depending on whether or not the accounts were standardised for length. A number of other related variables were also considered; for example, truthful accounts were longer than deceptive accounts in both duration and length and the number of words produced per second was significantly greater for truthful accounts. Implications for research and practise are discussed; though perhaps most important in this respect was the finding that, despite the overall success of RM in discriminating truthful from deceptive accounts, RM criteria were not generally discriminating after standardisation for word-count or length. Moreover, a number of the moderators affected RM scores regardless of whether they were derived from truthful or deceptive accounts. This suggests that we may still be a long way from developing a method (such as the use of normative criteria) that could be used in the field to classify individual cases. Nevertheless, in the meantime, at the very least, the present results suggest that when judging the veracity of accounts using RM criteria, the scoring and other moderating variables identified in this thesis should be investigated systematically, and measured and

applied consistently, if researchers wish to compare and replicate findings within and across studies.

Acknowledgments

I need to express my gratitude to my supervisor Professor Graham Wagstaff. He has been the best mentor and supervisor a researcher could ever ask for. Spending time with and learning from him during the past years has been the most valuable lesson for me. He was always supportive and patient with me and I cannot thank him enough for his input, flexibility and encouragement.

My thanks must also go to my partner Ismini Lefa who has been my family, friend and soulmate during most of this journey. I am not sure if I could have made it without you but I am confident that this adventure would have been so much rougher without your love and support.

To my parents Γιώργο and Φωτεινή: I owe it all to you.

Preface

Research into the detection of human deception has tended to centre round three main approaches: the identification of non-verbal cues to deception, the examination of physiological indicators of deception, and the analysis of verbal cues to deception. As yet, none of the three approaches has been able to provide an exact method for differentiating between liars and truth-tellers; however, a growing body of research indicates that a focus on or use of the verbal cues and the content of spoken and written accounts, can improve the lie-detection ability. This thesis, therefore, further examines what is potentially one of the more promising of these verbal approaches, Reality Monitoring (RM), which is derived primarily from the theoretical framework created by Johnson and Raye (1981). Within the context of lie-detection, RM suggests that memories based on real experiences (assumed to be more likely in truthful accounts) will be more vivid, realistic and easier to reconstruct than imagined or fabricated experiences (as will occur more often in deceptive accounts). Truthful accounts will also contain more perceptual, spatial, temporal and affective information. On the other hand, because liars attempt to internally generate logically consistent information, they are more likely to include information regarding cognitive operations (thoughts and reasoning).

The RM approach has advantages not only in terms of its underpinnings in memory research and theory, but also its ease of practical application. However**,** so far, as in many other areas of lie-detection research, research in the area of RM has lacked methodological standardisation; consequently, we know little of the potential effects of contextual and other moderating variables that may influence RM

outcomes. Given these considerations, the primary aims of the present thesis were as follows:

1. To assess whether the RM approach, conveyed in a standard format, has any value overall in distinguishing between truthful and deceptive accounts.
2. And if it does, to investigate the circumstances under which it might give optimal results; i.e. to assess what factors moderate its efficacy in this respect.

With regard to the latter, potential moderating factors receiving particular emphasis in the present thesis were the language proficiency of story-tellers, modality employed (i.e. written vs. spoken discourse), absence/presence of others effects, scoring system applied (rating scales vs. raw frequencies), possible demand characteristics (knowledge of purpose of study), and standardisation procedures used to control for account length. However, at the same time, the opportunity was taken to investigate the utility of some other potentially related cues to deception, including duration and speech rate, which have previously received support as potentially useful cues to deception (Dilmon, 2009).

The thesis is divided into three main parts, Part 1, Introduction and Literature Review; Part 2, The Empirical work; and, Part 3, General Discussion and Conclusions.

Part I consists of four chapters which introduce and discuss different types of lie-detection and provide the rationale for emphasising verbal lie-detection and, in particular, the RM approach. Thus in Chapter 1, the definitional framework and the different types of lie-detection are discussed. In Chapter 2, the various verbal lie-detection techniques and their effectiveness are explained and compared. In Chapter 3, the historical and theoretical basis of RM, and the development, definitions and

applications of the RM criteria are described and explained. And finally, Chapter 4 discusses the possible influence of potentially important moderators in the application of the RM approach which have yet to receive systematic scrutiny.

Part 2 consists of seven chapters; one on introductory and methodological considerations, and six covering the empirical research studies. Thus Chapter 5 introduces the empirical research, outlines the research aims and respective hypotheses and summarises the six empirical studies and the methodology employed in the thesis. Chapter 6 describes Study 1, which considered the relative efficacy (i.e. ability to discriminate between truthful and deceptive accounts) of rating scales and raw frequency RM scoring systems among a group of untrained coders, and examined possible second-language effects on the efficacy of the RM approach. Chapter 7 describes Study 2, which examined the relative efficacy of spoken and written statements in relation to the RM approach, and assessed the viability of account length, duration and speech rate as cues to deception. Chapter 8 describes Study 3 which assessed the possible influence of demand characteristics (blind coding) on the coding and efficacy of RM criteria. Chapter 9 describes Study 4 which examined the effects of standardisation for length and duration of accounts on the efficacy of the RM approach. Chapter 10 describes Study 5 which was a reconsideration of the data from Study 1 using an analysis that allowed the calculation of RM scores before and after standardization for word-count. Finally, Chapter 11 describes Study 6 which investigated the possible effects of the number of people in the room when the account was given, on the efficacy of RM. This study also looked further at how standardising for account-length differences might affect the usefulness of RM using two types of standardisation: word-count and duration standardisation.

Part 3 consists of Chapter 12, which is a general discussion of the findings and conclusions. This chapter revisits the aims and hypotheses of the research, and the major findings are summarised and discussed in relation to these. Methodological developments are highlighted, and new theoretical approaches are suggested. Finally, limitations of the research are addressed, and possible directions for further research are considered.

Part 1

Introduction and Literature Review

"You shall not bear false witness against your neighbour" (Exodus 20:16)

Chapter 1

Deceptive behaviour: definitions of and approaches to deception-detection

Deceptive behaviour has been long been portrayed in a variety of cultures and religions as a sinful act that must be avoided (Underwood, 1993). Obviously lying may have different consequences depending on context; nevertheless, it is assumed that all humans know what it is to lie, hence we are all potentially in a position to detect a liar, and especially if the liar is someone we know very well (Vrij, 2008a). Moreover, although it is commonly accepted that lie-detection is difficult, research suggests that many believe that lie-detection skills can be boosted in those exposed to and trained in certain lie-detection techniques (Colwell, Miller, Miller & Lyons, 2006; Vrij, 2000). If it is, indeed, possible to train people in deception detection then this would obviously have very important implications for forensic investigations; hence the idea of developing and examining a relatively simple and practical lie-detection tool was one of the considerations guiding the present thesis. However, before reviewing the literature on deception detection in the legal area, it is obviously important to consider what exactly constitutes lying and deceptive behaviour.

1.1. Definitional framework

Any attempt to define the complex act of deceiving must take into account the context in which such an act is observed and the individual parties involved. Within

the broader spectrum of biological sciences, deception has typically been construed as an evolutionary act that involves a false communication which aims to benefit the communicator (Bond & Robinson, 1988). Importantly, within this context, deception does not necessarily imply a conscious intention to deceive; for example, complex biological phenomena such as fish and lizard camouflaging (Allen, Mäthger, Barbosa, Hanlon, 2009; Stuart-Fox & Moussalli, 2008) or mimicry and advergence amongst insects and carnivorous plants (Jurgens, El-Sayed & Suckling, 2009; Pekar & Kraal, 2002) can be considered to involve adaptive deception. However, it is only when humans are considered that the notion of intentionality of a deceptive act becomes of central importance.

For example, according to Goffman (1974), within the domain of human interactions, types of deception can be broadly categorised into exploitative fabrications and benign fabrications; exploitative fabrications are those that aim to serve the communicator and/or harm the message receiver, whereas benign fabrications are primarily aimed to promote the interests of the deceived; i.e. the intention is not to harm the message receiver. Within a similar framework, Lindskold and Walters (1983) have argued that the deceivers' motivation can be conceptualised along a six-point scale which ranges from altruistic motivations to exploitative motivations. Categorisations of deception can be further elaborated in terms of the roles that deceivers adopt and their associated behaviours. For example, in the case of computer-mediated communication, particular deceptive acts can be conceptualised and categorised as identity concealers, category deceivers (e.g. claiming of being a young female in a forum when, in fact, the deceiver is a man in late-adulthood), trolls (i.e. coming up with a completely invented character whose entire purpose is to provoke and disturb the message receivers), or identity

theft/impersonators (for an overview of computer-mediated deception, see Utz, 2005).

However, central to these conceptualisations is the idea that deception involves intention; nevertheless, intentionality cannot be assumed to apply to all human instances in which individuals might miscommunicate or tell an 'untruth' (Vrij, 2008a). For example, an old person with poor memory skills or a schizophrenic patient who genuinely experiences stimuli that do not exist would not normally be classified as a 'liar'; indeed, if we adopt the Oxford English Dictionary's (OED) definition of a lie as "to make a false statement with the intention to deceive", then technically, they cannot be liars. For purposes of clarity, and to narrow down the focus of the investigation, therefore, within the present thesis, the definition of 'deception' given by Zuckerman, DePaulo, and Rosenthal (1981) will be adopted; i.e. deception or deceiving is "an act that is intended to foster in another person a belief or understanding which the deceiver considers false" (p. 3). In this sense, therefore, if we accept the OED definition of lying, lying and deception essentially involve the same thing; i.e. a deliberate attempt to mislead.

1.2. Types of lie-detection

The development of deception or lie-detection techniques in forensic settings has generally involved three main approaches: the detection of non-verbal cues to deception, the examination of physiological indicators of deception, and the analysis of verbal cues to deception (Granhag & Stromwall, 2004; Vrij, 2008a; 2008b). The different methods currently used by lie detectors, both practitioners and researchers, are based on the principle that the mental states of the liars are distinct from those of the truth-tellers. These differences in turn are assumed to affect the subjects'

expression of verbal and non-verbal behaviours and their psycho-physiological functioning patterns allowing a discrimination between truth-tellers and liars (Vrij & Granhag, 2007). Unfortunately, however, as yet, none of the three approaches has been able to provide an exact differentiation between liars and truth-tellers. Hence, even the most reliable cues to deception are only probabilistically, and rather weakly, linked to deceptive behaviour (Granhag & Stromwall, 2004; Vrij & Granhag, 2007).

Due to a variety of influences, including outdated and non-empirically based Police manuals, a number of what are considered by many researchers to be 'myths' have arisen concerning the ability of a range of cues to indicate deception (Colwell et al., 2006; Vrij & Granhag, 2007; Vrij, Granhag & Porter, 2010). Indeed, a substantial body of research has indicated that the use of these cues can sometimes drop accuracy rates below chance level, thus actually diminishing the ability to detect lies (Vrij, 2008a; 2008b).

The most common alleged myths about the behaviours of liars concern their non-verbal behaviours. For example, at various times it has been suggested that liars are likely to display more foot and leg movements (Gordon & Fleisher, 2002); avoid eye contact (Gordon & Fleisher, 2002; Inbau, Reid, Buckley, & Jayne, 2001), and display overall nervous behaviours such as fidgeting (Rabon, 1992) and touching of the nose (Gordon & Fleisher, 2002), ears and eyes (Macdonald & Michaud, 1992). Indeed, such ideas can still be found in a number of widely-used investigative interview approaches such as the Reid Approach (Inbau et al., 2001; Vrij, 2008b; Vrij, Mann, Kristen, & Fisher, 2007). Such ideas conflict with many empirical findings; for example, liars sometimes move less whilst lying, since they consciously try to control their behaviour in order to appear calm and genuine; they

also may engage in more eye-contact to counteract the popular assumption that gaze aversion is a feature of lying (Bond & DePaulo 2006; De Paulo Lindsay, Malone, Muhlenbruck, Charlton, & Cooper, 2003). Consequently, as yet, there appears to be a reasonable consensus amongst researchers that there is still no established, definitive collection of non-verbal cues that can be used to reliably discriminate liars from non-liars across a variety of contexts.

Of course, the fact that there is ambiguity surrounding the usefulness of the non-verbal lie-detection approaches does not mean that that future research on the topic is not likely to be fruitful. Indeed, a number of researchers have emphasized that the efficacy of non-verbal indicators has yet to be fully explored in 'high stakes' field-based research, and it is possible that some (including gaze aversion) may yet emerge as significant predictors in more real-life contexts (Wright-Whelan, Wagstaff & Wheatcroft, 2014). Also, some indicators, such as eye-blinks, may emerge as predictors only in specific circumstances; such as when cognitive load is applied (Vrij et al., 2008). Nevertheless, the perhaps unwarranted dominance of non-verbal-behaviour in the lie-detection literature has been highlighted numerous times (Vrij, 2008a; 2008b; Vrij et al., 2010).

Despite the continuing fascination of non-verbal cues to deception amongst researchers and practitioners, perhaps the most popular lie-detection approaches used in the field involve physiological measures; in particular, techniques using the 'polygraph'. Of these, the most characteristic and promising example is the Guilty Knowledge Test (GKT; Ben Shakhar & Elaad, 2003; Lykken, 1959, 1960; Honts, 2004; Vrij & Granhag, 2007). The GKT records physiological responses such as respiration, blood pressure and sweating of the fingers. It is the most frequently used polygraph–based test in Japan (5000 GKT per year), and its results are sometimes

used there as evidence in courts (Nakayama, 2002). The GKT technique is based on the assumption that guilty subjects' physiological responses will fluctuate, largely because of changes in anxiety and arousal, when they are presented with both crime-related and crime-unrelated features. If the test is designed meticulously, and the crime-related information is sufficient, so that a series of numerous crime-related questions can be formulated, then there is some evidence that the approach can discriminate significantly between truth-tellers and liars in both field (Bem-Shakhar & Elaad, 2003; Elaad , 1990; Elaad, Ginton, & Jungman , 1992) and laboratory studies Ben-Shakhar & Furedy, 1990; Ekman, 2001). However, critics have pointed out that, even if sometimes effective, the approach is limited in its practical application. For example, consensual and non-consensual sex cannot be discriminated since both innocent and guilty subjects are familiar with the case-related features. Similarly, it is difficult to apply the approach reliably to group-crimes (e.g. gang-rapes; group robberies and terrorist plots), to assign roles (i.e. leader, peripheral player, etc.) and degree of involvement to guilty suspects. In addition, the approach is based on the assumption that guilty subjects will be familiar with the questions and their answers and innocent subjects will not. This is logistically impossible to achieve in some cases as the crime-related details might have been released in the press or the information known to the Police might be so limited that it is impossible to design the test (Ben-Sakhar & Elaad, 2003; BPS, 2004). The GKT is also intrusive and it can be used only with the examinees' knowledge, hence it gives the suspect the chance to develop countermeasures (Bem-Shakhar & Elaad, 2003; Vrij & Granhag, 2007).

Other examples of physiological detection techniques include Voice Stress Analysers (VSA) and Thermal Imaging (TI). VSA detects voice indices such as

pitch, intensity, frequency, or micro tremors whilst TI detects temperature changes around the eye. The US National Research Council (2003) has suggested that, compared to standard polygraph measures, these techniques may be less intrusive, and more sensitive and effective indicators of physiological arousal accompanying lying. However, they are also more sensitive to environmental distractions and have yet to be empirically validated. At the moment, therefore, considering that the GKT is the only questioning method widely accepted by polygraph supporters themselves (Ben Shakhar & Elaad, 2003; BPS, 2004; Lykken, 1959, 1960) and that due to problems of distraction, VSA and TI cannot be reliably employed in conjunction with established questioning techniques, neither is considered sufficiently reliable to be used as evidence in court in the USA and UK (National Research Council, 2003; Nakanishi & Imai-Matsumura 2008; Nozawa & Tacano 2009; Tanaka, Ide,& Nagashuma, 2000).

The third major approach to lie-detection involves the use of verbal cues. As noted previously, traditionally, in the lie-detection role, non-verbal behaviours have tended to attract more attention from both researchers and lay people than verbal behaviours. Nevertheless, as alluded to previously, some evidence suggests that over-reliance on non-verbal cues, such as those ostensibly related to 'nervousness', may actually inhibit accurate detection (Colwell et al., 2006; Vrij, et al., 2010). In contrast, research clearly indicates that a focus on or use of the verbal cues and the content of the subjects' accounts, is more likely to boost accuracy rates than looking for non-verbal indicators of deceit (Bond & DePaulo, 2006; Vrij et al, 2010). It can be noted also that verbal lie-detection measures are also potentially more simple, and relatively easy and cheap, to test and implement than physiological measures which

require expensive equipment. Moreover, they are not invasive since audio or video transcripts can be analysed retrospectively and without the subjects' knowledge.

1.3 Conclusion

Given the points and issues raised in this chapter, it was decided to concentrate on verbal indicators of deception in the present thesis, so the next chapter attempts to explore this aspect of the detection of deception in more detail.

Chapter 2

The detection of deception through verbal behaviour

As noted in the previous chapter, although the focus of lie-detection research is broadly on non-verbal behaviour, there is growing interest in and appreciation of the verbal content of an account. Indeed, analysis of verbal behaviour may potentially be able to provide a way of detecting deception that is both simpler and cheaper than psychophysiological techniques, and more valid and efficient than the use of non-verbal cues. The aim of this chapter, therefore, is to overview and evaluate in more detail the different methods of verbal deception detection and their respective usefulness.

2.1. Using verbal cues to detect deception

A variety of studies have shown that analyses of verbal cues to deception can boost deception-detection accuracy rates above the chance level (Vrij, 2008b, Vrij, 2000; Vrij & Akehurst, 1998). Supporting evidence has been found in experimental (Porter & Yuille, 1996; Vrij, Edward, Roberts, & Bull, 2000), field (Adams & Jarvis, 2006; Mann, Vrij & Bull, 2002; Vrij & Mann, 2001a; 2001b) and meta-analytical studies (De Paulo et al., 2001; Masip, Sporer, Garrido & Herrerol, 2005; Vrij, 2015).

2.1.1. *Different verbal lie-detection approaches*

Research on verbal lie-detection involves two broad categories of studies: the first category concerns methods which involve an empirically and/or theoretically clustered collection of verbal deception-detection cues (Masip et al., 2005; see Vrij, 2008b for an overview). Within this category are a limited number of techniques which have received a degree of empirical support, such as Reality Monitoring (Masip et al., 2005; Sporer, 2004; Vrij, 2008b) and Criteria-based Content Analysis (Vrij, 2005, 2008b). A second subset of approaches within this category involves techniques which are less generally accepted by the academic community, (Vrij, 2008a; Vrij, Mann, & Fisher, 2006); these include techniques such as Scientific Content Analysis (Driscoll, 1994; Smith, 2001) and the Behavioural Analysis Interview (Horvath, Jayne, & Buckley, 1994). Even more obscure are techniques such as Investigative Resource Analysis, Verbal Behaviour Analysis, and Lexical Diversity (Adams & Jarvis, 2006).

The second category of techniques (e.g. Adams & Jarvis, 2006; De Paulo et al., 2003) concerns the use of more specific verbal and paralinguistic cues to detect deception on an individual basis. Cues which have received some empirical support as potential cues to deception include the use of equivocation; e.g. I think, I believed it to be, it was kind of, etc.(Adams & Jarvis, 2006; ten Brinke & Porter, 2012; De Paulo et al., 2003; Zuckerman et al., 1981), speech errors such as repetitions (Harpster, Adams & Jarvis, 2009; Vrij & Mann, 2001a), pauses (Mann et al., 2002; Vrij & Mann, 2001a), and short account length (Colwell, Hiscock-Anisman, & Memon, 2002; Elntib, Wagstaff & Wheatcroft, 2014; Memon, Fraser, Colwell, Odinot, & Mastroberardino, 2010; Vrij, Akehurst, Soukara & Bull, 2004). However, as yet these cues have not been integrated into a systematic technique for practical

application. In sum, perhaps the four best-defined, systematic, lie-detection approaches which include, or focus exclusively on, verbal-cues to deception are Reality Monitoring (RM), Criteria-Based Content Analysis (CBCA), Scientific Content Analysis (SCAN), and the Behaviour Analysis Interview (BAI). These are further detailed as follows.

2.1.2. *The mechanics and accuracy of verbal lie-detection approaches: an overview*

The Behaviour Analysis Interview (BAI) utilises both verbal and non-verbal cues to differentiate between liars and truth-tellers. It is widely applied in the United States and it is very influential in shaping beliefs regarding deceptive behaviour in both practitioners and lay people. Most importantly, it is the only published interview protocol that incorporates a lie-detection approach within it (Vrij et al., 2006). According to this technique, truth-tellers and liars will show different patterns of verbal and non-verbal behaviour across the interviewing session which consists of 15 standardised questions, each one assumed to discriminate deceptive from truthful suspects and witnesses on the basis of their verbal and non-verbal cue-responses. However, despite BAI's popularity and wide use in the US, it is extremely controversial and, according to many researchers, is based on a number of empirically unfounded assumptions (Vrij, et al, 2006). These include, for example, notions such as that, "liars are less evasive about the purpose of the interview, less emphatic in their denial of having committed the crime, more likely to deny any knowledge of who the culprit might be, less likely to name another suspect, less likely to admit that a crime has taken place and that there was the opportunity to commit the crime, more likely to voice negative feelings towards the interview

because innocent suspects have faith that they will be exonerated, less likely to admit to having thought about committing a crime similar to that under investigation, less likely to give a reasonable motive for the crime because guilty suspects do not want to reveal their own motives, less likely to suggest a serious punishment for the person who committed the crime, more likely to suggest that the guilty person should be given a second chance, more likely to answer in the third person, less likely to have informed their loved ones about the interview, being less confident in being cleared" (Vrij, et al, 2006, p.330).

The only published supportive work on BAI appears to be a study by Horvath, Jayne and Buckley (1994) in which videos of 60 real interviews with suspects were independently assessed by four observers using the BAI criteria. The researchers reported accuracy rates as high as 91% for truthful and 80% for deceptive suspects. However, although showing some apparent support for the technique, the study was subject to a number of methodological criticisms that weakened the acceptance of BAI by the global scientific community (Vrij et al., 2006). For example, critics argued that Horvath et al. used weak criteria to establish ground-truth (i.e. the actual truthfulness-status of the accounts-whether it was truthful or deceptive) namely confessions and suspects' personal details and characteristics such as biographical information and motivation. Critics also argued that the approach is inflexible since it treats all suspects similarly without taking into account individual vulnerabilities such as the developmental maturity of the suspect (Kostelnik & Reppucci, 2009; Redlich, 2007; Weinstock & Thompson, 2009). Indeed, since Horvath et al.'s study, several experimental studies have appeared to provide strong evidence against the assumptions of BAI (Kassim & Fong, 1999; Mann, Vrij & Bull, 2004; Vrij et al, 2006). For example, BAI postulates that liars

will be less emphatic in their denial of having committed the crime; however, the evidence suggests that truth-tellers often consider that their innocence will be obvious to others, i.e. they have an *illusion of transparency*; therefore, they are less likely to have a strategy and are likely to be less emphatic than the liars (Vrij, et al., 2006). In contrast, liars are more keen to be believed (Hartwig, Granhag, Stromwall, & Doering, 2010).

The Scientific Content Analysis or SCAN approach has been widely applied as a lie-detection tool in a number of countries including Australia, Belgium, Canada, Israel, Mexico, UK, US, the Netherlands, Qatar, Singapore, and South Africa and respective training sessions are delivered on a weekly basis in both Canada and the US (Nahari, Vrij & Fisher, 2012; Vrij, 2008). It is recognised as a cheap and practical technique (Driscoll, 1994; Sapir, 1987; Smith, 2001). The suspect or witness is first asked to hand-write his/her account on a piece of paper and the accounts are then analysed using a list of predetermined criteria some of which are presumed to more likely occur in deceptive accounts (e.g. a change or omission of pronouns, spontaneous corrections, use of emotional language near the peak of the story, not denying the allegations in the account, expressing lack of memory). In addition, the SCAN is one of the few approaches that include a number of criteria that are presumed to more likely occur in truthful accounts (e.g. appropriate use of pronouns, lack of spontaneous corrections, use of emotional language throughout the story, directly denying the allegations within the statement, lack of memory lapses). Unfortunately, one of only two studies reportedly finding support for this approach, i.e. Smith (2001), does not attempt to define or to classify the criteria beyond what is available in Sapir's (1987) original manual, which cannot be widely accessed online. Moreover, the other supportive study by Driscoll (1994)

admits that the list of criteria outlined in the paper is not exhaustive, being based on a limited range of criteria used most often in training. The accuracy rates reported by the two studies are remarkably high ranging between 65-95% (Driscoll, 1994; Smith, 2001); indeed, the criterion denial of allegation alone discriminated truth-tellers from liars in 80% of the statements. However, both studies have again been criticised for conceptual and methodological shortcomings. Again, ground-truth was not established for all statements; also, the theoretical underpinnings of the technique are not clear (Vrij, 2008a); neither Sapir nor other proponents of the technique have provided a theoretical rationale for why certain behaviours should be indicative of deception. Perhaps most importantly, Porter and Yuille (1994) failed to find any empirical support for the SCAN technique (although they used transcribed statements of oral testimonies rather than written statements per se as advised by Sapir).

The next approach, Criteria-Based Content Analysis (CBCA) has not only been the subject of extensive research, but has been practically applied in criminal investigations and in court proceedings in a few countries (Sweden, Germany, The Netherlands and Switzerland), particularly for testing accounts derived from child-witnesses/victims of (sexual) abuse (Vrij, 2008a; 2008b). The CBCA approach is based primarily on the assumption that "truthful, reality-based accounts differ significantly and noticeably from unfounded, falsified, or distorted stories" (Undeutsch, 1984, p 44); hence genuine statements will be different in terms of a number of content criteria from deceptive statements (Steller & Kohnken, 1989). The CBCA, therefore, involves the application of criteria for distinguishing truth-telling from lies derived actuarially (i.e. post hoc) from an analysis of a large volume of children's testimonies (Kohnken, 2004). The CBCA constitutes a core component

of a wider verbal-lie-detection approach known as Statement Validity Analysis (SVA), and it includes 18 criteria (for an overview see Table 2.1) which are focussed on cognitive and motivational precursors of deceptive behaviours; that is, it is assumed that deception occurs as a result of both cognitive and motivational factors (Vrij, 2008b).

Table 2.1. The Content Criteria for Statement Validity Analysis (from Raskin & Esplin, 1991; Vrij, 2008a)

General Characteristics
1. Logical Structure: Is the statement coherent? Do the different segments fit together? (Note: Peculiar or unique details or unexpected complications do not diminish logical structure)
2. Unstructured Production: Are the descriptions unconstrained? Is the report somewhat unorganized? Are there digressions or spontaneous shifts of focus? Are some elements distributed throughout? (Note: This criterion requires that the account be locally consistent.)
3. Quantity of Details: Are there specific descriptions of place and time? Are persons, objects, and events specifically described? (Note: Repetitions do not count.)
Specific Contents
4. Contextual Embedding: Are events placed in spatial and temporal context? Is the action connected to other incidental events, such as routine daily occurrences?
5. Interactions: Are there reports of actions and reactions or conversation composed of a minimum of three elements involving at least the accused and the witness?
6. Reproduction of speech: Is speech and conversation during the incident reported in its original form? (Note: Unfamiliar terms or quotes are especially strong indicators, even when attributed to only one participant.)
7. Unexpected Complications: Was there an unplanned interruption or an unexpected complication or difficulty during the sexual incident?
8. Unusual Details: Are there details of person, objects, or events that are unusual, yet meaningful in this context? (Note: Unusual details must be realistic)
9. Superfluous Details: Are peripheral details described in connection with the alleged sexual events that are not essential and do not contribute directly to the specific allegations? (Note: If passage satisfies any of the specific criteria 4–18, it probably is not superfluous).
10. Accurately Reported Details Misunderstood: Did the child correctly describe an object or event but interpret it incorrectly?
11. Related External Associations: Is there reference to a sexually toned event or conversation of a sexual nature that it is related in some way to the incident but is not part of the alleged offences?
12. Subjective Experience: Did the child describe feelings or thoughts experienced at the time of the incident? (Note: This criterion is not satisfied when the child responds to a direct question, unless the answer goes beyond the question.)
13. Attribution of the Accused's Mental State: Is there reference to the perpetrator's feelings or thoughts during the incident? (Note: Descriptions of overt behavior do not qualify.)
Motivation-Related Contents
14. Spontaneous Corrections or Additions: Were corrections offered or information added to material previously provided in the statement? (Note: Responses to direct questions do not qualify.)
15. Admitting Lack of Memory or Knowledge: Did the child indicate lack of memory or knowledge of an aspect of the incident? (Note: In response to a direct question, the answer must go beyond "I don't know" or "I can't remember.")
16. Raising Doubts About One's Own Testimony: Did the child express concern that some part of the statement seems incorrect or unbelievable?
17. Self-deprecation: mentioning personally unfavourable, self-incriminating details: "Obviously it was stupid of me to leave my door wide open because my wallet was clearly visible on my desk"
18. Pardoning the perpetrator: making excuses for the perpetrator or failing to blame him or her, such as a girl who says she now feels sympathy for the defendant who possibly faces imprisonment

In terms of cognitive factors, it is assumed that criteria 1-13 (i.e. logical structure, unstructured production, quantity of detail, contextual embedding, interactions, reproduction of speech, unexpected complications, unusual details, superfluous details, accurately reported details misunderstood, related external associations and attribution of accused's metal states) are more likely to be found in genuine accounts. The theoretical assumption here is that these criteria apply to the production of a "natural" account where the subject has cognitive access to a full range of memories, including both target and rich contextual information.

Motivationally influenced behaviours indicated by criteria 14-18 (i.e. spontaneous corrections or additions, admitting lack of memory or knowledge, raising doubts about one's own testimony, self-deprecation, pardoning the perpetrator) are also assumed to be found more often in truthful accounts, the underlying theoretical assumption being that truth-tellers will be less likely to attempt to display what they perceive to be stereotypically genuine behaviours; i.e. they feel less of a need to give a 'good impression to their audience in order to be believed (De Paulo et al., 2003; Vrij, 2008a; 2008b).

On the whole, empirical tests of the CBCA have been encouraging; for example, meta-analytical studies have reported accuracy rates ranging from 65-95% (Vrij, 2005, 2008a; 2008b). Some researchers have suggested that the approach can only be fully applied to child-abuse cases, since the criteria were primarily derived from this source (Honts, 1994; Horowitz, Lamb, Esplin, Boychuk, Krispin, & Reiter-Lavery, 1997). However, others have argued that at least some aspects of

CBCA, such as quantity of details and contextual embedding may be applied to adult testimonies (Porter & Yuille, 1996; Steller & Kohnken, 1989).

The final approach to be considered here, Reality Monitoring (RM), has been the subject of a reasonable amount of empirical research, but it has yet to be applied in practice. According to the RM theoretical framework, memories based on a real experience will be qualitatively different from memories based on fiction (Johnson & Raye, 1981). As with SCAN and CBCA, the RM approach retrospectively applies a number of predetermined criteria to verbal accounts to differentiate between truthful and deceptive accounts. From its inception, the RM approach has been deeply rooted in memory theory and research (Johnson & Raye, 1981). Underpinning the approach is the idea that, again because the truth-teller has access to a wider range of relevant target and contextual information, truthful accounts will be more vivid, more realistic and easier to reconstruct than deceptive accounts. They will also contain more perceptual, spatial, temporal and affective information. On the other hand, because liars attempt to internally generate logically consistent information, their accounts are more likely to include information regarding Cognitive Operations (thoughts and reasoning); for a description of the criteria according to Vrij (2000; 2008; 2015), see Table 2.2.

A variety of studies have shown that RM criteria can significantly differentiate between lying and truth telling, with accuracy rates typically ranging between 66-72%, though rates beyond 80% have also been reported (see, for example, Masip et al., 2005; Sporer, 2004; Vrij, 2008a; Vrij et al., 2004). However, although the RM approach is widely accepted as potentially one of the most, if not the most effective tool for verbal lie-detection, a number of researchers have argued that the criteria are still not sufficiently defined, which makes measurement difficult, particularly if the

technique is to be applied in the field. Indeed, as yet, there is no standard procedure for measuring the criteria (Granhag, Strömwall & Landström, 2006; Vrij, 2008a). Also, the evidence suggests that whilst some criteria (e.g. temporal information) in the RM technique seem to be particularly effective in discriminating between truth and deception, others (e.g. cognitive information) are not (Masip et al., 2005; Vrij, 2008a). It would, therefore, make sense to break the technique down into its primary elements and to test them individually under different conditions to maximise the technique's potential as a diagnostic tool.

Table 2.2 The Reality Monitoring Criteria (from Vrij, 2000; 2008)

1. Perceptual Information: the presence sensorial experiences such as sounds (e.g. 'he really shouted at him') or visual details (e.g. 'I saw him entering the room').
2. Temporal information: the presence of information about when the event happened (e.g. 'it was early in the morning') or explicitly describing a sequence of events (e.g. 'as soon as the guy entered the pub the girl started smiling').
3. Spatial information: the presence of information about locations (e.g. 'It was in a park') or the spatial arrangement of people/objects (e.g. 'the man was sitting left from his wife' or 'the lamp was partially hidden behind the curtains').
4. Remembered feelings (affect): how well the person remembers feelings (accounts of subjective mental states) from the event (e.g.' Joseph was very scared').
5. Cognitive operations: evidence in the narratives of various cognitive activities, such as thoughts or reasoning (e.g 'I must have had my coat on, as it was very cold that night') and cognitive suppositions of sensory experiences (e.g. 'She seemed quite clever'). This criterion also includes descriptions of inferences made by the participant at the time of the event (e.g. 'it made me think at the time how nice it could be if I have never been there').
6. Realism: This criterion is present if the story is plausible, realistic and makes sense.
7. Vividness/Clarity: this refers to the clarity and vividness of the statement. This criterion is present if the report is vivid, lively, clear and sharp instead of dim, faint, vague and indefinite.
8. Reconstructability: possibility to reconstruct the event on the basis of the information given.

2.1.3. *Verbal lie-detection approach comparisons*

In light of the previous considerations, despite having a measure of popular support, the BAI and SCAN techniques have generally considered by the scientific community to be inferior, both in terms of reliability and construct and empirical validity to the Criteria-Based Content Analysis and the Reality Monitoring approaches (for reviews see Bogaard, Meijer & Vrij, 2014; Granhag et al., 2006; Masip, et al., 2005; Vrij, 2008a, 2008b; Vrij et al, 2000); thus only the CBCA and RM approaches appear consistently to receive positive reviews claiming that their accuracy levels are significantly above chance (Bogaard et al, 2014; Vrij, 2008a, 2008b). Indeed, sometimes the predictions of the various approaches can go in opposite directions. For example, the SCAN suggests that liars are more likely to spontaneously correct themselves and to express lack of memory for an event and more likely to describe an event without sticking to its timeline and structure. These (empirically unsupported) assertions are in direct contradiction to the assumptions of the CBCA which suggest that these elements are more likely to be found within truthful accounts as a result of the different impression management strategies employed by liars and truth-tellers.

Nevertheless, both CBCA and RM techniques have relative strengths and weaknesses, and a number of studies have attempted to compare the two (Masip et al., 2005; Sporer, 1997; Vrij et al.,, 2000). For example, Masip et al. (2005) highlighted the theoretical superiority of the RM approach whilst emphasizing the weak theoretical foundations of CBCA. In particular, as noted previously, CBCA's principles were derived actuarially from child-interviewers' practical experience rather than from an existing theory. On the other hand, the RM framework has its roots in existing and extensively reviewed theory and research on memory.

Significantly, Johnson and Ray3's (1981) seminal theoretical perspective regarding genuine and imagined experiences has been adopted and used by numerous researchers from diverse scientific domains outside the field of lie-detection and it has received strong empirical support and nearly 2000 citations. In addition, the RM approach is easier to use as it has fewer criteria and more clear-cut definitions; hence it is considered relatively less time-consuming and generally more cost-effective in terms of interviewer training and administration (Sporer, 1997; Vrij, 2000). It has also been noted that CBCA is usually presented as a truth-detection rather than a lie-detection technique; hence all of the criteria are designed to indicate truthfulness rather than deception (Stromwall, Bengtsson, Leander & Granhag, 2004). Whereas RM includes a criterion (i.e. cognitive information) designed to increase suspicions that the account is deceptive (Vrij, 2008a).

In terms of empirical comparisons, Granhag et al. (2006) found that although two of the CBCA criteria, namely logical structure and quantity of details, had discriminative power in a lie-detection task, the total CBCA scores for each statement could not reliably distinguish between truthful and fabricated accounts. In contrast, RM criteria scores (including the total RM scores) were effective in discriminating the truthful from the deceptive accounts. Similar results have been obtained in other studies (Porter & Yuille, 1996; Stromwall et al., 2004). Also, inter-rater agreement comparisons between RM and CBCA indicate that there is overall higher agreement between raters using the RM approach (Sporer, 1997; Stromwall et al, 2004; Vrij et al., 2000; Vrij et al., 2004a).

2.2. Conclusions

To sum up, it appears that very few approaches to verbal lie-detection have been used in the field by practitioners, and of these CBCA has overall appeared to receive most positive reviews from both researchers and practitioners. RM has also received positive reviews from some researchers, but has yet to be applied in the field. However, CBCA is arguably weaker than RM in theoretical terms, since its formulation was not theory-driven like the RM approach. CBCA is also a more labour-intensive approach than RM, and is not more diagnostic of deception; indeed, one of CBCA's strongest criteria (i.e. contextual embedding) is also included in the RM criteria (spatial and temporal information); therefore, one might claim that RM is not only simpler, but it also contains one of CBCA's most powerful criteria.

With these considerations in mind, Reality Monitoring was selected as the main verbal deception technique for investigation in the present thesis.

Chapter 3

The Reality Monitoring approach

Having presented a rationale for focussing on verbal deception-detection cues using a Reality Monitoring approach in the present thesis, the objective of this chapter is to look again, in more detail, at the historical background of the RM approach and to expand upon the definitions and use of its criteria.

3.1. Historical and theoretical basis of RM

As mentioned in the previous chapter, like CBCA, RM rests on the assertion that genuine and fabricated memories will be quantitatively and qualitatively different. However, in the case of RM, the theoretical underpinnings are derived directly from the work of Johnson and her colleagues (Johnson & Raye, 1981; Johnson, Foley, Suengas, & Raye, 1988, Johnson, Hashtroudi, & Lindsay, 1993; Johnson & Suengas, 1989) regarding the origin of memories. To reiterate, according to Johnson et al.'s seminal RM theoretical framework, memories based on a real experience will be qualitatively different from memories based on fiction (Johnson & Raye, 1981). Memories of real events are obtained through perceptual processes, therefore, they are more likely to contain perceptual (e.g. sounds, colours, details of smell), and spatiotemporal details (e.g. details regarding the spatial arrangement of people or the time order of events) than imagined or fabricated memories (Sporer, 2004; Vrij, 2000). In contrast, memories about fabricated events are internally-derived; hence they are more likely to contain information regarding cognitive operations (e.g. thoughts, reasoning, and cognitive suppositions of sensory experiences).

Interestingly, the initial paradigms upon which this theor7 was tested were not related to lie-detection at all, rather they were aimed at exploring how and whether participants could separate internally (i.e. imagined) and externally (i.e. experienced) generated stimuli. The core notion of the original theory was that internally and externally derived memories were generated by two different mechanisms (Johnson & Raye, 1981). In particular, externally generated (i.e. experienced) memories involved sensorial (i.e. colours, sounds, etc) and contextual (i.e. details about the time and space of the event) information during the encoding-phase of the experienced event; therefore, recalling these memories will be more likely to also include such information. In contrast, imagined events will be richer in information reflecting cognitive operations such as reasoning, attribution of mental states to others and logical assumptions. This makes sense since no external information is available during the synthesis of the imagined event which is primarily a product of internal processes (Johnson et al, 1993; Johnson & Suengas, 1989).

It was soon recognised that the above assumptions might have implications for lie-detection research since one would expect that fabricated events, like imagined ones, would contain less sensorial and contextual information and more cognitive information. These ideas were, therefore, subsequently tested in lie-detection paradigms by Alonso-Quecuty and her colleagues, and later others, in a series of studies which found broad support for the value of RM in distinguishing between truthful and deceptive reports (Alonso-Quecuty, 1992, 1995; Barnier, Sharman, Mckay & Sporer, 2005; Masip et al., 2005).

It should be cautioned, however, that although most of the published literature indicates that at least some of the criteria used by researchers might reliably distinguish truthful from deceptive accounts, the bulk of the original

research literature on the use of RM as a lie-detection tool was initially conducted in Spanish and later in German and French. This in itself is not necessarily problematic, however, it is sensible to consider whether subsequent researchers were aware of the possible limitations of the original research. The first published translated summary of the research of Alonso-Quecuty and colleagues came as late as in 2005, so researchers might not have been aware that the first published study from Alonso-Quecuty (1992) involved only 22 participants with eleven participants per condition; moreover, in none of the studies published in Spanish by Alonso-Quecuty and colleagues was there any report of an inter-rater reliability analysis or definitional framework that future researchers could copy and use (Masip et al., 2005).

3.2. The Reality Monitoring criteria: development and definitions

Since the aforementioned studies of RM by Alonso-Quecuty and colleagues in the early 1990s, the RM approach has widely become accepted as potentially one of the most effective tools for verbal lie-detection (Masip et al., 2005; Sporer, 2004; Vrij, 2000, 2008; Vrij et al., 2004); nevertheless, despite its popularity, researchers have yet to agree upon the exact definitions and numbers of Reality Monitoring criteria that can usefully be applied (Granhag et al, 2006; Vrij, 2008). An indication of the fluid operationalization of the criteria can be seen in the gap between the original papers by Johnson and Raye (1981) and those who have since extended and developed the criteria for lie-detection purposes, such as research from Alonso-Quecuty et al., and Vrij, Sporer, and Granhag. For example, in their original paper, Johnson and Raye (1981) proposed four criteria, namely contextual, sensory, semantic information and cognitive operations. Alonso-Quecuty subsequently

operationalized the criteria in Spanish but, as noted in the previous section, gave little information in terms of their standardised definitions. Research into RM and lie-detection then proliferated in a variety of countries including Germany (Sporer & Hamilton, 1996), Canada (Porter & Yuille, 1996), Finland (Santtila, Roppola & Niemi,. 1998), France (Biland et al., 1999; cited in Masip et al., 2005), Sweden (Granhag et al., 2006) and the UK (Vrij, 2000). Nevertheless, only a few papers/textbooks have purported to provide a finite list of the criteria with clear definitions (e.g., Vrij, 2000; 2008; 2015). Indeed, it is somewhat remarkable that the RM procedures appear to have been successful in discriminating between truths and lies when there is no obvious consensus on the number of criteria to be applied and their definitions.

For example, contextual information has been reclassified as temporal and spatial information in the majority of studies conducted during the past 20 years (Elntib et al., 2014; Roberts & Lamb, 2009; Sporer, 1997), although the more inclusive term *contextual information* has still been used in some (Masip et al., 2005). In the same way, sensory or perceptual information has variously been classified as a single criterion (e.g.. Barnier et al, 2005; Elntib et al., 2014; Sporer, 1997; Sporer & Sharman, 2006), broken down into visual and audio (e.g.. Granhag, Stromwall & Olsson 2001; Vrij et al., 2000; Vrij et al, 2001), perceptual information about objects and people (i.e. Roberts & Lamb, 2010), and into visual, audio and smell(s) (Granhag et al., 2006). In yet other studies, sensory information is listed as a separate criterion from visual and auditory information (Stromwall & Granhag, 2005).

The number of criteria being used has also varied depending on the research group under investigation, and occasionally there have even been variations within

research groups. For example, the criteria of reconstructability and realism have been used by many (Santtila, at al., 1998; Sporer, 1997; Sporer & Hamilton, 1995; Sporer & Sharman, 2006) but not all researchers (e.g. Barnier et al, 2005). Similarly there are studies where some of Johnson et al.'s core criteria, such as cognitive (i.e Vrij, Edwards, & Bull, 2001) and spatial (i.e. Granhag et al., 2001) information have not been included at all. Other criteria such as rehearsal, actions and complexity have variously been included by researchers depending on their conceptualisation of the RM approach (e.g. Roberts & Lamb, 2010). Notably also, studies employing frequency measures of RM criteria rather have tended to ignore global criteria such as realism, reconstructability and clarity/vividness (e.g. Elntib et al., 2014; Memon et al, 2010; Otgaar, Candel, Memon, & Almerigogna, 2010; Stromwall & Granhag, 2005).

These variations and ambiguities obviously create difficulties for any study of RM. However, Vrij's (2000) classification of eight criteria, with their associated definitions shown in Table 2.2 in the previous chapter, provides a reasonably clear exposition of RM components that is concise yet comprehensive, and fits with criteria used in previous studies. Hence it was Vrij's classification that was used in the present thesis; i.e. Vividness, Perceptual Information; Spatial information; Affective information; Reconstructability; Realism; Temporal Information and Cognitive Operations.

3.3 Limitations of RM theory

However, within this context, it is also worth mentioning that questions have been raised about the degree of overlap between the original theory of Johnson et al. and the lie-detection approach derived from Johnson et al.'s original framework (Masip

et al., 2005). The assumption made by RM lie-detection researchers has been that imagined events are similar to deceptive accounts, hence the approach can be not only used to discriminate between experienced and imagined events but also between truthful and deceptive accounts. However, although this is an intuitively plausible idea, it is important not to beg the question. Indeed, research that has examined the usefulness of the RM approach to distinguish between truthful, imagined and deceptive accounts has indicated that imagined accounts have more cognitive and affective information than deceptive accounts, and, overall, imagined accounts are more similar to genuine rather than to deceptive accounts (Barnier et al., 2005). Hence, despite the apparent general effectiveness of the RM approach in lie-detection, this again raises the question of whether all of the criteria that were originally derived for distinguishing between real and imagined events can uniformly be applied to distinguishing between truthful and deceptive accounts; consequently, this issue points again to the importance of not only establishing a definitive list of criteria and definitions, but assessing which are effective, and which are perhaps not, in distinguishing between truthful and deceptive accounts.

3.4 Applications of the Reality Monitoring approach

Although, as yet, the Reality Monitoring approach has more or less been confined to experimental studies (Masip et al., 2005), some attempts have been made to raise the stakes of the research paradigm; for example, by using child witnesses (Granhag et al., 2006), transcripts from real allegations of sex-abuse made by children (Robert & Lamb, 2010), and suspects' statements (Porter & Yuille, 1996). However, as previously intimated, there are a number of key advantages that the RM approach might potentially have over alternatives both for research and in the field. In

particular, as well as being able to assess both written accounts and transcripts of oral testimonies. it potentially contains a manageable number of criteria, greatly simplifying training and administration; and giving greater potential for standardisation and reliability.

Nevertheless, if the aim of developing the RM approach is ultimately to develop a technique that will outperform alternatives such as CBCA in the field, it would make sense to first attempt to develop and test in the laboratory some sort of standardised RM protocol that could potentially be used by practitioners in the field. As indicated in the previous section, at present there appears to be no such standardised procedure; indeed, researchers do not even seem to agree on the basic criteria to apply. Hence the development of such a protocol that can be applied by interviewers/raters was one of the objectives of the present thesis. However, an equally important consideration is to define the parameters under which a protocol is most likely to be effective; which brings us to the issue of RM moderators.

CHAPTER 4

The problem of moderators

As emphasized in the previous chapter, there has been a very obvious lack of standardisation of studies in this area, which suggests that there are potentially a large number of moderating variables that could have affected the outcomes of various RM studies and might, therefore, limit their robustness and generality (though it can be noted that problems arising from failure to control for these possible moderating factors are unlikely to be confined to RM studies). These potential moderating factors include differences in the information disclosed to participants about the study, the stimulus materials and their mode and form of presentation, the actual participants providing the stimuli, scoring procedures employed and so on. Some of the most important of these are detailed and considered in this chapter.

4.1 Early procedural factors

A number of early decisions during the construction of a psychological study using the RM approach might be important outcome-moderators. Two factors in particular might affect the results: the first is motivation. Often lie-detection studies do not make any attempt to consider participants' motivation, either at recruitment or during the study (for examples see Bond, Omar, Mahmoud, & Bonser, 1990;

Burgoon & Buller, 1994; Ekman, Friesen, & Simons, 1985; Sporer, 1997; Vrij & Heaven, 1999; Zuckerman, Driver, & Koestner, 1982; Zuckerman, Kernis, Driver, & Koestner, 1984). When studies are considered that have taken motivation into account, three broad types of motivation induction can be identified; (i) identity-relevant motivation, where participants are told that convincing the message receivers that their accounts are genuine/deceptive is a highly desirable characteristic (for examples see Streeter, Krauss, Geller, Olsen, & Apple, 1977; Vrij, Semin, & Bull, 1996); (ii) instrumental motivation, where participants are offered rewards, often in the form of money or course credits (for examples see Bond, Kahler, & Paolicelli, 1985; Porter & Yuille, 1996; Stiff & Miller, 1986; Vrij, 1993,1995); and ii) a combination of identity-relevant and instrumental motivation (Miller, deTurck, & Kalbfleisch, 1983).

These considerations may be important given that DePaulo et al. (2003) found that, in lie-detection studies generally, the diagnostic value of cues to deception was better in designs where the participants were motivated to succeed in convincing the lie-detectors that their deceptive story is genuine. The authors also found that the effects of deceptive cues were stronger when identity relevant motivation was induced than when instrumental motivation (i.e. financial reward). However, to the author's knowledge, so far, the only RM study that has investigated motivation as an independent variable (Raskin & Esplin, 1991) found no effects of this variable on the accuracy of veracity assessments. Nevertheless, it has been suggested that, in RM research generally, if the results are to have any practical application, it makes sense to use procedures that allow stimulus participants (those providing the accounts) to be immersed or engaged in realistic events (Masip et al., 2005).

A second potential moderating variable concerns the information given to participants about the purpose of the study, and possible associated demand characteristics effects. It is generally assumed that, to avoid demand characteristics effects, it is preferable for participants not to be given information beforehand about the exact purpose of the study. Nevertheless, there appears to be considerable variability between studies in the instructions given to both stimulus-participants (i.e. participants who give the accounts) and lie-detectors; for example, in some studies it appears that participants have been informed the study is about lie-detection (Vrij et al., 2000), whereas in others they are variously told that the research they participate looks into the association between health and mood (Memon et al., 2010), or even more vaguely about personal experiences (Sporer & Sharman, 2006); though very often no details are given in this respect (see, for example, Bembibre & Higuearas, 2012; Granhag et al., 2006, Nahari & Vrij, 2014; Vernham, Vrij, Mann, Leal, & Hillman, 2014). Consequently, it is difficult to assess whether introductory instructions and the degree to which the participants were blind to the study's purpose might have affected the results.

Notably, in contrast, most RM studies give a better indication of the information and instructions given to the lie-detectors, though these are not standardised between studies (see, for example, Santtila et al., 1998; Sporer, 1997; Vrij et al., 2007). The common trend is to use very experienced or well-trained lie-detectors who are, therefore, familiar with the use of RM as a lie-detection tool. As a result, we do not know how RM coders might have scored the statements had they not been aware that they were taking part in a lie-detection study. It is possible, for example, that implicit global judgements of accounts as genuine or as deceptive might affect the coding of the presence or absence of RM criteria within the account

(perhaps reducing their accuracy in detection, or spuriously inflating accuracy). It was, therefore, one of the objectives of the present thesis to investigate whether keeping response participants (i.e. lie-detectors) blind to the purposes of the study might affect their scoring.

4.2. Training of lie-detectors

Related to the previous point, the training of lie-detectors or RM raters/coders is potentially a very important moderator since one might expect the coders' expertise to have an impact on the accuracy of applying the RM criteria. Consequently, researchers in this area often utilise experts in lie-detection or train postgraduate or undergraduate students to code the material. For instance, with the exception of a study by Sporer and Sharman (2006) where only basic training was provided for student coders, in all RM studies that have used two or three coders, extensive training was provided for the coders (see, for example, Santtila, et al.,, 1998; Sporer, 1997; Vrij et al., 2007). In fact, the level of coders' knowledge of verbal lie-detection literature in most studies would be considered exceptional by the standards of a lay person or ordinary police officer (see Sporer, 1997). Vrij et al. (2000) have argued that this may be justified; for example, they suggest that one of the reasons why performance in non-verbal lie-detection is relatively poor compared to in physiological and verbal-lie-detection may lie in the fact that studies which have tested the former often employed lay persons as detectors, whereas those which have tested the latter have employed experts, because only they can conduct such examinations. Nevertheless, it would clearly be desirable to evolve a technique that could be introduced into the criminal justice system and used effectively but involves minimal training and associated costs. Hence, although the effects of

training and expertise per se were not formally investigated in the present thesis, minimally trained coders from various backgrounds were involved as well as experts.

4.3. Stimulus participant factors: Second-language effect

There has been very little research on how the characteristics of story-tellers (i.e. stimulus participants) might affect the ability of lie-detectors to detect deception. Nevertheless, some researchers have suggested that there may be important individual differences not only in lying behaviour, but on verbal behaviour as a whole, that might make results difficult to interpret (Nahari & Vrij, 2014).

For example, although the evidence is not conclusive (Masip et al., 2005), it has been suggested that RM's diagnostic validity may be higher in adults since the comparatively weaker verbal skills of children might reduce the capacity of the technique to discriminate between truthful and deceptive accounts (Santtila et al., 1998). However, evidence regarding the influence of other sources of individual differences in RM research is extremely scarce. One relatively under-researched area in this respect is that of the language proficiency skills of stimulus-participants or story-tellers (Caldwell-Harris & Ayçiçeği-Dinn, 2009; Broadhurst & Hiu Wan Cheng, 2005). Typically, proponents of lie-detection techniques make no reference to whether or not these techniques are equally effective with people who are not using their first-language; victims, suspects and witnesses of crime are examined similarly irrespective of their first-language, even though this could affect the reliability of lie-detection technique used (U.S National Research Council, 2003). For instance, given that both lying and speaking a second-language require the employment of extensive cortical areas, it has been proposed that bilingual

individuals are hit with a double stressor when they lie, which might aid differentiation (Caldwell-Harris & Ayçiçeği-Dinn, 2009). It may, therefore, be worth investigating whether the fact that an account is given by someone using his/her first or second-language has an effect on the RM scores and their overall diagnostic strength.

4.4. Design-related factors

Another potentially moderating factor in RM research is that of experimental design; in particular whether different participants are asked to produce a truthful or a deceptive accounts(i.e. a between-subjects design), or whether participants are asked to produce both a truthful and a deceptive account (i.e. a within-subjects-design). Both designs have been used previously in lie-detection research and have various advantages and disadvantages. Thus, on the one hand, between-subjects designs (Sporer, 1997; Vrij et al., 2001; Wright-Whelan et al., 2014) resemble some real-life investigations where suspects or witnesses provide a single statement of which the truth-status is unknown to the investigators. However, within-subjects designs, whereby the same stimulus participants provide both truthful and untruthful accounts, are more efficient, allowing analysis of a larger set of statements, but most important, they more closely resemble the designs used in physiological lie-detection research, whereby baseline behaviours are compared with behaviours in reaction to crime-related questions (BPS, 2004; Porter & ten Brinke, 2010). In this respect, they could be considered more sensitive. Another of the advantages of within-subjects designs is that they minimise noise from individual differences (Nahari & Vrij, 2014). Hence, the latter design is to be found in the bulk of traditional laboratory-based (Alonso, 1992; Bembibre & Higuears, 2012; Manzanero

& Diges, 1995) and realistic high-stakes studies of lie-detection (Mann & Vrij, 2006; Mann et al., 2006;Villar, Arciuli, & Mallard 2012; Vrij & Mann, 2001a). Nevertheless, as yet, there has not been a comparison between the two designs in the RM literature. Given this, with regard to the present thesis, at the outset one aim was to look at the relative efficacy of the two designs in an RM context.

4.5. Nature of the stimulus materials/accounts

As mentioned in the earlier section on motivation, one might also expect the emotional significance and content of different topics to affect the nature and quality of accounts in terms of RM criteria, independently of motivational considerations. For example, one might predict that participants' descriptions of a genuine experience will differ depending on their role in the story they describe (such as whether they are a victim or a bystander), whether their experience was staged or was spontaneous, whether they were actively engaged in the event or merely watching an event from a safe distance or from a computer monitor (Alonso-Quecuty, 1992; 1995), or whether the event was autobiographical (e.g. Sporer & Sharman, 2006). However, most important, it has been suggested that, in RM research generally, if the results are to have any practical application, it makes sense to use procedures that allow participants to be immersed or engaged in realistic events (Masip et al., 2005).

4.6. Scoring procedures

With regard to scoring procedures, there are broadly two types of scoring systems commonly used in RM research: subjective ratings of RM criteria and raw frequencies (i.e. the exact number of times that each criterion is deemed to be

present). Hence some studies have primarily used rating scales (Santtila et al., 1998; Sporer, 1997; Sporer & Sharman, 2006), others frequencies (Alonso-Quecuty, 1992; 1995), and others, both (Granhag et al., 2006; Stromwall et al, 2004); however, as yet, no direct comparisons between the two have been made. In fact, as will be detailed shortly, when using raw frequency RM data, there has been such lack of consensus about how to deal with length-differences between the truthful and deceptive accounts that some authors have more or less given up using raw frequencies because of this confound (see, for example, Granhag et al., 2006; Strömwall et al., 2004). Given this, and for other reasons that are detailed later (i.e. complexity), it might be predicted that with relatively untrained raters/coders at least, rating scales might prove to be a more useful and effective tool in this context.

4.7. Oral and written accounts

Another neglected feature of RM studies concerns the modality of the stimulus materials in terms of whether the accounts are written or spoken. There are a number of ways in which oral and written statements might differ which may be relevant to lie-detection (Beaugrande, 1984; Kroll, 1977). Speakers, for example, have a much faster production rate than writers, whereas writers rarely interact with their audiences as much as speakers often do, hence they perceive their roles differently (Chafe, 1982). Moreover, oral narratives have overall lower *lexical density* than written accounts as they are often unplanned (Halliday, 1989; 2001; Tannen & Chafe, 1986). This creates one of the main differences between the two types of accounts: written discourse is more coherent and richer in content-words per clause than spoken language.

It has also been reported that oral accounts contain more *cause* or *because* clauses than written ones (Beaman, 1985; Pu, 2006). Such words in written accounts are used to describe causal relationships between different events (e.g. *The person knocked but the man didn't answer because the other person didn't ring the bell)*. These links tend to be objective and they are assumed to be an outcome of the writers' role and perceived obligation to explain unusual or unexpected events in the absence of the face-to-face chance to be seen as credible (Pu, 2006). On the other hand, in oral accounts, such words are used in a more elaborate fashion; on some occasions, they are also used to explain causal relationships but often to give subjective and imprecise explanations of events (e.g. *He didn't ring the doorbell cause I don't think he saw it*; Pu, 2006).

Nevertheless, despite these differences between spoken and written discourse, and their implications for lie-detection using verbal reports, there appear to have been only five studies utilising RM assessments of written statements to discriminate between truthful and deceptive accounts. Moreover, of these, only three utilised solely written statements (i.e. Barnier, et al , 2005; Nahari et al., 2012; Sporer & Sharman, 2006), whereas the remaining two studies (i.e. Granhag et al, 2006; Manzanero & Diges, 1995) used both written and spoken statements, but with no attempt to identify the different processes involved or assess their relative utility in terms of lie-detection. For example, in their study on children, Granhag et al. (2006) attempted to justify the use of both types of account (oral and written) by arguing that asking interviewees to recall information using different modes (e.g writing, talking, and drawing) mirrors the "type of differentiated memory work that may characterise the disclosure-process" (p.85), but they made no attempt to compare them. Indeed, the oral and written accounts were treated as equivalent. This

may be significant given that, although the findings of studies using both written and spoken accounts have been generally supportive of RM, some findings have been anomalous. For example, whilst Barnier et al. (2005) found that truthful accounts were clearer and contained more affective information than deceptive accounts, in direct opposition to the predictions of RM theorising, they also found that truthful accounts contained more information regarding cognitive operations than deceptive accounts. It can also be noted that, whilst the findings of Nahari et al. (2012) generally supported the use of written statements in RM assessments (since total RM scores were higher for truthful than for deceptive accounts), they did not detail the individual diagnostic validity of their criteria but merely compared total RM scores to SCAN scores. Clearly, there is a need to compare to relative efficacy of written and oral accounts in RM research.

4.8. Account Length

RM studies, like others that use raw frequency counts as primary data, seem consistently to underestimate the influence of a key methodological factor, namely the length of (number of words in) the account. It is generally assumed that lying is a cognitively demanding task, in which, in an attempt to appear credible, the liar must continually monitor his/her performance, by, for example, selectively concealing information and avoiding contradictions (DePaulo et al., 2003; Vrij, 2008; Zuckerman et al., 1981). One might, therefore, predict that deceptive accounts will be generally shorter than truthful accounts, because liars may be more likely to conceal information, or avoid providing potentially contradictory information, which may reduce their credibility. The cognitive load associated with lying might also generally 'slow down' their thinking time. As a result, the overall length of accounts

per se has often been proposed as a potentially useful cue to deception (DePaulo et al., 2003; Porter & Yuille, 1996; Vrij et al., 2004; Vrij et al., 2000).

However, findings relating to word-count have been mixed. Thus whilst a tendency for deceptive accounts to be shorter has been reported in some studies (Dilmon, 2009; Driscell, 2013; Nahari et al., 2012; Santtila et al., 1998; Stromwall, et al., 2004; Stromwall & Granhag, 2005; Vrij, Evans, Akerhurst & Mann, 2004), others have found that deceptive accounts may be longer than (Sporer & Sharman, 2006) or similar (Sporer, 1997; Vrij et al., 2007) to truthful accounts. It has also been suggested that the length of the account may be influenced by the time-interval between the recall phase and the experience being described (Alonso-Quecuty, 1993).

This issue has obvious implications for RM research. For whilst the "number of words" or "account length" is not of itself included as an RM criterion in most descriptions of RM in the literature, it may clearly have an impact on how RM criteria are applied; for example, other things being equal, longer accounts in terms of word frequency are likely to contain higher frequency counts of the relevant RM criteria than shorter accounts.

4.9. Standardising for length differences

Particular problems in this regard arise when frequency data are used to measure more discrete RM variables, such as visual and cognitive information. In such cases, results can vary considerably depending not only on whether the accounts have been standardised for word-count but also the actual method of standardisation (Masip et al., 2005). Consequently, decisions to standardise per se can affect the diagnostic

validity of the criteria in separating truths from lies, particularly when the lengths of the truthful and deceptive accounts differ.

Decisions as to whether or not to standardise also tend to be very ad hoc. It is a common practice, for instance, to control for word-count when statistically significant differences in length are found between truthful and deceptive accounts (Gnisci, Caso, & Vrij, 2010; Memon et al., 2010). Nevertheless, often length differences are found but no standardisation takes place (see, e.g. Nahari et al., 2012). Also, there are instances in the general literature where the accounts did not differ significantly in the amount of words they contained, but they were still standardised for length; and, in other studies, for no clear reason, some criteria were standardised and others not (see, e.g. Vrij et al., 2004; Vrij et al., 2000). Moreover, even when the decision to standardise is taken, methods of standardisation can differ considerably. For example, a particularly common standardisation method is to calculate the number of raw frequencies of a particular RM criterion contained per 100 words of the account (Larson & Granhag, 2005; Strömwall & Granhag, 2005; Vrij et al., 2004). But other alternatives have included presence of cues per 50 words (Vrij et al., 2000), the transformation of raw frequencies into a 5-point-rating scale (Memon et al., 2010; Vrij, Mann, Fisher, Leal, Milne & Bull, 2008) and even measuring raw frequencies as well as controlling for the duration of the account (in number of minutes) (Gnisci et al., 2010). Such is the complexity of and ambiguity associated with this issue that some authors have more or less given up and argued that when there are significant length differences between truthful and deceptive accounts, raw frequencies cannot be used because the raw criteria frequencies and the number of words used are confounded (e.g. Granhag et al., 2006; Strömwall et al., 2004). In the present thesis, therefore, an attempt is made to systematically

address the issues of account length and standardisation procedures in relation to RM scoring.

4.10. The presence of others

Another factor that might potentially influence the application of RM criteria is the presence of others when participants are producing their accounts. Of possible relevance here is the seminal work of Zajonc (1965, 1980) on social facilitation and inhibition (see also, Wagstaff, Cole, Brunas-Wagstaff, Blackmore & Pilkington, 2008). Zajonc proposed that the mere presence of others can enhance performance in simple or well-rehearsed tasks but impair performance in unfamiliar or complex tasks. His explanation is that the mere presence of others increases drive levels and inhibits emission of subordinate responses. Consequently, dominant and automatic responses are promoted, and overall performance is thereby determined by the efficacy and accuracy of these dominant responses. Thus, on familiar or simple tasks, the dominant responses are more likely to be appropriate and correct; therefore, performance is more likely to be enhanced. However, performance of a complex or a non-familiar task in the presence of others will probably be impaired as dominant and/ or automatic responses will tend to lead to suboptimal performance.

There have been a number of competing explanations for social facilitation and inhibition effects, including distraction (Aiello & Douthitt, 2001; Baron, 1986), and evaluation apprehension Cottrell (1972). However, Wagstaff et al. (2008) have suggested that the results from various studies may be best accommodated by a working memory model. The idea is that monitoring the presence of others may increase cognitive load distraction and arousal which have the effects of inhibiting executive (frontal) processing, whilst facilitating more automatic (temporal)

processing. In this way, the presence of others essentially puts the individual in a 'dual task' situation, such that novel and complex tasks, which require executive processing, are inhibited due to increases in cognitive load, and automatic familiar tasks, which involve only automatic processing, are facilitated due to increases in drive or arousal.

Wagstaff et al. (2008) suggest that this may make sense from a sociobiological perspective in terms of initiating the 'fight or flight' response. That is, human groups, especially strangers, may be perceived as a potential source of threat; hence, if the executive system is activated by being in a group (because one actively monitors the group), then one might predict that an executive task would be performed less well in a group situation; precisely in the same way as the administration of two executive tasks concurrently inhibits performance on one or both tasks. However, as a corollary, this might 'free up' other 'fight or flight' systems to respond automatically to environmental threat without intervention from a supervisory system (Wagstaff et al, 2008).

From this perspective, therefore, one might predict that if the presence of others increases cognitive load, and deception is a cognitively demanding task, then lying will be made more difficult (and the cues more obvious) in the presence of others. It may be significant, therefore, that in investigative interviews the practice of using two interviewers is common (Baldwin, 1993), and is considered advantageous, particularly on ethical grounds (Horgan & Horgan, 1979; Kincaid & Bright, 1957); however, the efficacy of using different numbers of interviewers (and having others in the room, such as lawyers, appropriate adults, etc.) in terms of gleaning accurate testimony has received very little empirical attention. For example, it could be argued that including a second interviewer might reduce

intimacy and rapport building which could otherwise encourage reluctant interviewees to ‘open up’ (Simmel, 1964). On the other hand, as just suggested, one might predict that the increased cognitive load resulting from the presence of others might make deception easier to detect.

What little research has been conducted on the effects of different numbers of interviewers suggests that the addition of a second interviewer does not affect rapport building or its indicators such as the lexical density of the accounts in words that reflect mutual attentiveness, positivity and coordination; however, transcripts tend to be longer when two interviewers are present as both interviewers and interviewees appear to speak more), thus more information is gathered (Driskell, Blickensderfer, & Salas, 2013; Tickle-Degnen & Rosenthal, 1990).

Given this, the present thesis also includes an investigation into the effects of the presence of others on RM outcomes.

4.11. Conclusion

To summarize, there appear there be a number of moderators that could potentially influence the outcomes of RM and are worthy of both investigation and control, yet, so far, have received little and sometimes no consideration in the literature. With this in mind, those receiving particular emphasis in the present thesis are listed as follows.

1. The language proficiency of story-tellers.
2. The type of modality used (i.e. written vs. spoken accounts).
3. The scoring system used (rating scales vs. raw frequencies).
4. The demand characteristics and their effects on using the RM criteria.

5. The type of standardisation used to control for account length and its effects on RM assessments.
6. Absence/presence of others effects: i.e. the number of interviewers in the room.

As pointed out, at the outset, it was also intended to make a systematic comparison of between and within-subjects methods; in the event, however, the disadvantages, including practical limitations, of the between-subjects method soon became so apparent that a systematic comparison was abandoned in favour of using solely the within-subjects method. Similarly, due to practical considerations and time limitations, the motivational characteristics of the stimulus materials was not investigated systematically, instead an attempt was made to use materials that would at least moderately engage the story-tellers.

PART 2

The Empirical work

Chapter 5

Introduction to the empirical research

As an introduction to the empirical work, the purpose of the present chapter is to give an overview of the main research aims and hypotheses, outline the empirical studies and some relevant methodological issues, and review ethical considerations.

5.1. Research Aims

In the light of the research findings discussed in Part 1 the core aims of the present thesis can be summarised as follows.

1. To assess whether the RM approach has any value overall in distinguishing between truthful and deceptive accounts.

2. If it does, to investigate the circumstances under which it might give optimal results; i.e. to assess what factors moderate its efficacy in this respect.

Within the remit of these broad aims, therefore, several more specific aims were formulated in relation to using RM to discriminate between truthful and deceptive accounts.

1. To compare the relative efficacy of rating scales and raw frequency RM scoring systems, particularly using untrained raters/coders.

2. To examine possible second-language effects on the efficacy of the RM approach.

3. To test the relative efficacy of using spoken and written statements with the RM approach.

4. To test the effects of standardisation for length and duration of accounts on the efficacy of the RM approach.

5. To test more generally the usefulness of length, duration and speech rate as cues to deception (i.e. as influences on, and possible adjuncts to or elaborations of, RM).

6. To assess whether demand characteristics (blind coding) may influence the coding and efficacy of RM criteria.

7. To test the influence on the efficacy of the RM approach of the number of people present in the room when accounts are given.

5.2 Hypotheses

Given these aims, there were, respectively, seven main research hypotheses formulated based on the considerations reviewed in Part 1, and additional literature reviewed before each study in this part of the thesis.

1. In terms of the efficacy of RM criteria in distinguishing truth from lies, due to factors including the complexity of the procedures, it was hypothesised that RM rating scales will be a more useful RM measurement tool than raw frequency items amongst a group of untrained raters.

2. Regarding the second-language effects on RM-based assessments, the approach here was more investigative than hypothesis driven, but it would seem reasonable to hypothesise that as second-language liars are engaging in a potentially more stressful and cognitively demanding task than first-language liars (trying to lie in another language), it may be easier to detect lying through RM in the former

3. It was hypothesized that, irrespective of whether the accounts are truthful or not, spoken accounts will receive higher RM scores than written accounts. However,

there appears to be no a priori reason why one should be any better generally at differentiating between truthful and deceptive statements.

4. It was hypothesized that the RM approach will be better at discriminating between truthful and deceptive accounts before standardisation for length and duration.

5. It was hypothesized that truthful accounts would be longer in length and duration, and would be associated with a faster speech rate than deceptive accounts, and these factors would influence the measurement of RM criteria.

6. It was predicted that raters/coders will be better in discriminating truthful from deceptive accounts when blind to the purpose of the study. The main rationale for this was that generic or global truthfulness assessments may affect the coding of RM criteria (i.e. even untrained raters might assume that truthful accounts differ in terms of different kinds of detail, and bias their ratings accordingly).

7. It was hypothesised that the speech rate of the accounts will be lower when two people are in the room and, as a result, the accounts will be shorter in length and have lower RM scores. Moreover, being in a room with two persons will induce the most cognitive load hence this is when the RM scores will discriminate the best between truthful and deceptive accounts

5.3. The empirical studies

Given these aims and hypotheses, the following empirical studies were conducted. These were as follows.

5.3.1. *Study 1*

Study 1 examined the potential of the RM technique to discriminate between truthful and deceptive written statements produced by individuals who used English as a first or as a second-language, and to investigate whether writing in a second-language affects the usefulness of the RM approach to discriminate between truthful and deceptive accounts. Also, both rating scales and raw frequencies were used to investigate which of the two scoring systems is most effective with untrained raters/coders.

5.3.2 *Study 2*

Study 2 examined the efficacy of the RM criteria, as diagnostic criteria to detect lies in two different conditions; i.e. transcriptions of spoken statements, and written statements. This study also investigated the efficacy in distinguishing truth from lies in using two indicators that could influence the application of RM criteria, and possibly be employed alongside RM, account length and speech rate (number of words produced per second).

5.3.3 *Study 3*

The aim of Study 3 was primarily to check whether coders using the RM are still able (or even better able) to discriminate between truthful and deceptive accounts when they are blind as to the exact purpose of the study.

5.3.4 *Study 4*

As noted previously, researchers into verbal lie-detection do not always standardise for account-length and when they do, they utilise a variety of different approaches.

The purpose of Study 4, therefore, was to test the usefulness of RM in discriminating between truthful and deceptive accounts before and after standardisation. Again, this variable has rarely been systematically assessed in verbal lie-detection research. The efficacy of the Reality Monitoring criteria, as diagnostic cues to detect lies in spoken and written statements was assessed again in this study.

5.3.5 *Study 5*

Following on from, and as a way of replicating the findings of Study 4, Study 5 was a reconsideration of the data from Study 1 using an analysis that allowed the calculation of RM scores before and after standardization for word-count.

5.3.6. *Study 6*

This final study, Study 6, investigated the possible effects of the number of people in the room when the account was given, on the efficacy of RM to detect lies. In addition, this study looked at how standardising for account-length differences might affect the usefulness of RM using two types of standardisation: word-count and duration standardisation. With regard to the latter, the main purpose was to investigate not only whether RM works better before or after standardisation, but also what type of standardisation is most effective in this respect. This study also considered the effects of the presence of others on the diagnostic value of RM and other indicators (i.e. speech rate and account length).

5.4. Methodological Considerations

5.4.1. *Stimulus Materials*

For Studies 1 and 5 (described in chapters 6 and 10 respectively), the stimulus materials were short video clips of a crime story. Similar materials were used in the initial RM studies by Alonso-Quecuty (1992; 1995). This approach has a number of advantages; for example, it is relatively easy to set up, and it allows considerable control over the ground-truth of accounts. It has also, produced a number of positive results with regard to the application of RM (Alonso-Quecuty, 1992; 1995; Vrij et al., 2000; 2001); hence some researchers have encouraged the use of videos over other materials (Masip et al., 2005).

One of the disadvantages of the video approach, however, is that participants may not emotionally engage with the materials as they might in a real-life situation, which compromises their ecological validity (for example, De Paulo et al., 2003). An alternative, therefore, is to use autobiographical accounts as stimulus materials that stimulus participants can lie or tell the truth about; hence autobiographical accounts are commonly used as stimulus materials in lie-detection research (for example, Ball & O'Callaghan, 2008; Barnier et al., 2005; Johnson et al, 1988; Masip et al., 2005; Sporer & Sharman, 2006). However, although autobiographical accounts have the advantage over videos that participants may be more actively and realistically engaged with them, they potentially present greater problems in terms of establishing ground-truth. Nevertheless, it could be argued that, unless participants were being deliberately disruptive and uncooperative, it seems unlikely that they would simply manufacture their truthful accounts, whether in whole or part. And even if they did, this would tend to reduce the distinction between truthful and

untruthful reports such that the findings would err on the side of caution, an outcome that could be construed as preferable in this area. Given these considerations, and experiences actually running the studies, in Studies 2, 3, 4 and 6 (described in Chapters 7, 8, 9 & 11, respectively), autobiographical events were ultimately used as the preferred stimulus materials.

5.4.2. *Design Considerations*

For reasons mentioned in Chapter 4, both between and within-subjects designs were used in the present thesis; i.e. stimulus participants gave either a truthful or a deceptive account (Studies 1 & 5), or each participant provided one truthful and one deceptive account (Studies 2, 3, 4 and 6). However, as mentioned also, whilst conducting the studies, the disadvantages of between-subjects designs in the generation of stimulus materials, in terms of both practical efficiency and sensitivity, soon became apparent, so ultimately the within-subjects approach was preferred.

5.5. Ethical Considerations

This research was conducted in accordance with BPS and APA research ethics guidelines and all studies (both methods and materials) were approved by the University of Liverpool Institute of Psychology Health and Society Research Ethics Committee. No vulnerable individuals were employed in the studies and the video materials were selected with ethical considerations in mind. For example, the videos used for Study 1 were taken from a film on general release and contained only moderate violence. Standard procedures for gaining informed consent were used, and in all studies participants were informed of their right to withdraw from the study at any point without having any obligation to explain their reasons for

withdrawing (see Appendix 1). They were also given extensive information about the study before they began and were appropriately debriefed at the conclusion (see respectively Appendices 2 and 3).

5.6. Setting

Data collection for Studies 1-5 was completed in University of Liverpool premises and Study 6 was conducted in both University of Liverpool and University of Huddersfield premises (both within their respective Psychology departments). In all cases, a small, quiet room was used, although, in Study 6, a slightly larger room was used to allow more space for the stimulus participants to describe the truthful and fabricated stories in the presence of different numbers of people.

Chapter 6

Study 1: Reality Monitoring in the assessment of written statements with attention to possible second-language and scoring method effects: A pilot study

6.1. Introduction

The aims of Study 1, reported in this chapter, were to investigate the potential of the RM technique to discriminate between truthful and deceptive written statements and whether writing in a second-language affects its usefulness in this respect. Also, both rating scales and raw frequency measures of RM were used to investigate which of the two scoring systems works best with untrained raters/coders.

It is perhaps important to emphasise here that this first study was very much an introductory pilot (hence the caveat in the title), run at the beginning of the project, before the researcher's aims, ideas and relevant expertise had been more fully developed and informed. As a result, and as will soon become apparent, it suffered from a number of methodological limitations, not only in design but in scoring. However, in many ways, this study was very informative in this respect, and, therefore, the decision was made to include it here, 'warts and all'.

6.1.1. *Using RM to assess deception in second-language accounts*

A variety of evidence suggests that a number of cultural factors may affect the ability to detect lies (Bond, et al, 1990; Broadhurst & Cheng, 2005). Notably, Bond

et al's (1990) study of non-verbal deceptive behaviours was one of the first to suggest that cues such as head movements and lack of gesturing might be seen as an indication of deceit amongst Americans but not amongst Jordanians. Hence lie-detection accuracy may vary depending on cultural context. In line with this, other research suggests that liars' psychophysiological responses may be different when they lie in their first-language compared to when lying in a second-language. For example, participants' electrodermal activity is higher when they lie in a second-language; though subjective reports indicate that lying in one's first-language is felt more strongly (i.e. evokes a greater emotional experience) than lying in a second-language (Caldwell-Harris & Ayçiçeği-Dinn, 2009). Further, as noted in Chapter 4, other studies have found that when considering truthful accounts, veracity cues are diagnostically weaker when people speak in a non-native language than when using their own first-language (Broadhurst & Cheng, 2005; DaSilva & Leach, 2013; Leach & DaSilva, 2013). Moreover, these effects persist regardless of whether the cue judgments are made by either lay persons or police officers (Leach & DaSilva, 2013). However, the reverse seems to be the case with deceptive accounts; i.e. veracity cues are more accurate/discriminating when liars speak their second-language (Broadhurst & Cheng, 2005).

The above results suggest, therefore, that language effects in lie-detection judgments may vary depending on whether the speaker is lying or telling the truth. This could, in part, be due to the demonstrable lie-bias that exists towards second–language speakers (Evans & Micheal, 2014; DaSilva & Leach, 2013; Leach & DaSilva, 2013). That is, if participants are more likely to judge second-language speakers as deceptive and first-language speakers as truthful, irrespective of their truth status, then they will be more likely to be accurate in detecting lies in second-

language accounts and truths in first-language accounts. However, so far evidence for these biases has been derived from studies of non-verbal behaviours, so it is not known whether the findings are applicable to analysis of verbal behaviour.

In general, the consensus view seems to be that the ability to detect lies deteriorates in cross-cultural, face-to-face interactions (Bond et al., 1990; Vrij, 2000). Nevertheless, despite a call for more cross-cultural research on lie-detection (Zhou & Lutterbiem, 2005), virtually no research has been conducted on how using a second-language may influence the lexical structure of accounts and their respective interpretation using content-based approaches to lie-detection. One of the main aims of the present study, therefore, was to investigate the effects of using a second-language on the efficacy of RM in detecting lies. Since verbal behaviours are not as salient as non-verbal behaviours (Vrij, 2008b) one might predict that second-language effects might be diminished in verbal lie-detection; moreover, second-language speakers may not have the vocabulary to describe details in ways that might be picked up in RM scoring procedures, diminishing the ability of RM criteria to detect deception. Nevertheless, it would seem reasonable to propose that second-language liars are still engaging in a potentially more stressful and cognitively demanding task than first-language liars (trying to lie in another language), and, as such, it may be easier to detect lying through RM in the former.

6.1.2. *RM scoring systems: raw frequencies vs rating scales*

As mentioned previously, RM researchers have tended to use two main scoring methods, raw frequencies which potentially measure the exact number of occurrences of the RM criteria, and Likert-style rating scales. The latter constitute a more global and subjective form of assessment, which, on first consideration, might

seem less sensitive as a method of discriminating between truthful and deceptive accounts. Recognizing this, some researchers have attempted to use rating scales that include more rating points or categories; i.e. from 1-5 to 1-7, or even 1-10 (see, for example, Sporer & Sharman, 2006; Szechtman, Woody, Kenneth, Bowers & Nahmias, 1998). However, they have not described or explained how each of the points on the scales is to be labelled or conceptualised (for example, the difference between 6 and 7 on a 1-10 scale). Studies that have used both rating scales and frequencies exist, however, they are rare and have not focused on comparisons between the two types of assessment. Granhag et al. (2006) and Stromwall et al. (2004), for example, included both types of assessment, however, they subsequently seemed to abandon the raw frequencies, as they realised that these could be confounded by the number of words contained in the accounts; i.e. longer accounts are likely to contain more RM information, irrespective of their truth status.

Another potential drawback to using raw frequencies concerns the training demands involved. Training individuals to measure raw RM frequencies requires a very comprehensive understanding of the definitions and scoring of the criteria. RM rating scales also require training in criteria definitions, but arguably, expertise in assigning global ratings is easier to obtain. However, perhaps most important, ratings can capture the full gamut of RM criteria whereas raw frequencies can capture only the criteria that can be quantified (i.e. temporal information, spatial information, affective information, cognitive information, perceptual information), and not those that are very difficult to quantify numerically (such as realism, vividness and reconstructability).

Given these considerations, in addition to a second-language effect, it was tentatively hypothesised that rating scales might be a more useful RM measurement tool than raw frequency items amongst a group of untrained coders.

6.2. Method

6.2.1. *Participants*

Two sets of participants were employed in this study, eight Stimulus Participants (SPs) and 13 Response Participants (RPs). Of the eight SPs (three males and five females), six were undergraduate and postgraduate students from the University of Liverpool, and two were members of the general public (*M* age = 26.39; range = 23-30; *SD* = 2.81); also, of these, four were native and four non-native speakers of English; in the latter case these were individuals who spoke English as a foreign language but who scored over 5.5 in an *International English Language Testing System* examination. Of the 13 RPs, seven were also postgraduate students from the University of Liverpool but from non-psychological backgrounds, and the remainder were members of the general public (*M* age = 29. 92; range = 18-47; *SD* = 8.11); there were five males and eight females. It can be noted that, although all RPs were fluent in speaking and writing in English, two used English as their second language. However, given that the aim of the study was to control for the language proficiency of the persons giving the accounts and not of the participants evaluating them, this variable was not controlled for or investigated as a moderator at this stage.

6.2.2. *Materials and Procedure*

The study was conducted in two phases. In the first phase, the Stimulus Participants (SPs) were invited to take part in a lie-detection study. Two video clips were prepared (Video 1 and Video 2, see Appendix 7 for a description of the two Videos), each lasting approximately two minutes, from a film that contained a crime-related story. The videos were from an Irish film on general release (Greenhalgh & Branigan, 2008); it contained moderate violence. These participants were then randomly assigned to one of two stimulus conditions: truthful and deceptive (N = 4 in each). In each condition, two SPs were shown Video 1 (one native and one non-native English speaker), and two were shown Video 2 (one native and one non-native English speaker). In the truthful condition, the SPs were asked to provide a written statement of their full recollection of the video-scene. In the deceptive condition, the SPs were asked to make up or fabricate an account of what they had seen in the video story and to provide a written statement of what they had seen in the video, based on this fabricated story (for a sample-summary of the accounts see Appendix 4). There was no time limit for this in either condition (although none of the SPs took more than 20 minutes to write the statement).

In Phase 2, 13 participant judges, were randomly assigned to two further conditions; in the first condition judges (N = 8) were asked to examine the written statements (both truthful and deceptive) that derived from the SPs who used English as their first-language. In the second condition (N = 5) judges were asked to examine the written statements (both truthful and deceptive) derived from the SPs who used English as their second-language. The statements were presented in a different random order for each participant (see Appendix 5 for participant instructions and scoring sheets).

For RM rating and coding, an RM framework was devised. The idea behind this was to come up with a relatively succinct but effective tool to elicit RM ratings from individuals both who may have been extensively trained in evaluating RM criteria, or may have received little if any training. This consisted of a sheet listing eight RM criteria (perceptual information, temporal and spatial information, affective information cognitive operations, temporal information and, realism, vividness and reconstructability) and a set of descriptions of their definitions derived from Vrij (2000; 2008; 2015). The criteria (with examples) were defined as follows (See Appendix 6a).

1. Perceptual Information: the presence of sensorial experiences such as sounds (e.g. 'he really shouted at him') or visual details (e.g. 'I saw him entering the room').

2. Temporal information: the presence of information about when the event happened (e.g. 'it was early in the morning') or explicitly describing a sequence of events (e.g. 'as soon as the guy entered the pub the girl started smiling').

3. Spatial information: the presence of information about locations (e.g. 'It was in a park') or the spatial arrangement of people/objects (e.g. 'the man was sitting left from his wife' or 'the lamp was partially hidden behind the curtains').

4. Remembered feelings (affect): how well the person remembers feelings (accounts of subjective mental states) from the event (e.g.' Joseph was very scared').

5. Cognitive operations: evidence in the narratives of various cognitive activities, such as thoughts or reasoning (e.g 'I must have had my coat on, as it was very cold that night') and cognitive suppositions of sensory experiences (e.g. 'She seemed quite clever'). This criterion also includes descriptions of inferences made

by the participant at the time of the event (e.g. ‘it made me think at the moment how nice it could be If I have never been there’).

6. Realism: This criterion is present if the story is plausible, realistic and makes sense.

7. Vividness/Clarity: this refers to the clarity and vividness of the statement. This criterion is present if the report is vivid, lively, clear and sharp instead of dim, faint, vague and indefinite.

8. Reconstructability: this refers to the possibility to reconstruct the event on the basis of the information given.

RPs were, therefore, given an information sheet presenting these eight RM criteria with the associated definitions. They were then asked to state to what extent they believed that the person who wrote the statement was telling the truth (overall global truthfulness rating), and to rate the statements on the eight RM criteria, both on a five-point Likert scale, from 4 indicating definitely truthful, 2 unsure/don’t know, to 0 indicating definitely lying. RM ratings also ranged from 0-4 with high scores representing high presence of the criteria in the statement and low scores low presence of the criteria. In addition, judges were asked to count the number of occasions that five of these criteria (i.e. perceptual information, spatial information, affective information, temporal information and information regarding cognitive operations) were present in the statements; that is, to score their raw frequencies. For the RM criteria of vividness, reconstructability and realism, frequencies were not measured since they refer more to global characteristics of the accounts that are not readily measured using frequency counts. Participant instructions and scoring sheets are included in Appendix 5

6.2.3. *Design*

To summarise, a 2 x 2 x 2 mixed design was employed; truthfulness (truthful/deceptive) and videos 1 and 2 were the within-subjects factors, and language proficiency (English as a first/second-language) was the between-subjects factor. The dependent variables were the ratings of the truthfulness of the accounts, the RM criteria ratings, the five RM frequency-scale items, and the Total RM scores (on both measures). The design is summarized in Table 6.1.

Table 6.1. Study Design

Judges		Video 1		Video 2	
		D	T	D	T
N=8	E1	SP1	SP2	SP3	SP4
N=5	E2	SP5	SP6	SP7	SP8

D=Deceptive account, T= Truthful account
E1= English as a first-language, E2= English as a second-language
SP= Stimulus Participants

6.3 Results

Given that there is no non-parametric test that will deal adequately with interactions, where appropriate, parametric ANOVA procedures were preferred here (and throughout this thesis) working on the assumption that such procedures are relatively robust in the face of violations of the normality and homogeneity of variance assumptions with fixed levels of the independent variable and equal or approximately equal cell frequencies (Kirk, 1968; Glass & Stanley, 1970; Howell, 1992; Schmider, Ziegler, Danay, Beyer, & Bühner, 2010; Shavelson, 1996). A series of ANOVAs was also chosen rather than MANOVAs, as there were more dependent

variables than cases per cell, thus violating the minimal sample size requirement for MANOVA (Tabachnick & Fidell, 2001).

Nevertheless, in all the following analyses, where appropriate, checks were made for homogeneity of variance using Levene's test. In the majority of cases these were not significant (i.e. the assumption of homogeneity of variance was considered satisfied). In cases where they were significant, alternative non-parametric tests were also run on any significant main effects; in all cases effects were equivalent.

Preliminary analysis showed that although, on average, the truthful accounts (M = 154.20, SD = 42.36) contained fewer words than the deceptive ones (M = 180.50, SD = 127.70), this difference was not significant, $t(6) = 0.31$, $p > .05$. Nevertheless, as mentioned previously, because of possible differences between conditions in terms of the length of statements, some have argued that raw frequency scores are not an appropriate measure for RM analyses since RM criteria frequency and the number of words per se may be confounded (Granhag et al., 2006; Stromwall et al., 2004; Porter & Yuille, 1996). To standardise word-count, therefore, the RM raw scores were re-calculated per 100 words of account; i.e. the raw RM scores were multiplied by 100 and then divided by the number of words contained in the statement (for examples of this method, see Larsson & Granhag, 2005; Strömwall & Granhag, 2005; Vrij et. al, 2004).

Preliminary 2 x 2 x 2 analyses showed no significant simple main effects; however, the overall trends were somewhat different for Videos 1 and 2; i.e. a number of statistically significant interaction effects were found involving the particular Video assessed and a number of Reality Monitoring criteria. Although assessing the effects of stimulus videos per se was not pertinent to the aims of this thesis, given these effects, it seemed misleading to simply combine the results for

the two videos; for clarity, therefore, results are presented below in different sections for each video (i.e. data for each video were analysed separately).

6.3.1. *Inter-rater reliability*

Judges 1 to 8, and judges 9 to 13, individually rated written statements 1 to 4 and 5 to 8 respectively (see Table 6.1). The inter-rater reliability was independently assessed for each written statement. Separate analyses were conducted for the rating-items and for the frequency-scale-items; therefore, 16 tests were conducted overall. Kendall's coefficient of concordance (W) showed that there was significant inter-rater reliability for both first and second-language accounts for all raw frequency measures, whereas agreement across the subjective ratings was somewhat lower (though none approached zero or were negative). These results are presented in Table 6.2.

Table 6.2 Inter-rater reliability for the RM criteria

	Video 1		Video 2	
	T(W)	D(W)	T (W)	D (W)
1st language accounts				
Rating-scale items	.44**	.16	.15	.72**
Frequency-scale items	.86**	.66**	.60**	.90**
2nd language accounts				
Rating-scale items	.63**	.11	.20*	.43**
Frequency-scale items	.59**	.63**	.64**	.77**

(Degrees of freedom for all rating and raw-frequency-scale items were 7 and 12 respectively)
T/D(W): Truthful/Deceptive accounts (Kendall's W)
*$p < .05$; **$p \leq .01$

Despite these discrepancies, because this was the first study and in many respects explorative, it was decided to proceed with further analyses rather than re-run the whole study again with different videos and participants.

6.3.2 *Analysis for Video 1 material*

The analysis for the Video 1 material is presented below.

6.3.2.1 *Rating scales for Video 1*

The RM rating scale data for Video 1 were analysed using a series of 10, 2 (truthfulness: truthful/deceptive) × 2 (English proficiency of statement's author: English as first/second-language) mixed ANOVAs with repeated measures on the first factor; i.e. one ANOVA on the subjective truthfulness rating, one in the total RM rating scores, and one on each of the eight criteria. Total RM scores were calculated by adding scores for vividness, realism, reconstructability, perceptual, spatial, affective, and temporal information and deducting scores for cognitive operations. Due to the large number of analyses only those significant are reported here. Also, because of the large number of tests performed consideration was given to adopting a more stringent criterion for significance than $p < .05$ in this particular study; however, again as in many respects this first study was explorative, it was decided to keep the $p < .05$ criterion to identify any particular trends that might be worth following up. Also, given that the $p < .05$ criterion is still generally accepted as the most practical compromise between sensitivity and feasibility in psychological research (see Bross, 1971) it was also adopted in all subsequent

studies as the number of tests conducted was considerably smaller; however, all post-hoc analyses were Bonferroni-corrected.

Main effects for truthfulness were found for Total RM scores $F(1,11) = 12.52$, $p = .005$, affective information $F(1,11) = 21.81$, $p = .001$ and cognitive information $F(1,11) = 10.46$, $p = .008$. Contrary to predictions, deceptive statements received higher Total RM ratings than truthful statements, higher affective ratings, and lower cognitive information ratings than truthful accounts (Table 6.3 below).

Table 6.3. Global Truthfulness and RM mean (SD) ratings as a function of truthfulness

RM Criterion	Deceptive	Truthful	η^2_p
Global Truth	2.54 (0.66)	1.77(1.30)	.18
Perceptual	2.08 (0.76)	1.30 (1.31)	.29
Vividness	2.92 (0.76)	2.31 (1.49)	.13
Realism	2.46 (0.78)	2.08 (1.03)	.12
Reconstructability	2.77 (0.83)	2.30 (1.25)	.13
Spatial	2.38 (1.32)	2.31 (1.44)	.01
Affective	2.77 (1.17)	0.46 (0.96)	.67**
Cognitive	2.23 (1.17)	3.38 (0.87)	.49**
Temporal	2.77 (1.01)	2.08(1.44)	.13
Total	15.92(5.12)	9.46(6.45)	.53**

*$p < .05$; **$p < .01$

No main effects were found for English proficiency in any of the ANOVAs, (see Table 6.4 below). However, two significant interactions were found. First, there

was a significant interaction between truthfulness ratings and language proficiency, $F(1,11) = 7.10$, $p = .02$.

Table 6.4. Global Truthfulness and RM mean (SD) ratings as a function of language proficiency (L1, first-language; L2 second-language).

RM Criterion	L1 accounts	L2 accounts	η^2_p
Global Truthfulness	1.87 (0.89)	2.60 (0.66)	.28
Perceptual	1.94 (1.02)	1.30 (0.96)	.14
Vividness	2.56(1.23)	2.70(0.98)	.01
Realism	2.00 (0.88)	2.70 (0.80)	.26
Reconstructability	2.30 (1.14)	2.90 (0.79)	.12
Spatial	2.00 (1.45)	2.90 (1.15)	.14
Affective	1.56 (1.25)	1.70 (0.64)	.02
Cognitive	3.00 (0.67)	2.50 (1.21)	.10
Temporal	2.10 (1.34)	3.00(0.77)	.13
Total	11.44 (6.06)	14.70 (5.25)	.11

Further *F* tests showed a significant effect for global truthfulness for the ratings of first-language statements $F(1,7) = 7.00$, $p = .02$, but not for second-language statements $F(1,4) = 1.00$, $p > .05$. When first-language statements were assessed, deceptive accounts received higher truthfulness ratings than truthful statements. That is, deceptive first–language accounts were more likely to be judged as truthful, and truthful accounts as deceptive (Table 6.5). Second, there was a significant interaction between cognitive ratings and language proficiency $F(1,11) =$

6.61, p = .026. A significant effect for truthfulness was found for the cognitive ratings of the first-language statements, $F(1,7) = 49.00$, p = .001, but not for second-language statements $F(1,4) = 0.91$, $p > .05$. Contrary to expectations, when first-language statements were assessed, deceptive accounts received lower cognitive ratings than truthful statements (see Table 6.5).

Table 6.5. Means for Interaction effects: global truthfulness, cognitive information ratings and language proficiency (L1, first-language; L2 second-language).

Rating scores	Language Proficiency	Truthful	Deceptive	η^2_p
Global	L1	1.12 (1.25)	2.62 (0.52)	.56*
Truthfulness	L2	2.80 (0.45)	2.40 (0.89)	.20
Cognitive	L1	2.12(0.99)	3.87 (0.35)	.87**
	L2	2.40(0.89)	2.60 (1.51)	.02

*$p < .05$; **$p < .01$

6.3.2.2 *Raw frequency-scores for Video 1*

As mentioned earlier, to standardise word-count, the RM raw scores were re-calculated per 100 words of account, as in previous research (for example, Larson & Granhag, 2005; Strömwall & Granhag, 2005; Vrij et al., 2004). These frequencies after standardisation were analysed using a series of six 2 × 2 (language proficiency: first versus second-language accounts × truthfulness: truthful/deceptive) mixed ANOVAs with repeated measures on the second factor; one for each of the five RM criteria, and one for the Total RM score. Analyses of the RM scores after

standardisation showed significant effects only for affective information, $F(1,11) = 51.85$, $p = .001$, but again, contrary to expectations, deceptive accounts received significantly higher scores for affective information than truthful accounts. However, although no other main effects for truthfulness were found, total RM scores for truthful accounts were overall higher (see Table 6.6).

Table 6.6. RM mean (SD) raw frequencies as a function of truthfulness

RM Criterion	Deceptive	Truthful	η^2_p
Perceptual	2.80 (1.62)	5.58 (4.83)	.28
Spatial	3.18 (2.09)	5.15 (3.41)	.19
Affective	2.02 (0.91)	0.29 (0.59)	.82**
Cognitive	1.01 (1.70)	0.20 (0.33)	.24
Temporal	1.93 (0.95)	1.44 (1.48)	.20
Total	8.92(3.35)	12.26 (6.56)	.21

**$p < .01$

No significant main effects or interactions for language proficiency were found (Table 6.7).

Table 6.7. RM mean (SD) raw frequencies as a function of language proficiency (L1, first-language; L2 second-language).

RM Criterion	L1 accounts	L2 accounts	η^2_p
Perceptual	4.08 (2.78)	4.35 (4.02)	.01
Spatial	3.99 (2.72)	4.45 (1.69)	.01
Affective	1.14(0.76)	1.19 (0.51)	.01
Cognitive	0.24(0.26)	1.19 (1.47)	.30
Temporal	1.34(1.19)	2.25 (1.01)	.17
Total	10.30 (5.17)	11.05 (4.60)	.01

6.3.3 *Analysis for Video 2 material*

The analysis for the Video 2 material is presented as follows.

6.3.3.1 *Rating scales for Video 2*

The RM rating scale data for Video 2 were analysed in the same way using a series of 10, 2 x 2 (truthfulness: truthful/deceptive × English proficiency of statement's author: English as first/second-language) mixed ANOVAs with repeated measures on the first factor; i.e. one ANOVA on the subjective truthfulness rating scores, one in the total RM rating scores, and one on the scores for each of the eight criteria.

Table 6.8. Global Truthfulness and RM mean (SD) ratings as a function of truthfulness

RM Criterion	Deceptive	Truthful	η^2_p
Global Truthfulness	2.15 (.90)	2.15 (1.21)	.00
Perceptual	1.77 (1.10)	2.53 (0.66)	.47**
Vividness	1.85 (1.28)	2.85 (0.55)	.38*
Realism	2.70 (0.75)	2.31 (0.94)	.13
Reconstructability	2.38 (0.65)	2.15 (0.69)	.09
Spatial	2.39 (0.77)	2.70 (0.63)	.39*
Affective	.85 (1.46)	1.07 (1.18)	.01
Cognitive	3.54 (0.78)	2.46 (0.87)	.56**
Temporal	0.77 (0.83)	2.23 (1.01)	.65**
Total	9.15 (3.61)	13.38 (1.94)	.46*

*$p < .05$; **$p < .01$

Main effects for truthfulness were found for Total RM scores $F(1,11) = 9.34$, $p = .01$, perceptual information $F(1,11) = 9.63$, $p = .01$, vividness $F(1,11) = 6.80$, $p = .025$, spatial $F(1,11) = 7.00$, $p = .023$, temporal $F(1,11) = 20.74$, $p = .001$ scores, and cognitive information $F(1,11) = 14.05$, $p = .003$. In line with the expectations and RM theory, truthful statements overall received higher RM ratings in each case, with the exception of cognitive information ratings which were higher for deceptive accounts (Table 6.8).

No significant main effects for language proficiency were found (Table 6.9 below); however, there were two significant interactions.

Table 6.9. Global truthfulness and RM mean (SD) ratings as a function of language proficiency(L1, first-language; L2 second-language).

RM Criterion	L1 accounts	L2 accounts	η^2_p
Global Truthfulness	2.06 (1.19)	2.30 (0.78)	.03
Perceptual	2.07 (0.73)	2.30 (1.03)	.14
Vividness	2.38 (0.08)	2.30 (1.16)	.01
Realism	2.69 (0.87)	2.20 (0.78)	.13
Reconstructability	2.50 (0.63)	1.90 (0.42)	.29
Spatial	2.56 (0.62)	2.50 (0.45)	.01
Affective	0.69 (0.70)	1.40 (1.10)	.22
Cognitive	3.19 (0.89)	2.70 (0.64)	.13
Temporal	1.31 (0.81)	1.89 (0.92)	.12
Total	11.00 (2.23)	11.70 (3.57)	.02

There was a significant interaction between spatial information ratings and language proficiency $F(1,11) = 21.01$, $p = .001$. Further F tests showed a significant effect for truthfulness for ratings of the second $F\ (1,4) = 32.67$, $p = .005$ but not the first-language statements, $F(1,4) = 3.00$, $p < .05$. When second-language statements were assessed, truthful accounts received higher ratings for spatial information than deceptive statements (Table 6.10).

Table 6.10 Means for interaction effects: spatial and affective information ratings and language proficiency (L1, first-language; L2 second-language)

RM	Language proficiency	Truthful	Deceptive	η^2_p
Spatial	L1	2.37 (0.52)	2.75 (0.71)	.22
	L2	3.20 (0.45)	1.80 (0.45)	.89**
Affective	L1	1.37 (1.41)	0.00 (0.00)	.52*
	L2	0.60 (0.57)	2.20 (1.64)	.46

*p < .05; **p < .01

Second, there was a significant interaction between affective information ratings and language proficiency $F(1,11) = 6.61$, $p = .026$. A significant effect of truthfulness was found for the affective ratings of the first-language statements $F(1,7) = 7.06$, $p = .028$, but not for second-language statements, $F(1,4) = 3.37$, $p > .05$. In line with expectations, when first-language statements were assessed, truthful accounts received higher ratings for affective information than deceptive statements (Table 6.10).

6.3.3.2 *Raw frequency-scores for Video 2*

The raw frequency scores were again analysed using a series of six 2 × 2 (language proficiency: first/second-language accounts × truthfulness: truthful/deceptive)

mixed ANOVAs with repeated measures on the second factor; one for each of the five RM criteria, and one for the Total RM score.

Main effects for truthfulness were found for Total RM scores $F(1,11) = 5.03$, $p = .046$, and spatial information $F(1,11) = 10.46$, $p = .008$. Contrary to expectations, deceptive statements overall received higher Total RM raw frequency scores and higher spatial scores than truthful statements (Table 6.11).

Table 6.11. RM mean (SD) raw frequencies as a function of truthfulness

RM Criterion	Deceptive	Truthful	η^2_p
Perceptual	11.90 (8.96)	6.38 (6.63)	.15
Spatial	6.55 (2.89)	3.54 (1.80)	.45*
Affective	1.26 (2.09)	0.78 (0.69)	.31
Cognitive	0.89 (1.45)	0.81 (0.51)	.06
Temporal	1.73 (1.32)	2.40 (1.15)	.25
Total	20.54 (9.42)	12.29 (8.01)	.31*

*$p < .05$

Significant main effects for language proficiency were found for affective, $F(1,11) = 14.56$, $p = .003$, and perceptual, $F(1,11) = 4.95$, $p = .048$, information scores (Table 6.12 below). Second-language statements received significantly higher raw scores for affective information than first-language statements, and first-language statements received significantly higher raw scores for perceptual information than second-language statements. No interaction effects were found.

Table 6.12. RM mean (SD) raw frequencies as a function of language proficiency(L1, first-language; L2 second-language).

RM Criterion	L1 Accounts	L2 Accounts	η^2_p
Perceptual	11.22 (3.17)	5.81 (5.96)	.42*
Spatial	5.44 (1.94)	4.42 (1.51)	.13
Affective*	0.25 (0.30)	2.25 (1.41)	.62**
Cognitive	0.50 (0.62)	1.41 (1.15)	.28
Temporal	1.62 (0.89)	2.77 (1.46)	.27
Total	18.03(4.06)	13.84 (6.18)	.27

*$p < .05$; **$p < .01$

6.4. Discussion

The findings can be summarised as follows: The general predictions of RM theory were supported only with regard to subjective ratings of Video 2 accounts. When the RM criteria were examined individually for Video 2, six were able to discriminate significantly between liars and truth-tellers; these were Total RM ratings, vividness, spatial, temporal, perceptual and cognitive information). In contrast, for Video 1 accounts there were only three equivalent significant effects, but these were in the opposite direction to the predictions of RM theory. In particular, truthful accounts were judged to contain significantly more information regarding cognitive operations, lower amounts affective information, and were given lower overall Total RM scores.

With regard to the raw frequencies, there were fewer significant overall findings. There were, however, significant main effects for perceptual information

for Video 1 accounts and Total RM and spatial information for Video 2 accounts, but these were in the opposite direction to that predicted by RM theory; i.e. deceptive accounts received higher scores than truthful accounts. In sum, not all RM tools were successful in discriminating between truthful and deceptive accounts and, even then, not necessarily in the direction predicted by RM theory.

Whether the statements were derived from individuals who use English as a first or second-language had varied effects for the two types of RM assessment and differed between videos. In Video 1 ratings, deceptive first–language accounts were more likely to be judged as truthful, and truthful accounts as deceptive, and (contrary to the predictions of RM theory) deceptive accounts received lower cognitive ratings than truthful statements. In Video 2 ratings, when second-language statements were assessed, truthful accounts received higher ratings for spatial information than deceptive statements; also, more in line with expectations, when first-language statements were assessed, truthful accounts received higher ratings for affective information than deceptive statements (Table 6.10).

When raw frequencies were considered, the only effects to emerge were in response to Video 2; i.e. second-language statements received significantly higher raw scores for affective information than first-language statements, and first-language statements received significantly higher raw scores for perceptual information than second-language statements, regardless of whether the statements were true or not.

The finding from the Video 1 ratings that deceptive first–language accounts were more likely to be judged as truthful and truthful first-language accounts as deceptive was particularly unexpected since past research has shown that the ability to detect lies is generally poorer when second-language accounts are investigated

and stronger when first-language accounts are assessed (Broadhurst & Cheng, 2005; DaSilva & Leach, 2013; Leach & DaSilva, 2013). However, all that can really be said is that no obvious overall patterns with regard to the effects of language proficiency seemed to emerge from these findings.

In this study generally, it would obviously have helped both methodologically and in terms of interpretation if responses to the two videos had been similar. Overall, only Video 2 produced results consistent with the predictions of the RM approach, and then only for rating scales, which leads to the issue of whether there was some quality about Video 2 which aided more accurate discrimination between truth-tellers and liars. This raises the issue of what it might have been then about the videos that could have affected the scoring of truthful accounts so differently. Both Videos were derived from the same film; hence two of the actors in Video 1 also take part in Video 2. However, Video 1 involves fewer actors, but the interactions between them are very intense (for example, all are involved in a fight). On the other hand, in Video 2 there are more actors but some are passive observers. Video 1 is also slightly longer (i.e. by 42 seconds) than Video 2. Nevertheless, it is difficult to see how these differences could have systematically affected scores on the RM criteria; moreover, if it really is the case that the accuracy of RM criteria in discriminating truth-tellers from liars depends on factors such as type and duration of the story being described, number of protagonists being involved, number and type of interactions between the protagonists, then arguably they are not going to be of much practical use.

Another possibility is that the differences between the two video conditions came about simply because the statements came from different sets of individuals; i.e. it was a feature of the very limited sample size. However, this again draws

attention to a possible inherent weakness in the RM approach. That is, if the application of RM techniques (or at least those used here) can be influenced systematically so much by individual differences in the way people write accounts, irrespective of the truth status of the accounts, then this also potentially severely limits its practical application, especially to individual cases. This raises another issue regarding the general design of the study; i.e. notwithstanding the sample size problem, it could have been the case that the between-subjects design may have overemphasized confounding individual difference variables and militated against obtaining consistent and reliable results.

It can be noted that this same issue arises particularly in studies employing the psychophysiological detection of deception, which are often by necessity conducted on small numbers of cases (BPS, 2004; Porter & ten Brinke, 2010). To overcome the 'between-subjects effect' investigators commonly employ a within-subjects design; that is, they record and contrast the physiological responses of the same person both lying and telling the truth. Although individual differences can still affect the final results (Vrij, 2008a), each case is considered separately; hence such a procedure can also be applied independently to single cases. As noted previously, because of the increased sensitivity of this approach this design is to be found in the bulk of traditional laboratory-based (Alonso, 1992; Bembibre & Higuears, 2012; Manzanero & Diges, 1995) and realistic high-stakes studies of lie-detection (Mann & Vrij, 2006; Mann et al., 2006;Villar et al., 2012; Vrij & Mann, 2001a). It was, therefore, decided to adopt this approach in the next study.

In addition, as also mentioned in the previous chapter, videos, in any case, can potentially be problematic in that, unless they are of individuals telling lies in high stakes situations, their ecological validity may be compromised (De Paulo et al.,

2003; Masip et al., 2005). An alternative, therefore, is to use autobiographical accounts as stimulus materials (see, for example, Ball & O'Callaghan, 2008; Barnier et al., 2005; Johnson et al.,1988; Masip et al., 2005; Sporer & Sharman, 2006); hence this latter approach was also adopted in the next study.

Overall, the findings of this study were very mixed. They are also subject to an obvious limitation in terms of the sample size used. When the study was first designed the idea was that it would be exploratory, to see whether any clear patterns emerged that might be worth following up. Initially using two videos with eight stimulus participants and 13 response participants was considered adequate for this purpose. However, as noted, when clear and significant differences appeared in responses to the videos it seemed most sensible to look at them separately. However, this limited the sample sizes for the stimulus participants even further; so that, for each video, essentially there was only one stimulus participant in each stimulus cell (i.e. first-language truthful, first-language deceptive, etc.). Methodologically this obviously made the study extremely weak; however, it can be noted that small stimulus sample sizes are not unusual in the lie-detection literature; indeed, some studies have used samples of honest and deceptive behaviour from only one individual (Villar et al., 2012; Vrij & Mann, 2001a). The rationale for using small numbers of stimuli is that, in the absence of normative data, if a technique is to have application in the field, it should be powerful enough able to distinguish between truth and lies in small numbers of, if not single, cases. Nevertheless, an alternative way of proceeding would have been to have started again with different videos and a larger sample size; but given the very mixed bag of results, including some low inter-rater reliability indicators, both within and between the videos, and further consideration of the usefulness of videos generally in this context, it was decided

that an extension of the study would have been unlikely to have been very productive and would not have been cost-effective.

To summarize, in many respects, this first study turned out to be a valuable exercise in the problems of conducting RM deception-detection research rather than producing a set of informative and generalisable results. Nevertheless, an attempt was made to apply the lessons learned to the other studies in this thesis. Taken into consideration, in particular, were the nature of the stimulus materials, which should be more variable and ecologically valid, and the use of a within as distinct from between-subjects design; the former being potentially both more sensitive to differences between truthful and deceptive accounts and less sensitive to confounding individual difference variables, and the small sample size of stimulus materials. The significance of the issue of standardisation in scoring was also something that only really emerged as the thesis developed. Moreover, given the rather sparse and conflicting effects for language proficiency, and the fact that there were other perhaps more influential moderating effects to be investigated, the language proficiency variable was not investigated further at this stage (though it is revisited in Study 5). Nevertheless, to avoid any confounding effects of language proficiency, it was decided to use only first-language accounts in the next three studies; moreover, in retrospect, given the possibility that the language proficiency of response participants might affected the evaluation of the accounts, it was decided to apply this criterion to RPs also (i.e. English as a first language only).

Chapter 7

Study 2: Detection of deception by Reality Monitoring, account length and speech rate: Analysis of transcriptions of oral accounts and written statements

7.1. Introduction

The main aims of Study 2 described in this chapter were to investigate whether there is a spoken versus written modality effect in the ability to detect lies using the RM approach. However, in addition, it was decided to investigate the effects of introducing some additional objective lie-detection criteria, namely the length, the duration and the speech rate of the accounts, which have been used in some previous studies of cues to deception, and could be used alongside RM (Dilmon, 2009; Driscell et al., 2013; Nahari et al., 2012; Santtila et al., 1998; Stromwall, et al., 2004; Stromwall & Granhag, 2005; Vrij, et al., 2004; Vrij, et al.,2000). Also, since a cohort of minimally trained coders was again used, it was decided to use rating scale measures of the RM criteria, particularly bearing in mind there were some indications of success using these with regard to Video 2 in Study 1.

7.1.1. *Spoken vs. Written narratives*

Given the many differences between spoken and written language outlined in Chapter 4 there is clearly a need for more research using written statements in RM assessments. As has been noted, RM assessments of written statements are rarely

reported (Barnier et al., 2005; Granhag et al., 2006; Manzanero & Diges, 1995; Nahari et al., 2012; Sporer & Sharman, 2006), and studies that have used them have not differentiated between written statements and transcripts of spoken statements. It is, therefore, unclear why researchers tend to emphasise spoken statements over written statements.

One of the key differences between the oral and written modalities that would be of direct relevance to this chapter is that speakers' speech rate is 10 times faster than writers' (Chafe, 1982); hence speakers' accounts might be expected to be longer in length and likely to contain more RM information than written accounts. Also, in their efforts to offer their self-evaluations of the events they describe, speakers often give subjective views, and use sentences that are richer in words that may reflect uncertainty ('kind of',' may be...'), and contain more *cause* or *because* clauses and first person pronouns: I believe, I assume, etc., (Beaman,, 1984; Burgoon & Buller, 1994; Hancock, Curry, Goorha & Woodworth, 2008; Newman, Pennebaker, Berry & Richards, 2003; Pu, 2006). It seems likely that these three distinctive aspects of spoken language could readily be interpreted as indicators of internal operations, and, therefore, might be coded by judges as cognitive information (an RM indicator of deception). Overall, therefore, one might expect that the RM scores of spoken accounts will be higher than those of written irrespective of their truth status. If so, then if modality is not considered, this could spuriously affect the interpretation RM scores. However, if modality is taken into consideration, there appears to be no a priori reason why one should be any better generally at differentiating between truthful and deceptive statements.

7.1.2. *Using objective measures to detect deception: speech rate, length and duration of the account.*

Although modality effects related to account length and speech rate may potentially confound RM scores, if they do differ between truthful and deceptive accounts then they could presumably also be used alongside RM scores as independent predictors in their own right. For example, they could be combined with more subjective RM ratings to provide an additional more objective, but fairly simple, form of measurement that might be more simple to apply than more complex computer based measures (for examples of the latter see Hauch, Blandón-Gitlin, Masip, & Sporer, 2014; McQuaid, Woodworth, Hutton, Porter, & ten Brinke, 2015; Zhou, Burgoon, Nunamaker, & Twitchell, 2004). Hence, the length and duration of the accounts and the speech rate of the story tellers were also investigated in the present study as possible cues to deception in their own right. On the basis of previous findings in the literature it was predicted that truthful accounts will tend to contain more words and be longer in duration (Dilmon, 2009; Driscell et al, 2013; Nahari et al., 2012; Santtila et al., 1998; Stromwall, et al., 2004; Stromwall & Granhag, 2005; Vrij, et al., 2004).

Speech rate has been also used as a cue to deception in its own right in past research. Typically it has been defined as the number of spoken words divided by the length of interview minus latency period. However, as yet, using this definition, no significant effects for deception-detection have been found (Vrij, et al.,2004; Vrij, et al.,2000). Nevertheless, when Vrij et al. (2008) revised their definition of speech rate by excluding the latency period from their definition (and placing latency as a separate measure) the cue was reliable; that is, when speech rate was

defined simply as the number of words produced per second, the authors showed that liars who were asked to present their account in a reverse order displayed a slower speech-rate than truth-tellers. However, as yet no studies have examined further the utility of this revised definition within the context of a standard RM study; it was hypothesised, therefore, the speech rate of truth-tellers will be faster than that of deceivers.

7.2. Method

7.2.1. *Participants*

Given the weaknesses associated with the design of the first study, and in particular the use of a rather high number of judges evaluating a very small number of statements, it was decided to adopt a more conventional procedure involving a larger sample of statements evaluated by fewer judges (see, for example, Vrij, 2000; 2008; 2015). Consequently, of the two sets of participants employed in this study, 21 were stimulus participants (SPs) and only three were response-participants (RPs). It can be emphasised here that in lie-detection research it is usual to have only two or three people to rate and code behaviours (see, for example, Harpster, et al., 2009; Koper & Sahlman, 1991; Mann et al., 2002; ten Brinke & Porter, 2012; Villar, et al., 2012; Vrij & Mann, 2001a), as it is normally assumed that, if the ratings are not reliable using a few selected individuals, they would be unlikely to statistically discriminate between liars and truth-tellers and have little practical significance.

The 21 SPs were obtained through opportunity sampling from members of the general public and University of Liverpool students (*M* age = 25.80; range = 18-42; *SD* = 7.05); there were eight males and 13 females. There were no exclusion criteria other than participants had to be older than 17 years and speaking English as their

first language. All participants volunteered in response to an advertisement posted in the University website. No imbursement was offered. Of the three RPs or raters (one male and two female), two were research students of Forensic Psychology and one was a chartered Forensic Psychologist (*M* age = 35. 00, range = 31-40, *SD* = 4.58). Although all RPs or judges were familiar with the lie-detection literature, none of the RPs had conducted research in the field of the Reality Monitoring and none had received formal training in RM techniques beyond the instructions provided in this study.

7.2.2. *Materials*

The stimuli were devised as follows (see Appendix 8)

Life Experiences Inventory (LEI).

As noted previously, given the issues raised with regard to the use of videos in Study 1, autobiographical memories were used as the stimuli in the present study; these have been used as stimulus materials in numerous lie-detection studies (for example, Ball & O'Callaghan, 2008; Barnier et al., 2005; Johnson et al, 1988; Santtila et al., 1998; Masip et al., 2005; Sporer & Sharman, 2006). An adaptation of the kind of Life Experiences Inventory (LEI) used by Garry, Manning, Loftus and Sherman (1996) and Paddock, Noel, Terranova, Eber, Manning and Loftus (1999) was, therefore, devised to help participants to generate the stimulus information.

The LEI protocol listed three types of events, 1) having an indoors or outdoors accident, 2) being attacked by an insect/animal, and 3) having an unpleasant medical operation. Some examples of the first and third categories were also provided (such as sports injury, pet run over by a car, lost in a public space for more than an hour,

home broken into, painful dental surgery). Participants were instructed to look at the list of the three types of event, consider if they had previously experienced any of them, and then to perform two tasks according to the following instructions.

First, 'Please describe, in as much detail as possible, *one* of these events that you have experienced in the form of a narrative. If you realise that you have been involved in more than one of these events, please describe the one you remember the best. Your response will be audio recorded and timed. Feel free to ask as many questions as you wish before the task starts BUT remember that no questions will be answered after the timer starts'.

And second, 'Please identify which of these events you have *never* experienced. Please identify *only one* of the events you have never experienced and generate an imaginary story around it. In other words, please create a whole fictitious story and enrich it with as many details as possible to make it look like a genuinely true experience. We would like you to talk about this event so that if someone who did not know whether this event had happened to you were to read your account, they would believe that this event had in fact happened to you. Please remember that your accounts should be freely invented. *You should not* describe friends' experiences, events taken from books or films, personal experiences that had been modified. Your response will be audio recorded and timed. Feel free to ask as many questions as you wish before the task starts BUT remember that no questions will be answered after the timer starts'.

7.2.3 Procedure

The study was conducted in two phases. In the first phase, the stimulus participants (SPs) were invited to take part in a lie-detection study. The participants were

unknown to the experimenter and the exact purpose of the study was unknown to them. Participants were then given the LEI protocol as previously described. When describing a truthful event, participants were also reminded to report an event only of which they were 100% sure. Ten participants were asked to report their accounts orally and the remaining 11 were asked to write down their accounts. Within the constraints of the sample size, deceptive and truthful conditions were also counterbalanced within conditions, so 42 accounts were ultimately recorded. As intimated previously, this was a larger stimulus sample than that used in many deception studies (see, for example, Mann, et al, 2002; Mann & Vrij, 2006; Mann, Vrij, et al, 2006; Villar, et al, 2012; Vrij & Mann, 2001a). Spoken accounts were audio-recorded and transcribed into written form, and all accounts were timed.

In Phase 2, the RPs were given the same information and response sheet used in Study 1 (see section 6.2.2), but they were not asked to score the RM raw frequencies. Hence they were asked only to rate the extent to which they judged the person giving the account to be lying/truthful, and were then presented with the eight RM criteria (i.e. vividness, realism, perceptual information, spatial information, reconstructability, temporal, information regarding cognitive operations and affective information) with their definitions (see Section 6.2.2). Also, following feedback from the RPs in Study 1, the definition of vividness was revised. This was because in its original form the definition encapsulated two criteria, namely clarity and vividness (i.e. "refers to the clarity and vividness of the statement. This criterion is present if the report is vivid, lively, clear and sharp instead of dim, faint, vague and indefinite"). This appeared to confuse some of the coders, so the definition was revised to refer more specifically to vividness; i.e. "vividness: this refers to the

vividness of the statement. This criterion is present if the report is vivid and lively instead of dim and faint" (see Appendix 6b for the revised definition).

No more information was offered to the RPs. Each of the RPs was asked to score all 42 statements.

7.2.4. *Design*

To summarise, a mixed 2 x 2 (modality: written accounts vs. oral accounts x truthfulness: real event vs. fabricated event) design was used with Truthfulness (Truthful/Deceptive) as the within-subjects factor, and Modality (Written statement/Oral testimony) as the between-subjects factor. The dependent variables were the mean ratings of truthfulness of the accounts, and the individual and Total RM scores.

7.3. Results

7.3.1 *Inter-rater reliability for the rating data*

Kendall's coefficient of concordance (W) tests on the Likert ratings showed that there was significant inter-rater agreement between the three RPs for all but one RM criterion (i.e. reconstructability). The results are presented in Table 7.1. Although the concordance ratings were not particularly high (.43 to .66), overall, the level of agreement amongst the judges on the rating-scaled items was considered satisfactory, particularly bearing in mind the fact that minimal training took place beforehand. Consequently, the means of the three RPs ratings were used as data.

Table 7.1. Inter-rater reliability for Global Truthfulness and RM ratings

RM Criterion	*W*
Global Truthfulness	.47*
Vividness	.66*
Perceptual	.63*
Spatial	.50*
Affective	.55*
Reconstructability	.43
Realism	.53*
Temporal	.58*
Cognitive	.59*

*$p < .05$; **$p \leq .01$

7.3.2 *Preliminary analysis of accounts*

Preliminary analyses of objective data recorded directly by the researcher (i.e. not obtained from the RPs) was conducted using 2 x 2 (modality: written accounts vs. oral accounts x truthfulness: real event vs. fabricated event) mixed ANOVAs with repeated measures on the second factor. Latency period was not taken into account when measuring duration (i.e. number of seconds in the account) and speech rate (i.e. the number of words produced per second). Analyses showed that the truthful accounts contained significantly more words (M = 382.62, SD = 260.74) than the deceptive accounts (M = 305.33, SD = 266.67); F (1,19) = 6.94, p = .016; η^2_p =.27. Similarly, the truthful accounts were longer in terms of time spent (in seconds) producing them (M = 362.62, SD = 271.16) than the deceptive accounts (M =

301.14, *SD* = 224.04); *F* (1,19) = 7.54, *p* = .013; η^2_p = .28. Truthful accounts were also more fluent, producing significantly more words per second (*M* = 1.69, *SD* = 1.35) than deceptive accounts (*M* = 1.60, *SD* = 1.33); *F (*1, 19) = 8.30, *p* = .010*;* η^2_p = .30. None of these effects was influenced by the order in which the accounts were presented.

Between-subjects main effects were found for modality. Specifically, oral accounts contained significantly more words (*M* = 510.00, *SD* = 274.90) than written accounts (*M* = 193.04, *SD* = 97.79); *F (*1, 19) = 12.90, *p* = .002, $\eta^2_{p.}$= .40. The written accounts were longer in terms of time spent (in seconds) producing them (*M* = 477.95, *SD* = 250.78) than the oral accounts (*M* = 171.20, *SD* = 83.60); *F* (1, 19) = 13.54, *p* = .002, η^2_p = .42. Speakers' rate of word production was thus higher, producing significantly more words per second (*M* = 2.99, *SD* = .42) than writers (*M* = .42, *SD* = .10); *F (1, 19)* = 395.04, *p* = .001, η^2_p .= . 95. There were no significant interaction effects.

To further explore the relation between the above objective measures of truthfulness (i.e. account-length in words, duration in seconds, words produced per second) and the truth status of the accounts, Pearson's correlations were conducted between these measures. Significant correlations were found between the length in terms of number of words of truthful and the deceptive accounts, *r* = .87, *p* < .01; duration in seconds of truthful and the deceptive accounts, *r* = .93, *p* < .01); and words produced per second in truthful and deceptive accounts (*r* = .99, *p* < .01). These findings indicate that, although differences emerged between truthful and deceptive accounts, there was a tendency for individual differences in participants' responses on these variables to be similar for both truthful and deceptive accounts

(e.g. those who wrote longer truthful accounts also wrote longer deceptive accounts, etc.).

7.3.3. *Global Truthfulness and Reality Monitoring rating scales*

Results for the rating scale data (obtained from the RPs) showed that when asked whether the person describing the accounts was truthful or deceptive (global truthfulness), participants were more accurate in classifying truthful accounts than deceptive accounts, $\chi^2(1) = 8.00$, $p = .005$ (see Table 7.2); however, overall accuracy was not above the chance level $\chi^2(1) = 1.87$, $p > .05$. This reflects a truth bias effect found in previous lie-detection level research (see, for example, Vrij, 2008a).

Table 7.2. Accuracy rates of subjective account classifications as truthful or deceptive (Global Truthfulness)

	Deceptive Accounts (N=21)	Truthful accounts (N=21)	All accounts (N=44)
Correct classifications	38%	81%	60%
Incorrect classifications	62%	19%	40%

The rating scale data were then analysed using a series of 10, 2 x 2 (truthfulness: real event vs. fabricated event × modality: written accounts vs. oral accounts) mixed ANOVAs; i.e. one ANOVA on the assessment truthfulness

judgment item (i.e. rate whether you think the person was a liar/truth-teller), one on the total RM rating scores, and one on each of the eight RM criteria.

Main effects for Truthfulness were found for the Total RM criteria $F(1,19) = 9.15$, $p = .007$ and for vividness $F(1,19) = 6.84$, $p = .018$. Truthful accounts received higher Total RM and vividness scores than deceptive accounts (see Table 7.3).

Table 7.3. Global Truthfulness and RM mean ratings and SD as a function of truthfulness

RM Criterion	Deceptive	Truthful	η^2_p
Global Truth	2.14 (0.54)	2.30(0.59)	.08
Perceptual	1.80 (0.83)	2.00 (0.77)	.04
Vividness	2.05 (0.70)	2.41 (0.64)	.26*
Realism	2.24 (0.54)	2.41 (0.53)	.10
Reconstructability	2.25 (0.38)	2.22 (0.45)	.01
Spatial	1.60 (0.62)	1.57 (0.61)	.01
Affective	1.34 (0.61)	1.35 (0.67)	.00
Cognitive	2.61 (0.87)	2.31 (0.81)	.16
Temporal	1.46 (0.63)	1.76 (0.75)	.15
Total	10.11 (3.61)	11.74 (2.98)	.32**

**p <.05; ** p < .01*

Main effects for Modality were found for the Total RM criteria $F(1,19) = 4.58$, $p = .045$, affective information $F(1,19) = 4.67$, $p = .044$, temporal information $F(1,19) = 8.73$, $p = .008$ and cognitive information $F(1,19) = 6.17$, $p = .016$. In each

case spoken accounts were given higher RM ratings for all criteria except for cognitive information, irrespective of the truth status of the accounts (see Table 7.4). No significant interactions between truth status and modality were found.

Table 7.4. Global Truthfulness and RM mean ratings and SD as a function of modality

Criteria	Written	Oral	η^2_p
Global Truthfulness	2.04 (0.57)	2.41(0.51)	.15
Perceptual	1.68 (0.87)	2.13 (0.66)	.14
Vividness	2.06 (0.74)	2.41 (0.56)	.10
Realism	2.18 (0.46)	2.48 (0.56)	.12
Reconstructability	2.18 (0.41)	2.29 (0.42)	.03
Spatial	1.56 (0.56)	1.62 (0.69)	.01
Affective	1.13 (0.64)	1.58 (0.56)	.20*
Cognitive	2.83 (0.88)	2.06 (0.59)	.27*
Temporal	1.30 (0.59)	1.95 (0.62)	.31**
Total	9.65 (3.39)	12.32 (2.68)	.19*

**p<.05; ** p<.01*

7.4 Discussion

The findings of this study can be summarised as follows. In line with expectations, overall Total RM ratings were significantly higher for truthful than deceptive accounts, though only the criterion of vividness significantly discriminated between the truthful and the deceptive accounts. In these respects, RM assessment

outperformed global truthfulness ratings in discriminating between truthful and deceptive accounts.

With regards to modality, again as predicted, written accounts received lower ratings than spoken accounts for all the RM criteria, apart from cognitive information, and consequently they had lower Total RM scores, irrespective of their truth status. However, the differences in the ratings for the individual criteria reached statistical significance only for the criteria of affective, temporal and cognitive information.

The fact that Total RM and vividness rating scores significantly discriminated between truthful and deceptive accounts is particularly interesting given that previous studies using RM assessments have almost exclusively used highly trained individuals such as judges (see, for example Santtila, et al, 1998; Sporer, 1997; Vrij et al., 2007). In contrast, the response participants of this study were not specifically trained in the RM approach.

The fact that differences in RM ratings were found between written and spoken accounts, regardless of whether the accounts were truthful or deceptive, suggests that modality may provide a potentially highly confounding effect on RM assessment if it is not controlled. The finding of no influence of modality on the ability to discriminate lies from truth is notable given that very few RM studies have used written statements in their procedure (see, for example, Barnier et al, 2005; Granhag et al, 2006; Manzanero & Diges, 1995;Nahari et al., 2012; Sporer & Sharman, 2006). If the present results are at all generalisable, then there appears to be no obvious justification for favouring spoken accounts in this area.

Additional findings indicated that truthful accounts were longer than deceptive accounts in both duration and length. Moreover, the number of words

produced per second was significantly greater for truth-tellers than for liars. This finding was consistent for both writers and speakers. Again it is perhaps somewhat surprising that such criteria have not been considered for RM assessments when the related criterion of number of details is incorporated in the CBCA approach (i.e. quantity of details); indeed, it is one of most diagnostically strong cues (Vrij, 2008a).

It can also be noted again that, in previous studies, speech rate did not discriminate between truthful and deceptive accounts when the definition of speech rate included the time-interval between the interviewer's question and the answer (Vrij et al, 2000; 2004). However, like those of Vrij et al. (2008), the present findings suggest that if the definition of speech rate does not include the latency period it may be diagnostically stronger in distinguishing between truthful and deceptive accounts. Interestingly also, the usefulness of speech rate to discriminate between truthful and deceptive accounts was not affected by modality. Thus although speakers were overall more fluent in terms of speech rate than writers, truth-tellers were also more fluent than liars in both written and spoken accounts, and no interaction between modality and fluency was found. This is a somewhat unexpected finding as one might have predicted that, because of more immediate and salient audience evaluation effects, speakers would have been more eager to appear credible than writers, and might have experienced greater cognitive load than writers. In contrast, writers would have tended to be less involved with their audiences; they were also given more opportunity to process their accounts holistically, and make corrections at their own pace, hence potentially experiencing less cognitive load than the speakers.

Having achieved rather more positive results in this study than the previous one, and with raters who, although having knowledge of forensic psychology, had

received minimal training in RM specifically, attention was turned in the next study to whether results might actually be even better, if raters actually had no knowledge at all of the nature of the task they were to perform.

Chapter 8

Study 3: Does keeping participants blind to the purpose of the study affect RM scores?

8.1. Introduction

Orne (1962) uses the term 'demand characteristics' to refer to those cues in the experimental context that convey information about the exact nature of the study and the direction of the experimental hypothesis. As Orne and others demonstrated, if participants are clearly aware of the purpose of the study and the nature of the measures in this respect, this may substantially confound the results as participants both intentionally and inadvertently interpret and respond to these cues (see, for example, Orne, 1962; Wagstaff, 1981). It may be significant, therefore, as in Studies 1 and 2 here, it is a very common practice in the wider field of lie-detection to ask the judges trained in RM to subjectively ascertain whether the statements are truthful (i.e. produce an overall or global rating) before they score the presence of the RM criteria in the statements (Vrij, 2008a). This presents the possibility that if one is aware of the association between higher RM scores and truthfulness, the global rating may affect RM judgments and vice versa. As a possible illustration of this, consider the data from the previous study; correlational analysis showed that the RPs' subjective truthfulness judgments for the truthful accounts were positively and significantly correlated with the RM ratings for the truthful accounts ($r = .54$, $p < .05$). Similarly, the RPs' subjective truthfulness judgments for the deceptive accounts were positively and significantly correlated with the RM ratings for the deceptive accounts ($r = .58$, $p < .01$). However, in neither case did the truthfulness

ratings themselves significantly discriminate between truthful and deceptive accounts. Although there are various ways these data could be interpreted, they are consistent with the view that RPs might have biased their RM ratings, at least to some extent, to fit with their (apparently inaccurate) preconceptions as to whether the accounts were truthful or not; and this in turn might have reduced the efficacy of the RM ratings in terms of differentiating between truthful and deceptive accounts.

This creates a practical challenge for research on RM using subjective ratings; on the one hand one might expect psychologists or research students with a background in forensic and investigative psychology to be more adept at applying the RM criteria, but, on the other hand, they will also be more likely to be aware of the association between RM scores and truthfulness. The outcome for the application of RM criteria could then go either way depending on how accurate raters are in their global estimates of truthfulness; i.e. if very accurate 'global' estimators are used, this may inflate the accuracy of RM measures in discriminating between truth and lies, but conversely, if very inaccurate global estimators are used this could render the application of the RM criteria relatively ineffective. The purpose of the study described in this chapter, therefore, was to replicate the procedures used in the previous study, but this time keeping participants experimentally blind to the purpose of the study. Given the failure of the global truthfulness ratings to distinguish truthful from deceptive accounts in the previous study, it was tentatively hypothesised that keeping participants blind to the nature of the experiment would allow for a less biased and more objective assessment of the accounts, allowing a more successful discrimination between truthful and deceptive accounts using the RM criteria.

8.2. Method

8.2.1. *Participants*

The SPs (and thereby stimulus accounts) and were the same as those used in Study 2 (see section 7.2.1). However, three new RPs (one male and two females) were recruited (*M* age = 33.00, range = 30-37, *SD* = 3.60. The RPs were postgraduate and post-doctoral researchers, trained in Astrophysics and Philosophy rather than Psychology, to guarantee that they were unfamiliar with the RM approach and blind to the true purpose of the study. Indeed, following the advice of Orne (1962), debriefing after the study indicated that they were unfamiliar with the theory and procedures and were not aware of the purpose of the study. All three RPs used English as their first language.

8.2.2. *Materials and Procedure*

Essentially the procedure was also identical to that used in the previous study, i.e. RPs were asked to score the RM within the statements using the same autobiographical materials; however, there were two major variations.

First, the RPs were told that they were taking part in a study regarding the process of recollecting autobiographical events. Specifically, they were told that the aim of the study was to explore the quality of information found in traumatic autobiographical memories and to assess how and if the type of event described can influence the quality and quantity of memories. And second, the instruction to rate the accounts for truthfulness was removed, as this would obviously affect

participants' perceptions of the purpose of the study (for instructions see Appendix 5).

8.2.3. *Design*

As in the previous study, a mixed 2 x 2 (modality: written accounts vs. oral accounts x truthfulness: real event vs. fabricated event) design was again used, with Truthfulness (Truthful/Deceptive) as the within-subjects factor, and Modality (Written statement/Oral testimony) as the between-subjects factor. The dependent variables were the mean ratings of the individual and Total RM subjective rating scores.

8.3 Results

8.3.1. *Inter-rater reliability for the RM ratings*

Table 8.1 Inter-rater reliability for the RM ratings

RM Criterion	*W*
Vividness	.42*
Perceptual	.50*
Spatial	.51*
Affective	.54*
Reconstructability	.40
Realism	.50*
Temporal	.54*
Cognitive	.61*

*$p < .05$

Kendall's coefficient of concordance (*W*) tests showed that there was significant inter-rater agreement between the three RPs for all but one criterion (i.e. reconstructability). These results are presented in Table 8.1. Overall, the level of agreement amongst the judges on the rating-scaled items was considered satisfactory (W = .40 to .61); although agreement was overall somewhat lower in comparison with that of the previous study it was not significantly so (Sign Test, p = .289). Consequently, the mean scores of the RPs were used as data.

8.3.2 *Reality Monitoring rating scales*

Table 8.2 RM mean ratings and *SD* as a function of truthfulness

RM Criterion	Deceptive	Truthful	η^2_p
Perceptual	1.84 (0.81)	2.02 (0.67)	.03
Vividness	2.16 (0.62)	2.16 (0.48)	.00
Realism	2.43 (0.80)	2.21 (0.59)	.07
Reconstructability	2.28 (0.65)	2.06 (0.68)	.07
Spatial	1.78 (0.78)	1.78 (0.60)	.00
Affective	1.51 (0.54)	1.49 (0.60)	.01
Cognitive	2.16 (0.87)	2.00 (0.83)	.34
Temporal	1.69 (0.90)	1.88 (0.86)	.39
Total	11.52 (3.89)	11.58 (3.10)	.00

The rating scale data were analysed using a series of nine, 2 (truthfulness: truthful/deceptive) × 2 (modality: written accounts/oral accounts) mixed ANOVAs;

i.e. one ANOVA on the total RM rating scores, and one on each of the eight Reality Monitoring criteria. No main effects were found for Truthfulness or Modality (see Tables 8.2 and 8.3), and no interaction effects between truthfulness and modality were found.

Table 8.3 RM mean ratings and *SD* as a function of modality

Criteria	Written	Oral	η^2_p
Perceptual	1.94 (0.75)	1.90 (0.77)	.01
Vividness	2.22 (0.52)	2.10 (0.56)	.02
Realism	2.49 (0.58)	2.14 (0.78)	.10
Reconstructability	2.29 (0.51)	2.04 (0.40)	.06
Spatial	1.73 (0.70)	1.84 (0.69)	.01
Affective	1.37 (0.73)	1.65 (0.66)	.08
Cognitive	2.22 (0.86)	1.94 (0.84)	.04
Temporal	1.84 (0.85)	1.72 (0.75)	.01
Total	11.65 (3.07)	11.45 (3.98)	.00

8.3.3 *A comparison of the two data sets: blind and non-blind participants*

Thus far, the results from this and the previous study appear to indicate that, not withstanding any negative effects of including a global truthfulness rating measurement, participants with a psychology background could discriminate accurately truthful from deceptive accounts to some extent but equally qualified judges with no psychology background and with no knowledge of the purpose of the

study could not. A statistical comparison of the two groups was, therefore, conducted to clarify whether their scores differed significantly.

The rating scale data were, therefore, analysed using a series of nine, 2 (truthfulness: truthful/deceptive) × 2 (modality: written/spoken accounts) x 2 (blind/not blind) mixed ANOVAs; i.e. one ANOVA on the total RM rating scores, and one on each of the eight RM criteria. No main effects for truthfulness, modality or blindness were found.

However, in recent years a number of researchers have stressed the advantages of using effect sizes to estimate the size or strength of a phenomenon when comparing different samples. One of the main advantages of effect size in this respect is that provides an estimate of the strength of phenomena independently of sample size, which can be critical when, as here, the sample sizes were small (Cohen, 1988; 1990).

In the present context the most critical comparison is that which concerns the relative ability of the different types of RPs to use RM ratings to discriminate truthful from deceptive accounts; i.e. the results in Tables 7.3 and 8.2. For clarity these are presented again alongside each other here in a single table, Table 8.4.

According to Cohen's (1988) criteria, four of the effect sizes for the non-blind sample were large and in the direction predicted by RM theory (vividness, cognitive, temporal and total). In contrast, only two of the effects sizes in the blind sample could be classified as large (cognitive and temporal); but particularly notable was the fact that whereas the total RM scores showed a large effect size in the non-blind sample, the effect size was more or less zero in the blind group.

Table 8.4 Global Truthfulness, RM mean ratings (with SD) as a function of truthfulness for experimentally blind and non-blind response participants.

Non-blind RPs

RM Criterion	Deceptive	Truthful	η^2_p
Perceptual	1.80 (0.83)	2.00 (0.77)	.04
Vividness	2.05 (0.70)	2.41 (0.64)	.26*
Realism	2.24 (0.54)	2.41 (0.53)	.10
Reconstructability	2.25 (0.38)	2.22 (0.45)	.01
Spatial	1.60 (0.62)	1.57 (0.61)	.01
Affective	1.34 (0.61)	1.35 (0.67)	.00
Cognitive	2.61 (0.87)	2.31 (0.81)	.16
Temporal	1.46 (0.63)	1.76 (0.75)	.15
Total	10.11 (3.61)	11.74 (2.98)	.32**

**p<.05; ** p<.01*

Blind RPs

RM Criterion	Deceptive	Truthful	η^2_p
Perceptual	1.84 (0.81)	2.02 (0.67)	.03
Vividness	2.16 (0.62)	2.16 (0.48)	.00
Realism	2.43 (0.80)	2.21 (0.59)	.07
Reconstructability	2.28 (0.65)	2.06 (0.68)	.07
Spatial	1.78 (0.78)	1.78 (0.60)	.00
Affective	1.51 (0.54)	1.49 (0.60)	.01
Cognitive	2.16 (0.87)	2.00 (0.83)	.34
Temporal	1.69 (0.90)	1.88 (0.86)	.39
Total	11.52 (3.89)	11.58 (3.10)	.00

8.4. Discussion

In contrast with those of Study 2, the results showed that when judges were kept blind to the purpose of the study, RM assessments did not significantly distinguish

truthful from deceptive accounts on ratings of any of the RM criteria. Also, the modality effects found in the previous chapter were eliminated here.

Once again one could question the generality of the results because of the small samples of raters; but as noted previously, as here, studies in this area typically employ only one, two or three raters (see, for example, Harpster et al., 2009; Koper & Sahlman, 1991; Mann et al.,2002; ten Brinke & Porter, 2012; Villar, et al., 2012; Vrij & Mann, 2001). Moreover, in both this and the previous study, the RPs were selected carefully in terms of their competence to understand the instructions and conduct the tasks assigned to them (all were University graduates), and overall the two sets of raters showed a satisfactory level of agreement between themselves in their ratings. Again also, the point can be made that if the kinds of RM rating procedures used here are so sensitive to individual differences in the way highly educated and competent raters apply the criteria as to render the results meaningless, then such procedures would clearly have little practical value.

In sum, the results of the present study find no support for the idea that using raters who are blind to the purpose of the study facilitates the accurate application of the RM criteria; i.e. there was no evidence that the blind RPs actually performed better than the non-blind RPs. It could be argued, therefore, that if the results are valid, either the ratings of non-blind judges were not influenced by explicit or implicit global judgments about the truthfulness of the stimulus information, or, if they were, this was perhaps offset by the influence of their expertise in the area.

In fact, what trends there are indicate that having some background knowledge about the subject matter may, on balance, be beneficial in applying the RM criteria ratings (see also Masip et al., 2005).

Chapter 9

Study 4: The role of account length in detecting deception in written and spoken autobiographical accounts using Reality Monitoring

9.1. Introduction

Having concentrated on subjective ratings in the last two Chapters, the aim Study 4 as described in this chapter was to return to the issue of the usefulness of RM raw frequency counts in discriminating between truthful and deceptive accounts, and in particular, the effects of standardisation.

As mentioned previously, despite some evidence indicating that the length of accounts per se may be a reliable cue to deception (DePaulo et al., 2003; Porter & Yuille, 1996; Vrij, et al, 2004; Vrij et al., 2000), to the author's knowledge, there has been no research in the area of RM that has looked systematically at the effects of standardisation of word-count on RM outcomes. In fact, the application of word-count standardisation appears to have been very arbitrary and ad hoc. For example, sometimes authors standardise the RM scores per 100 (Larson & Granhag, 2005; Strömwall & Granhag, 2005; Vrij et al., 2004) or 50 words (Vrij et al., 2000), sometimes they do not control for length differences and instead standardise per account duration (Gnisci et al., 2010) or do not standardise at all (Nahari et al., 2012). In other words, there appear to be multiple ways to standardise for account differences in length (or duration) but there is no consensus as to whether, when and how to do it. This creates confusion amongst researchers who have at times even given up using raw frequencies when account differences in length are significant (Granhag et al; 2006; Stromwall et al., 2004).

It is also important to note that the decision to standardise for word-count or not is not simply a statistical or methodological issue; it is also conceptual. For instance, although some writers have advocated some kind of standardisation as a general principle when the length of accounts differs (for example, Strömwall & Granhag, 2005; Sporer, 2004), it does not necessarily follow that this makes sense conceptually. The basic rationale for standardisation is that RM differences between truthful and deceptive accounts could simply be an artefact of general differences in length and density of words contained in the account. However, whilst this might appear logical, it arguably makes little sense to correct for word-count if length per se is considered to be a reliable cue to deception (Colwell et al, 2002; Memon et al., 2010; Vrij et al., 2004). Moreover, RM was originally formulated on the idea that "memories based in perception have better spatial, temporal, and sensory information" (Johnson & Raye, 1981, p.82). Notably, there is no reference in the seminal RM papers by Johnson and colleagues (Johnson, et al, 1993; Johnson & Raye, 1981) to the idea that truthful accounts will be richer in the *density* of RM criteria; but rather they will overall contain "more perceptual, spatial and temporal, semantic and affective information and less information about cognitive operations" (Johnson et al., 1993, p.4). It could be argued, therefore, that standardising for word-count differences is essentially an intervention that is not supported by the original theory, as it alters one of the core qualities of lies (i.e. they generally contain less information).

Given this lack of clarity about how to deal with what appears to be a fundamental methodological issue in the operation of RM, it is difficult to see how one could possibly operationally apply RM measures within the criminal justice system as a way of discriminating truth from lies in accounts. Yet, although a

number of researchers have recognised the problem, as just noted, there appears to have been few, if any, systematic research conducted on this issue. In particular, we need to know exactly how the standardisation of accounts for number of words affects the role of RM criteria (both singularly and in combination) in discriminating lies from truths. Though, interestingly, in the few instances where results before and after standardisation for word-count are available (Larson & Granhag, 2005; Masip et al., 2005), the trend has been for unstandardised raw frequency scores to be more accurate in terms of applying RM criteria (Elntib et al., 2014; Memon et al., 2010).

The effects of standardisation may also vary according to modality. For example, one might expect spoken accounts to display more information relevant to the RM criteria; i.e. unless standardised for length, oral accounts should receive higher RM scores than written accounts, irrespective of their veracity as overall they contain more words. However, in general, the evidence suggests that, when standardised for length, oral narratives tend to have lower lexical density than written accounts, as they are not as well planned (and corrected, etc.); hence they generally contain numerous pauses, false starts, incomplete sentences repetitions and hesitations (Chafe & Tannen, 1987; Halliday, 1989; 2001). Thus written discourse is more coherent and dense in terms of content-words per clause than spoken language. It might, therefore, be predicted that, without standardisation, oral accounts, being longer, will tend to receive higher RM scores. However, as they are not as dense lexically, when word-count standardisation takes place, they will tend to receive lower RM scores than written accounts. An exception to this might be found with the cognitive operations criterion. This should be present more often in oral accounts than written accounts, both before and after standardisation, as oral accounts tend to overall contain more first person pronouns, silent pauses and verbal fillers (e.g. “um,

uh"), and words that may reflect uncertainty and hesitation (e.g. "kind of, may be...") and subjective assumptions (e.g. "it seemed to me that…"; Pu, 2006). On the other hand, written accounts are generally better prepared; hence they tend not to contain words that reflect hesitation and uncertainty (Pu, 2006). This may be important, as it has been suggested that accounts rich in words that reflect equivocation or uncertainty in response to an open question are generally interpreted as associated with deception in lie-detection settings (Adams & Jarvis, 2006; DePaulo et al., 2003). And significantly, within the RM framework, words used to express uncertainty and subjectivity ("I think that he must have been present because…") are coded as cognitive operations. The presence of such words will, therefore, tend to increase both the density and presence of cognitive operation items coded in the accounts.

Given these considerations, the aim of the present study was to conduct a brief preliminary investigation to determine systematically whether standardising accounts for word-count affects the usefulness of the RM approach in discriminating between truthful and deceptive accounts, and whether this is moderated by the modality of the accounts; i.e. whether they are oral or written. Specifically, it was hypothesised that the RM criteria will be more effective in discriminating between truthful and deceptive accounts before word-count standardisation than after. In addition, oral accounts will tend to produce higher RM scores than written accounts before word-count standardisation (because they are longer), but lower RM scores after word-count standardisation (because they are less dense). And finally, oral accounts will be richer in information regarding cognitive operations both before and after word-count standardisation.

9.2. Method

9.2.1 *Participants*

The stimulus participants SPs (and thereby stimulus materials) were as described in the previous two studies (see Section 7.2.1). Two RPs who used English as their first language were employed (one male and one female). Both were postgraduate research students of Forensic Psychology with general expertise in lie-detection (*M* age = 36. 50, range = 33-40, *SD* = 4.95). As in the previous studies no formal training was given beyond the standard instructions used here; however, both RPs were considerably well-trained in coding cues to deception and had previous knowledge of RM-based lie-detection. Considering their expertise (and following their very high inter-coder agreement) it was decided that a third judge was not needed. The decision to use more expert coders here was based partly on the results of the previous study, but more particularly because the aim of the present study was to look specifically at the effects of standardisation, not to examine the efficacy of RM procedures with relatively untrained judges.

9.2.2. *Materials and procedure*

The stimulus materials were also as described in the previous two studies; i.e. 42 autobiographical accounts derived from an adaptation of the LEI (see section 7.2.2). The RM coding framework used in the two previous studies was also used.

The two RPs were given an opportunity to familiarise themselves with the particular RM coding protocol used here before beginning the scoring process. Both coders were blind as to the truth status of the accounts, or whether they were oral or written. It can be noted that verbal fillers (for example *um, uh*) were only present in the oral transcripts; these fillers were not removed from the transcripts, but the

coders were not aware of their purpose and function, or that they were confined to oral testimonies.

9.2.3. *Design*

To summarise, a mixed 2 x 2 (modality: written accounts vs. oral accounts x truthfulness: real event vs. fabricated event) design was used, with Truthfulness (Truthful/Deceptive) as the within-subjects factor, and Modality (Written statement/Oral testimony) as the between-subjects factor. The five RM criteria and the Total RM frequency scores (unstandardised and standardised per 100 words) were the dependent variables.

9.3. Results

9.3.1 *Inter-rater reliability for the raw frequency data*

The principal coder scored all accounts whereas the secondary coder scored 25% of the accounts, including truthful, deceptive, oral and written accounts. Intra-class correlation agreement and Pearson's correlations showed that, in terms of applying the RM criteria to the various measures, there was high and significant inter-coder agreement between the two judges (ICC = .90-.96; r = .84-.96). Considering the nearly-perfect inter-coder agreement for the random sample, frequency counts produced by the principal coder were used in all analyses.

9.3.2. *RM results before word-count standardisation*

The RM frequencies before word-count standardisation were analysed using a series of six 2 x 2 (modality: written accounts vs. oral accounts x truthfulness: truthful/deceptive) mixed ANOVAs with repeated measures on the second factor;

one for each of the five RM criteria, and one for the Total RM score. Total RM scores were calculated by adding scores for perceptual, spatial, affective, and temporal information and deducting scores for cognitive operations.

As predicted by RM theory, mean scores were higher for the truthful accounts for all RM criteria with the exception of cognitive operations. In particular, significant effects were found for Total RM scores $F(1,19) = 18.05$, $p = .001$, spatial information $F(1,19) = 17.79$, $p = .001$ and temporal information $F(1,19) = 8.32$, $p = .01$ (see Table 9.1).

With regard to modality, for all RM criteria, frequencies were higher for oral accounts before standardisation; however, a significant main effect for modality was found only for cognitive information $F(1,19) = 8.38$, $p = .009$, though there was a near significant trend for temporal information $F(1,19) = 4.22$, $p = .054$ (see Table 9.2). No interactions between truthfulness and modality were found for any of the analyses.

9.3.3. *Results after word-count standardisation*

To standardise word-count, the RM raw scores were re-calculated per 100 words of account (for examples of this method, see Larsson & Granhag, 2005; Strömwall & Granhag, 2005; Vrij et. al, 2004). The resulting frequencies after standardisation were again analysed using a series of six, 2 X 2 (modality: written accounts/oral accounts X truthfulness: truthful/deceptive) mixed ANOVAs with repeated measures on the second factor. Analyses of the RM scores after standardisation showed no significant effects, or effects approaching significance, for truthfulness, for any of the RM criteria, including Total scores (see Table 9.1).

Table 9.1. RM mean (*SD*) frequency counts as a function of truthfulness before and after standardisation of word count

RM Criterion	Deceptive	Truthful	η^2_p	Deceptive	Truthful	η^2_p
	Before standardisation			After standardisation		
Perceptual	5.36 (4.50)	6.40 (4.01)	.04	2.00 (1.41)	2.05 (1.16)	.01
Spatial	11.48 (9.55)	16.48 (10.89)	.47**	4.26 (2.12)	4.91 (2.12)	.11
Affective	3.79 (3.91)	4.26 (3.93)	.01	1.34 (1.66)	1.12 (1.23)	.01
Cognitive	6.48 (9.33)	5.64 (6.71)	.03	1.65 (1.03)	1.27 (0.96)	.11
Temporal	8.24 (6.80)	11.50 (5.49)	.31*	2.98 (1.59)	3.24 (1.12)	.04
Total	22.38 (14.56)	32.48 (18.00)	.49**	8.94 (4.35)	9.93 (3.88)	.06

*$p < .05$; **$p < .01$

Table 9.2. RM mean (SD) frequency counts as a function of modality before and after standardisation of word count

RM Criterion	Written	Oral	η^2_p	Written	Oral	η^2_p
	Before word count standardisation			After word count standardisation		
Perceptual	5.30(4.06)	6.56 (4.56)	.04	2.65 (1.34)	1.34 (0.76)	.55*
Spatial	10.34 (5.67)	17.98(12.75)	.16	5.61 (2.32)	3.47 (1.07)	.34*
Affective	3.04 (3.55)	5.10 (4.09)	.10	1.47 (1.94)	.98 (0.56)	.06
Cognitive	2.01 (2.32)	10.45 (9.70)	.31*	1.06 (1.06)	1.92 (0.73)	.28*
Temporal	7.00 (3.75)	13.00 (9.64)	.18	3.79 (1.28)	2.37 (1.00)	.39*
Total	23.11 11.58)	32.18(19.80)	.10	12.33 (3.30)	6.25 (1.93)	.76*

*$p < .01$

Moreover, whereas very substantive effect sizes were found for three of the RM measures before standardisation (spatial temporal and total), none was found after standardisation. This supports the key hypothesis regarding the superior diagnostic strength of the RM approach before standardisation for word-count.

Main effects for modality were found for perceptual information $F(1,19) = 23.19$, $p = .001$; spatial information $F(1,19) = 9.69$, $p = .006$; cognitive information $F(1,19) = 7.41$, $p = .014$, temporal information $F(1,19) = 12.12$, $p = .002$ and Total RM scores $F(1,19) = 59.09$, $p = .001$ (see Table 9.2). In contrast with results before standardisation, in each case, written accounts were denser in information relating to RM criteria. Moreover, the effect sizes for modality effects after standardisation were generally considerably larger than those before standardisation, suggesting that standardisation exaggerates the differences between spoken and written accounts. No significant interactions were found.

9.4. Discussion

The present results suggest that the ability of RM criteria to discriminate between truthful and deceptive accounts is affected by word-count or length standardisation. As hypothesised, total RM scores, spatial information and temporal information were more effective in discriminating between truthful and deceptive accounts before word-count standardisation than after. In fact, none of the criteria differentiated between truthful and deceptive accounts after standardisation. Moreover, these effects were not moderated by modality. In general, therefore, the present results appear to lend some support to previous findings suggesting that,

when raw frequencies are used, some RM criteria are more able to discriminate between truthful and deceptive accounts if there is no attempt to control for word-count (for example Masip et al., 2005; Larsson & Granhag, 2005)/

If the present results have any generality in this respect, they may have some interesting implications for the diagnostic use of RM criteria. As mentioned previously, it could be argued that standardising accounts for word-count constitutes an intervention which is not fully justified by the original theory (Johnson & Raye, 1981; Johnson et al., 1993). However, one of the drawbacks of using unstandardized frequencies, is that they make it difficult to establish normative criteria for comparisons within and across studies, and for the assessment of individual cases. In contrast, if all relevant studies used a standardised measure of the raw frequencies (for example per 100 words) then, in principle, researchers might be able to establish normative data for truthful and deceptive accounts against which individual cases could be compared. But this might also present something of a paradox for researchers and practitioners; there is little point standardising scores if to do so would mean rendering RM criteria relatively ineffective in predicting truthfulness.

Importantly, the present results also showed that there was no significant difference in the ability of oral and written accounts to discriminate between truthful and deceptive accounts (i.e. there were no significant interactions between truthfulness and modality), either before or after standardisation. This would suggest that either could potentially be used to help establish truthfulness at a broad statistical level, if the RM criteria are applied without standardisation for word-count. However, the task of establishing RM criteria through which to judge individual cases, remains a very significant challenge if RM is to be applied in the field.

In addition to the findings regarding the effects of word-count, there was some support for the prediction that, regardless of whether accounts are truthful or deceptive, oral accounts tend to be longer and, therefore, produce higher RM raw frequency scores than written accounts before word-count standardisation (all means were in the appropriate direction, but a significant effect was only found for cognitive information). Remarkably, this position was reversed after word-count standardisation; thus, after word-count standardisation, regardless of the truthfulness of the accounts, total RM scores were significantly higher for written accounts; i.e. written accounts were denser in terms of temporal, spatial, perceptual and total RM scores. This is in line with the rationale provided earlier regarding the different roles of speakers and writers (Halliday, 1989, 2001; Pu, 1996); and an obvious implication of these findings is that oral and written accounts should never be treated as equivalent either within, or across studies (see, for example, Granhag et al, 2006; Manzanero & Diges, 1995). Moreover, if written accounts overall tend to be more dense in terms of RM criteria than the oral accounts, and are easier to process, one might question why written accounts are used so rarely in RM research.

Finally, it can be noted that there was support for the prediction that oral accounts will be richer in information regarding cognitive operations both before and after word-count standardisation. The results showed a significant effect of modality on cognitive information both before and after word-count standardisation, such that oral accounts showed more cognitive information. Given the general failure of cognitive information to predict truthfulness, either before or after word-count standardisation, this emphasizes again the importance of not assuming that written and oral accounts are equivalent in terms of their effects on RM scores. For example, a simple comparison of an oral with a written account might give the

spurious impression that the oral account is more likely to be deceptive if the criterion of cognitive information is used as a cue. This may be particularly significant in that the cognitive information criterion is often considered the weakest RM criterion for predicting truthfulness (Granhag et al., 2006; Masip et al., 2005; Vrij et al., 2000; Vrij, 2008a) and there has been some scepticism surrounding it's use (i.e. Masip et al., 2005; Sporer & Sharman, 2006; Vrij et al, 2004). For example, contrary to the RM theory, cognitive information scores have often been found to be higher in truthful accounts than in the deceptive ones (Masip et al., 2005).

To conclude, at the very least, the present results suggest that when judging the truthfulness of accounts using RM criteria, treatment of word-count, and the modality in which an account is presented, appear to be variables that should be taken into consideration if researchers wish to compare and replicate findings within and across studies.

Chapter 10

Study 5: Standardisation as a moderator: Revisiting data from Study 1

10.1 Introduction

As highlighted in the previous Chapter, researchers in verbal lie-detection, and RM in particular, have generally neglected the issue of standardisation; however, in the few instances where results before and after standardisation for word-count were presented the trend has been for unstandardised raw frequency scores to be more predictive in terms of applying RM criteria (Elntib et al., 2014; Memon et al., 2010). The findings from the previous study supported this general trend. Study 5 as described in this chapter, therefore, revisited this issue with regard to the data from Study 1 (Chapter 6), which for reasons most salient at the time used only standardised scores (frequencies per 100 words) for the RM raw frequency data (on the recommendation of researchers such as, Granhag et al., 2006; Stromwall et al., 2004; Porter & Yuille, 1996). Study 5, therefore, examines whether the relevant results from Study 1 might have been different, and more in line with the predictions of RM, if unstandardised raw frequency scores had been used.

10.2. Method

The participants, materials and procedure were those from Study 1, as follows (for full details see Section 6.1).

10.2.1. *Participants*

There were eight Stimulus Participants (SP) and 13 Response Participants (RPs).

10.2.2. *Materials and Procedure*

To recap, the study was conducted in two phases. In the first phase, four native and four non-native stimulus participants were asked to view one of two videos and then to either provide a written statement involving their full recollection of the video-scene, or to fabricate a story and then provide a written statement based on this fabricated story. In Phase 2, the 13 RPs were asked to code the truthful and deceptive statements.

10.2.3. *Design*

A 2 x 2 x 2 mixed design was employed; truthfulness (truthful/deceptive) and Videos 1 and 2 were the within-subjects factors, and language proficiency (English as a first/second-language) was the between-subjects factor.

10.3. Results

For reasons stated in Chapter 6, the data for the two videos were again analysed separately.

10.3.1. *Raw frequency-scores for Video 1 before standardisation.*

Results for Video 1 before and after standardisation are presented in Table 10.1 (note that the results after standardisation have already been described in Chapter 6). The RM frequencies before word-count standardisation were analysed using a series of six 2 × 2 (language proficiency: first/second-language accounts × truthfulness: truthful/deceptive event) mixed ANOVAs with repeated measures on the second factor; one for each of the five RM criteria, and one for the Total RM scores.

Table 10.1. RM mean (SD) raw frequencies as a function of truthfulness before and after standardisation

Unstandardized				Standardised per 100 words		
RM Criterion	Deceptive	Truthful	η^2_p	Deceptive	Truthful	η^2_p
Perceptual	8.23 (5.31)	6.54 (6.10)	.03	2.80 (1.62)	5.58 (4.83)	.28
Spatial	8.84 (4.83)	5.90 (5.39)	.37*	3.18 (2.09)	5.15 (3.41)	.19
Affective	5.81 (2.50)	0.31(0.63)	.85**	2.02 (0.91)	0.29 (0.59)	.82**
Cognitive	2.69 (4.03)	0.46 (0.48)	.27	1.01 (1.70)	0.20 (0.33)	.24
Temporal	5.46 (2.37)	2.46 (2.66)	.65**	1.93 (0.95)	1.44 (1.48)	.20
Total	25.59 (9.69)	14.76 (7.93)	.59**	8.92(3.35)	12.26 (6.56)	.21

*$p < .05$; **$p < .01$

Total RM scores were calculated by adding scores for perceptual, spatial, affective, and temporal information and deducting scores for cognitive operations. Contrary to RM theory, scores were higher for the deceptive accounts for all RM criteria. Significant effects were found for Total RM scores $F(1,11) = 15.51$, $p = .001$, spatial information $F(1,11) = 6.55$, $p = .027$, affective $F(1,11) = 62.56$, $p = .001$ and temporal information $F(1,11) = 20.73$, $p = .001$. There was no evidence that the use of unstandardised scores improved the predictive efficacy of RM in the direction suggested by RM theory; indeed, if the number of significant results and related effect sizes are anything to go on (see Table 10.1), they appear to make the situation considerably worse.

No significant effects for language proficiency were found; both first and second-language accounts receiving similar scores (Table 10.2). And no interactions between truthfulness and modality were found for any of the analyses.

10.3.2. *Raw frequency-scores for Video 2 before standardisation.*

The Video 2 RM frequencies before word standardisation were then analysed in the same way with a series of six 2 × 2 (language proficiency: first/second-language accounts × truthfulness: truthful/deceptive) mixed ANOVAs. In line with the RM theory, scores were higher for the truthful accounts for all RM criteria. Significant effects were found for Total RM scores $F(1,11) = 12.01$, $p = .005$, perceptual information $F(1,11) = 5.12$, $p = .045$, spatial information $F(1,11) = 8.81$, $p = .013$ and temporal information $F(1,11) = 42.73$, $p = .001$ (see Table 10.3 which again also shows the results for standardised scores).

Table 10.2. RM mean (*SD*) raw frequencies as a function of language proficiency (L1, first language; L2 second language)

	Unstandardized			Standardised per 100 words		
RM Criterion	L1 accounts	L2 accounts	η^2_p	L1 accounts	L2 accounts	η^2_p
Perceptual	7.50 (5.12)	7.20 (5.94)	.01	4.08 (2.78)	4.35 (4.02)	.01
Spatial	6.81 (4.20)	8.30 (3.55)	.05	3.99 (2.72)	4.45 (1.69)	.01
Affective	3.25 (1.74)	2.80 (1.20)	.03	1.14(0.76)	1.19 (0.51)	.01
Cognitive	0.81 (0.88)	2.88 (3.47)	.24	0.24(0.26)	1.19 (1.47)	.30
Temporal	3.12 (2.18)	5.30 (2.40)	.24	1.34(1.19)	2.25 (1.01)	.17
Total	19.87 (9.61)	20.80 (7.41)	.01	10.30 (5.17)	11.05 (4.60)	.01

Table 10.3. RM mean (*SD*) raw frequencies as a function of truthfulness before and after standardisation

Unstandardized				Standardised per 100 words		
RM Criterion	Deceptive	Truthful	η^2_p	Deceptive	Truthful	η^2_p
Perceptual	6.30 (4.19)	11.92 (11.76)	.32*	11.90 (8.96)	6.38 (6.63)	.15
Spatial	4.15 (1.34)	6.69 (3.27)	.44*	6.55 (2.89)	3.54 (1.80)	.45*
Affective	1.31 (2.17)	1.46 (1.26)	.01	1.26 (2.09)	0.78 (0.69)	.31
Cognitive	0.85(1.46)	1.54(0.97)	.15	0.89 (1.45)	0.81 (0.51)	.06
Temporal	1.38 (1.39)	4.54 (2.02)	.79**	1.73 (1.32)	2.40 (1.15)	.25
Total	12.31 (2.72)	23.07 (17.57)	.52**	20.54 (9.42)	12.29 (8.01)	.31*

*$p < .05$; **$p < .01$

On balance, the effect sizes for the significant results tended to be bigger for the unstandardised scores; in general, therefore, one could argue that the unstandardised scores had the advantage in terms of the predictive efficacy of the RM criteria.

Video 2 second-language accounts overall received higher RM scores than first-language accounts. Significant main effects for language proficiency were, nevertheless, found only for affective, $F(1,11) = 14.56$, $p = .003$, and temporal, $F(1,11) = 4.95$, $p = .048$, information (Table 10.4). However, in general, looking at the results in Table 10.4, it is difficult to detect any systematic patterns of differences between the unstandardised and standardised accounts in terms of the effects of language proficiency. For example, results appear to go in distinctly opposite directions for perceptual and affective information, making it difficult to draw any meaningful conclusions.

Table 10.4. RM mean (*SD*) raw frequencies as a function of language proficiency (L1, first language; L2 second language)

	Unstandardized			Standardised per 100 words		
RM Criterion	L1 accounts	L2 accounts	η^2_p	L1 accounts	L2 accounts	η^2_p
Perceptual	8.75 (3.40)	9.70 (9.85)	.01	11.22 (3.17)	5.81 (5.96)	.42*
Spatial	4.87 (2.34)	6.30 (2.14)	.15	5.44 (1.94)	4.42 (1.51)	.13
Affective	0.50(0.60)	2.80 (1.70)	.57**	0.25 (0.30)	2.25 (1.41)	.62**
Cognitive	0.81 (0.70)	1.80 (1.38)	.25	.50 (0.62)	1.41 (1.15)	.28
Temporal	2.31 (1.01)	4.00 (2.06)	.31*	1.62 (0.89)	2.77 (1.46)	.27
Total	15.62 (5.24)	21.00 (10.07)	.16	18.03(4.06)	13.84 (6.18)	.27

*p < .05; **p < .01

10.4. Discussion

This reanalysis of the results of Study 1 did not support the view that unstandardised RM frequency scores would predict truthfulness better than standardised scores, at least in the direction predicted by RM theory. Indeed, if anything, the use of unstandardised scores seemed to exaggerate the different effects for the two videos reported previously for Study 1. Thus for Video 1, the unstandardised scores produced a number of effects directly opposing the predictions of RM theory; whereas, for Video 2, the trends were generally more supportive of the predictions of RM theory. Once again, it is difficult to understand what it was about the particular videos, or individuals observing them, that might have produced these results, or why now they were exaggerated when the scores were not standardised. However, with the benefit of hindsight, it may be possible to come up with a relatively simple but plausible explanation based on account length.

As a general finding, it has been established that, other things being equal, longer accounts tend to be richer in terms of RM criteria, and regardless of whether those producing them are telling the truth or being deceptive, some individuals give longer accounts than others. This is less of a problem when using a within-subjects design, where SPs produce both truthful and deceptive accounts, as this will tend to minimise the effects of individual differences in this respect. It is, however, potentially very problematic for between-subjects designs, especially when small numbers of SP accounts are used, as in Study 1. To illustrate this, Table 10.5 shows the design for Study 1, together with the unstandardised word-counts for each of the accounts produced by the eight SPs (i.e. it should be remembered that each of these is an individual SP).

Table 10.5. Design used in Study 1

(number of words in accounts)

Judges		Video 1		Video 2	
		D	T	D	T
N=8	E1	SP1	SP2	SP3	SP4
		(332)	(107)	(50)	(200)
N=5	E2	SP5	SP6	SP7	SP8
		(236)	(132)	(104)	(178)

D=Deceptive account, T= Truthful account
E1= English as a first-language, E2= English as a second-language
SP= Stimulus Participants

As one might expect, a Pearson's correlation for the eight SPs between the length of the accounts and the mean Total RM scores for their respective accounts, is positive and significant, $r = .44$, $n = 8$; $p < .01$. This may be critical, since Table 10.5 shows that the deceptive accounts deriving from Video 1 were at least twice as long as the truthful accounts. Hence it is perhaps not surprising that they were found to be richer in RM criteria than truthful accounts, particularly before standardisation. On the other hand, the truthful accounts deriving from Video 2 were twice as long as the deceptive accounts, and were correspondingly also richer in RM criteria. So when truthful accounts were longer than deceptive accounts (i.e. Video 2 stimuli), RM total scores distinguished between truthful and deceptive accounts in line with RM theory, whereas when deceptive accounts were longer than truthful accounts (i.e. Video 1 stimuli), deceptive accounts were accordingly richer in RM criteria.

Of course, correlation is not the same as causation, but these findings are consistent with the interpretation that the RM results for Video 1 in particular were an artefact of individual differences in the lengths of accounts given by participants, irrespective of whether they were telling the truth or not. So basically, these findings point again to the impracticalities associated with using unstandardized RM frequency scores with single cases, or with a small number of statements, in a between-subjects design.

The results for language proficiency were not very illuminating; it seems that language proficiency appeared to have some effect, but showed no meaningful pattern that distinguished results before and after standardisation. That is, it is difficult to come up with an interpretation for or explanation of the findings that would enable one to make informative conclusions. As it is, therefore, perhaps all that can really be concluded is that this re-analysis was 'worth a try'; but given the questions raised previously about the methodology of Study 1, perhaps not too much attention should be paid to the outcome.

Chapter 11

Study 6: The effects of standardisation and the presence of others on Reality Monitoring based lie-detection

11.1 Introduction

Thus far, a number of possible moderator variables have been investigated in an attempt to identify some of the optimal conditions for the use of the RM approach in detecting lies in verbal accounts. This next study considers another potential moderator that has been almost completely neglected in previous research in the area; this concerns the number of people present when someone is providing an account; i.e. whether the numbers of others present alter the lexical profile of accounts such that RM scores are affected.

Relevant research on this subject is rather limited and has tended to focus on changes in the lexical density of accounts. For example, some preliminary findings have indicated that the presence of a second interviewer does not affect the lexical density of the accounts in words when rapport is present. Nevertheless, other results suggest that triadic interactions in interviews can lead to discourse that contains a higher number of words; thus both interviewers and interviewees may independently produce more words in triads than when in groups of two (Driskell et al, 2013). Given the previous findings that longer accounts tend to produce higher RM scores, one might, therefore, expect that triadic relationships would increase the frequency of RM scores. However, these studies involved interactive discussions and did not involve manipulating deception. This contrasts with standard RM procedures which

typically require participants to give uninterrupted free-recall accounts of past events which may be fabricated. The direct relevance of this work to RM studies of deception is, therefore, somewhat questionable. However, perhaps equally or more relevant here is Zajonc's seminal work on social inhibition and facilitation, which has since been replicated and extended by other researchers.

Zajonc (1965, 1980) proposed, and found evidence to support the view that, compared to when working alone, the presence of others inhibits performance on some tasks and facilitates performance on others. Such tasks have been categorised in a number of ways but, typically, social inhibition has been shown to occur most frequently when tasks are complex, involve novel stimuli, require the suppression of dominant responses and require the detection and correction of errors, whereas social facilitation occurs when the opposite conditions exist (see Wagstaff et al., 2008).

Zajonc's main explanation for these effects was in terms of the selective enhancement and suppression of dominant responses, however, Wagstaff et al. (2008) have argued that the effects are perhaps better accommodated in terms of more modern conceptions of Working Memory. Within the latter framework, tasks subject to social inhibition can be defined as executive tasks in terms of the literature on Working Memory, and social inhibition can then be explained by the effects of the presence of others on executive functioning. For instance, any threat to well-being may necessitate careful executive monitoring of the situation in order to select appropriate strategies for 'fight or flight'. From a sociobiological perspective, human groups, especially strangers, may be perceived as a potential source of threat; consequently, if the executive system is activated by being in a group, one might predict that an executive task would be performed less well in a group situation;

precisely in the same way as the administration of two executive tasks concurrently inhibits performance on one or both tasks. However, this might also 'free up' other 'fight or flight' systems to respond automatically to environmental threat without intervention from a supervisory system. In other words, being in the presence of others, of itself, may increase cognitive load, which inhibits performance on executive tasks, but facilitates performance on non-executive tasks. And, indeed, Wagstaff et al. (2008) found evidence for this in that the presence of others decreased performance on an executive phonemic fluency task, and enhanced performance on a non-executive, more automatic confidence-accuracy task. Given this, one might predict that being in a room with others (at least in a verbally non-interactive context) would be associated with higher cognitive load, resulting in slower speech rate and shorter accounts with consequent lower overall RM scores, regardless of whether accounts are truthful or not. At the same time, however, one might expect that the presence of others, if it increases cognitive load, might allow a better discrimination between truthful from deceptive accounts using RM criteria. This would follow from recent research which suggests that, as lying is a cognitively demanding task, inducing cognitive load may improve the lie-detection ability using RM as well as other criteria (Vrij, 2008a; Vrij et al., 2008; Walczyk et al., 2012). The idea here is that producing a false account full of the kinds of details tapped by the RM criteria is a difficult task at the best of times; it is made even more difficult, however, if one has to do this in the presence of others. It was, therefore, hypothesized that RM criteria would be more diagnostic of deception when participants are in a room with others present. One might also predict that, presumably, these effects will be greater the more people are present. For example, in terms of an inhibiting influence on executive performance, Wagstaff et al. (2008)

found that the addition of one extra experimenter-observer to a group of two or three individuals was sufficient to produce a significant effect.

In addition to testing how the presence of others may affect RM scores, this study again investigated the effects of standardisation on the efficacy of the RM approach in discriminating between truthful and deceptive accounts using the RM approach. However, as well as standardisation per 100 words, another form of standardisation was included, standardising scores per duration of account. Although this approach has been used in past research (see, for example, Gnisci et al., 2010), as yet, no study has compared results before and after standardisation for duration, or compared the effects of this form of standardisation with standardisation for word-count and no standardisation (of raw frequencies). Notwithstanding the rather anomalous results of the last study, in line with previous predictions, it was again hypothesized that the RM approach will be more effective in discriminating truths from lies before standardisation for length or duration.

Finally, in this final study, the opportunity was taken to look again at the diagnostic validity of length, duration and speech rate, as cues to deception. In line with previous findings from Chapter 7, it was hypothesized that compared to those fabricating accounts, truth-tellers will produce accounts longer in terms of number of words and duration, and display a faster speech rate.

11.2 Method

11.2.1. *Participants*

An opportunity sample consisting of 31 University of Liverpool and University of Huddersfield students (mean age = 24.12; *SD* = 9.05) was employed as SPs; there were 12 males and 19 females. There were no exclusion criteria other than that

participants had to be older than 17 years. All Liverpool participants volunteered in response to an advertisement posted in the University of Liverpool website. University of Huddersfield students were approached via an e-mail invitation sent through the Director of a Psychology course known to the researcher. No reimbursement was offered.

There were two response participants; these were the same as those employed in Study 4 (see Section 9.2.1). Both (one male and one female) were postgraduate research students of Forensic Psychology with expertise in lie-detection.

11.2.2. *Materials and Procedure*

The study was conducted in two phases. The two phases were similar to those outlined in Chapters 7, 8 and 9. Thus in Phase 1, the stimulus participants were shown the same list of significant autobiographical events deriving from the adaptation of LEI described earlier (see section 7.2.2). They were asked to consider if they had previously experienced any of the listed events and to then describe two events: one that they had previously experienced (truthful condition) and one that they had never experienced (deceptive condition). All SPs were subjected to both deceptive and truthful conditions in a counterbalanced order.

In this study, however, participants were asked to report their accounts orally in three different conditions: when alone (n = 10), with one interviewer in the room (n = 10); and with two interviewers in the room (n = 11). After the free recollection of the event was initiated, no further interaction took place between interviewer(s) and interviewee. In the third condition where SPs were joined by two persons in the room, no interaction at all took place between the second

interviewers and the interviewee; however the second interviewer was introduced as a note-keeper at the beginning of the session. All accounts were (audio) recorded and timed, and subsequently transcribed.

In Phase 2, the two RPs were asked to code the frequencies of responses in the transcribed accounts according to the usual five RM criteria (i.e., perceptual information, spatial information, affective information, temporal, information and information regarding cognitive operations; see Appendix 6b for the RM definitions used and Appendix 5 for instructions and scoring sheets). The RPs were blind as to the experimental conditions, and to whether the accounts were truthful or deceptive but they were aware that this was a study in lie-detection.

11.2.3 *Design*

A 3 x 2 mixed design was used (presence of others: no one/one or two people present in the room x truthfulness: truthful/deceptive event); mixed ANOVAs with repeated measures on the second factor, were conducted. The five RM criteria frequency scores (unstandardised, standardised per 100 words, and standardised for duration), the Total RM scores, word-count, duration and speech rate were the dependent variables. Preliminary analyses found no effects of the order in which the accounts were presented.

11.3 Results

11.3.1. *Inter-rater reliability for the raw frequency data*

As in Study 4, the principal coder scored all accounts whereas the secondary coder randomly scored 25% of the accounts. Intra-class correlation agreement and Pearson's correlations showed that, in terms of applying the RM criteria to the

various measures, there was high and significant inter-coder agreement between the two judges (ICC = .82 - .91; r = .79 - .90). Again, considering the very good inter-coder agreement for the random sample, frequency counts produced by the principal coder were used in all analyses.

11.3.2. *Preliminary analyses of word-count, duration and speech rate*

Analyses using 3 x 2 (presence of others: no one/one or two people present in the room with truthfulness: truthful/deceptive event) mixed ANOVAs with repeated measures on the second factor, showed that the truthful accounts contained significantly more words (M = 376.45, SD = 236.47) than the deceptive accounts (M = 281.55, SD = 223.88); F (1,28) = 11.826, p = .002; η^2_p = .30. Similarly, the truthful accounts were longer in duration, i.e. time spent in seconds to produce them (M = 137.16, SD = 74.79), than the deceptive accounts (M = 106.84, SD = 66.47); F (1,28) = 13.67, p = .001; η^2_p = .33. Truthful accounts were also more fluent, producing significantly more words per second (M = 2.71, SD = 0.49) than deceptive accounts (M = 2.58, SD = 0.56); F (1,28) = 4.45, p = .044; η^2_p = .14.

Significant main effects were also found for presence of others. In particular, account length F (1,28) = 7.37, p = .003; η^2_p = .34; account duration F (1,28) = 5.25, p = .012; η^2_p = .27 and fluency F (1,28) = 4.15, p = .026; η^2_p = .22. There were no significant interaction effects.

Further Bonferroni-corrected post-hoc t-tests showed that accounts generated in a room with one person present were significantly longer in terms of the number of words they contained (M = 510.00, SD = 287.76) than accounts generated in a room with no persons present (M = 218.85, SD = 298.38), p =.004, or two persons present (M = 264.59, SD = 96.30), p =.014). No other between-group differences

were found. Similarly, accounts generated in a room with one person present were significantly longer in duration measured in seconds (M = 171.20, SD = 87.94) than accounts generated in a room with no persons present (M = 91.10, SD = 40.64), p =.014, but only marginally longer in duration than accounts produced when two persons were present (M = 105.36, SD =49.78), p =.05. No other group differences were found. Finally, speech rate amongst story tellers who were in a room with one person was marginally but not significantly faster (M = 3.00, SD =0.42) than the speech rate of story tellers who were in a room with no person present (M = 2.47, SD = 0.50), p =.053) or two persons present (M = 2.49, SD = 0.53), p =.058). No other group differences were found.

To summarise, overall the accounts produced in a room with one person contained more words, tended to be longer in duration and had a higher speech rate than the accounts produced in the other two conditions; however, the largest differences were found between the conditions where participants were alone and where one person was present. Contrary to predictions, there were no clear differences between accounts from participants who were in a room with two persons and those who were alone. Moreover, the lack of significant interactions indicated that the presence of others did not affect the ability to distinguish between truthful and deceptive accounts using these criteria.

11.3.3. *RM frequency results before word-count standardisation*

The RM frequencies before word-count standardisation were then analysed using a series of six 3 x 2 (presence of others: no one/one/two persons present with truthfulness: truthful/deceptive) mixed ANOVAs with repeated measures on the second factor; one for each of the five RM criteria, and one for the Total RM score.

As before, Total RM scores were calculated by adding scores for perceptual, spatial, affective, and temporal information and deducting scores for cognitive operations. The means and *SD* are shown in Table 11.1.

As predicted, before standardisation, mean scores were higher for the truthful accounts for all RM criteria with the exception of cognitive operations. Of those, significant effects were found for Total RM scores $F(1,28) = 28.34$, $p = .001$, perceptual information $F(1,28) = 5.24$; $p = .030$; spatial information $F(1,28) = 17.46$, $p = .001$ and temporal information $F(1,28) = 12.73$, $p = .001$ (see Table 11.1). The effect for affective information also approached significance, $F(1,28) = 4.06$, $p = .053$.

A significant main effect for Presence of others was found for the Total RM $F(1,28) = 4.23$, $p = .025$ but also for the criteria of affective information $F(1,28) = 4.26$, $p = .024$, cognitive information, $F(1,28) = 3.36$, $p = .049$, spatial information, $F(1,28) = 4.48$, $p = .020$ and temporal information $F(1,28) = 6.35$, $p = .005$ (see Table 11.2). Bonferroni-corrected post-hoc t-tests found that accounts generated in a room with one person present were richer in Total RM than accounts generated in a room with no persons present, $p = .034$). Similarly, accounts generated in a room with one person present were richer in affective information than accounts generated in a room with no persons present, $p = .022$. Also, accounts generated in a room with one person present were richer in spatial information than accounts generated in a room with two persons present, $p = .038$. And accounts generated in a room with one person present were richer in temporal information than accounts generated in a room with two, $p = .006$, or no persons present, $p = .003$.

Table 11.1. RM mean (*SD*) frequency counts as a function of truthfulness before and after standardisation of word count and duration for Deceptive (D) & Truthful (T) accounts

	Unstandardized			Standardised per 100 words			Standardised per minute		
RM Criterion	D	T	η^2_p	D	T	η^2_p	D	T	η^2_p
Perceptual	4.87	7.01	.16*	1.89	2.19	.03	3.00	3.55	.03
	(3.63)	(4.41)		(1.05)	(1.29)		(1.99)	(2.32)	
Spatial	9.98	13.64	.38**	3.69	3.68	.01	5.77	5.98	.01
	(8.13)	(10.02)		(1.58)	(1.64)		(2.85)	(2.59)	
Affective	2.69	3.72	.13	.89	.94	.03	1.44	1.48	.00
	(3.30)	(3.36)		(0.70)	(0.65)		(1.17)	(1.04)	
Cognitive	7.14	5.92	.07	2.40	1.60	.31**	3.70	2.51	.26**
	(7.75)	(5.65)		(1.15)	(1.09)		(1.75)	(1.52)	
Temporal	5.98	9.52	.31**	2.09	2.44	.05	3.11	4.09	.24**
	(6.05)	(8.15)		(1.21)	(0.99)		(1.54)	(1.89)	
Total RM	16.39	27.99	.50**	6.51	7.65	.08	9.62	12.58	.26**
	(12.94)	(18.48)		(3.00)	(2.80)		(5.49)	(4.96)	

*$p < .05$; **$p < .01$

Table 11.2. RM mean (SD) frequency counts as a function of the number of people present before standardization.

	Number of people present			
RM Criterion	0	1	2	η^2_p
Perceptual	5.95	6.52	5.41	.02
	(3.42)	(4.56)	(3.84)	
Spatial	9.00	19.77	8.77	.24*
	(4.56)	(12.75)	(4.89)	
Affective	1.60	5.10	1.95	.23*
	(1.61)	(4.08)	(2.65)	
Cognitive	5.45	10.45	3.95	.19*
	(4.05)	(9.73)	(3.05)	
Temporal	4.80	13.02	5.64	.31**
	(3.51)	(9.33)	(5.60)	
Total	15.90	32.17	18.82	.23*
	(9.57)	(19.79)	(11.05)	

*p<.05;**p < .01

Although, the cognitive RM scores were highest when there was one person present in the room, none of the pair-wise comparisons was significant. No other differences were significant. These results again illustrate a general trend for RM scores to be highest for accounts generated in a room with one person present.

Only one significant interaction between truthfulness and presence of others was found, that for affective information $F(1,28) = 3.94$, $p = .032$; $\eta^2_p = .22$ (see Table 11.3).

Table 11.3. Means (*SDs*) for interaction effect for affective information scores and presence of others in the room

RM	Number of people present	Truthful	Deceptive	η^2_p
Affective	No one else present	1.80 (1.87)	1.40 (1.34)	.06
	One person present	4.95 (3.64)	5.25 (4.53)	.01
	Two persons present	4.36 (3.61)	1.54 (1.69)	.44*

*$p < .05$

Further post-hoc *t* test comparisons between the truthful and deceptive accounts showed a significant effect only for the two person present condition (p <.05), and this was in the direction predicted by RM theory. i.e. the ability to detect lies using the affective information criterion was only significant in the two persons present condition. Otherwise, the lack of significant interactions indicated again that the presence of others did not affect the ability to distinguish between truthful and deceptive accounts using the RM criteria.

11.3.4. *RM frequency results after word-count standardisation*

Standardization per 100 words: To standardise word-count, the RM raw scores were re-calculated per 100 words of account; i.e. raw frequencies per 100 words (for examples of this method, see Larsson & Granhag, 2005; Strömwall & Granhag, 2005; Vrij et. al, 2004). The analysis used in the previous section was then repeated. None of the main effects for truthful and deceptive accounts was significant with the exception of cognitive information, which, as predicted, was higher in deceptive accounts, $F(1,28) = 12.44$, $p = .001$ (see Table 11.1).

Main effects for presence of others were found only for the criteria of perceptual information $F(1,28) = 10.55$, $p = .001$, and cognitive information, $F(1,28) = 4.43$, $p = .021$; no other group differences were found (Table 11.4). Bonferroni-corrected post-hoc *t*-tests found that accounts generated in a room where no person was present were richer in perceptual information than accounts generated in a room with one person present ($p < .001$). Also, accounts produced when no-one was present were also richer in cognitive information than accounts generated in a room with two persons present ($p < .05$). None of the other comparisons or interactions was significant.

Table 11.4. RM mean (SD) frequency counts as a function of the number of people present after word-count standardization

RM Criterion	0	1	2	η^2_p
Perceptual	2.70	1.34	2.07	.43**
	(1.06)	(0.75)	(1.28)	
Spatial	4.21	3.46	3.42	.09
	(1.67)	(1.06)	(1.56)	
Affective	0.71	0.98	1.04	.08
	(.65)	(0.56)	(0.76)	
Cognitive	2.60	1.88	1.55	.24*
	(1.51)	(0.76)	(0.67)	
Temporal	2.28	2.30	2.23	.01
	(1.32)	(0.96)	(1.00)	
Total	7.29	6.19	7.70	.09
	(3.39)	(1.94)	(3.16)	

* p <.05; **p < .01

These results show a fairly clear divergence from those found with the unstandardised scores in terms of both statistical significance and effect sizes; i.e. the ability to discriminate between truthful and deceptive accounts is generally reduced; also, the influence of the one person present condition seems to be replaced by a (lesser) influence of being alone. But again, and perhaps most notably, the presence of others did not appear to affect the ability to distinguish between truthful and deceptive accounts using these criteria.

Standardization per duration of accounts: To standardise for duration, the RM raw scores were re-calculated per minute of account (i.e. RM scores were divided by the total duration of accounts in minutes). In accordance with the standardisation described in Chapter 7, latency period was not taken into account. The same analysis used in the previous sections was then conducted.

Significant main effects for truthfulness were found for Total RM scores, $F(1,28) = 9.70$, $p = .004$; temporal information, $F(1,28) = 9.70$, $p = .006$ and cognitive information, $F(1,28) = 9.81$, $p = .004$, in the directions predicted by RM theory (see Table 11.1). Somewhat remarkably, therefore, although, compared to using unstandardised raw frequencies, standardising for duration appeared to reduce the diagnostic validity of RM (effect sizes were generally lower), nevertheless, RM criteria could still discriminate between truthful and deceptive accounts using this standardised measure.

A main effect for presence of others was found only for cognitive information $F(1,28) = 3.43$, $p = .047$ (Table 11.5); Bonferroni-corrected post-hoc *t*-tests found that none of the pairwise comparisons was significant ($p > .05$); however, the means indicated that cognitive information frequency counts after standardisation for duration were lowest in the two person condition (the other two conditions differed little). None of the interactions was significant; so once again, the presence of others did not affect the ability to distinguish between truthful and deceptive accounts using these criteria.

.

Table 11.5 RM mean (SD) frequency counts as a function of the number of people being present after standardization for account duration

RM Criterion	People Present			
	0	1	2	η^2_p
Perceptual	4.18	2.56	3.10	.17
	(2.02)	(1.08)	(2.49)	
Spatial	6.41	6.12	5.01	.06
	(3.38)	(2.14)	(2.56)	
Affective	1.03	1.65	1.61	.13
	(0.93)	(1.05)	(1.21)	
Cognitive	3.63	3.38	2.36	.18*
	(1.86)	(1.47)	(1.29)	
Temporal	3.15	4.21	3.44	.09
	(1.56)	(1.91)	(1.64)	
Total	11.13	11.33	10.87	.01
	(6.50)	(3.84)	(5.44)	

*$p<.05$

11.4. Discussion

In accordance with previous results in this thesis, truth-tellers' accounts were longer in duration and word-count and exhibited a faster speech rate than those of liars, thus further endorsing the possible use of these measures as cues to deception in verbal accounts. However, the present results also showed that when one interviewer was present, the accounts tended to be longer, both in terms of their duration in time and length, than when no or two interviewers were present. The story tellers' speech rate was also fastest when the account was given in a room with one person than when in

a room with no persons or two persons present. Significantly, however, the presence of others did not affect the ability to detect deception when applying these measures.

Before standardisation the RM results appeared to follow a similar pattern; in accordance with the predictions of RM theory, mean scores were higher for the truthful accounts for all RM criteria with the exception of cognitive operations, and most were significant, or nearly significant in this respect. Also, there was again a general trend for RM scores to be highest for accounts generated with one person present. However, the presence of others did not affect the ability to distinguish between truthful and deceptive accounts using the RM criteria.

The results for the standardised scores showed a general reduction in the ability to discriminate between truthful and deceptive accounts. Also, the influence of the one person present condition seemed to disappear. And once again, the presence of others did not appear to affect the ability to distinguish between truthful and deceptive accounts using these criteria. In some respects, however, the situation was reversed slightly when standardisation for duration was applied. Compared to unstandardised raw frequencies, standardising for duration appeared to reduce the diagnostic validity of RM (effect sizes were generally lower), but it was still somewhat better than when standardisation for word-count was applied.

Arguably the key finding of this study is that that the results of RM assessment through recording frequencies may vary not only according to whether standardisation is applied, but also the type of standardisation used. As such, they suggest that the findings of Study 4 may be reasonably robust; i.e. RM criteria distinguish best between truthful and deceptive accounts when not standardised. Consequently, standardisation might be a key variable in accounting for differences between RM studies when they have occurred. However, the issue still remains as to

why differences appeared to occur between the two methods of standardisation (word-count and duration).

As pointed out in the introduction to Study 4, the basic rationale for avoiding standardisation is that it makes little sense to correct for word-count if length per se is considered to be a reliable cue to deception (Colwell et al, 2002; Memon et al., 2010; Vrij et al., 2004). Moreover, there is no reference in the seminal RM papers by Johnson and colleagues (Johnson et al, 1993; Johnson & Raye, 1981) to the idea that truthful accounts will be richer in the *density* of RM criteria; but more simply that they will overall contain "more perceptual, spatial and temporal, semantic and affective information and less information about cognitive operations" (Johnson et al., 1993, p.4). It could be argued, therefore, that standardising for word-count differences alters one of the core qualities of lies (i.e. they generally contain less detailed information). This, in turn, might help explain the difference between the results for the two standardisation methods. Whereas standardisation for number of words clearly negates the core quality of 'more information' and thus more details relevant to RM criteria in accounts, standardisation according to duration (RM criteria per minute) is less likely to do so. This is because, as demonstrated in the results here, truthful accounts also tend to reflect a faster speed rate; that is, truth-tellers are able to relate the information they have at a faster rate and are thus able to report more details within the time units under consideration. In this way, standardisation for duration only partly controls for word-count, and thus one might expect results after standardisation for duration to approximate more to those for no standardisation than standardisation for word-count.

Significantly, although the number of people in the room also appeared to influence RM assessments, the effects were not consistently linear as might have

been predicted, and they varied according to standardisation. Importantly, there was no overall support for the view that, by increasing cognitive load, the presence of others would improve the ability to discriminate between truthful and deceptive accounts. Rather, regardless of the veracity of the accounts, scores for the various non-RM indicators, and RM indicators before standardisation, tended to be highest when one interviewer was in the room. For the RM indicators, however, this trend appeared to disappear after both forms of standardisation. Without further investigation, it is difficult to come up with a ready explanation for these findings. However, one possibility is that, with these particular tasks and this situation, any adverse effects of increasing cognitive load that might be expected to interfere with executive processing might have been counterbalanced by a social facilitation effect for speech production. Speech by its very nature is fundamentally a social tool; hence people are more used to speaking out loud to another than they are, sat alone, talking to themselves. Hence speaking to another may actually be construed as a more comfortable, automatic and less demanding task than talking to oneself, or speaking alone into a microphone. However, when an extra observer is present, the situation possibly becomes more uncomfortable and demanding again, and social inhibition sets in. If this argument has any validity, then it is possible one might obtain different results using written accounts.

As they stand, however, the present results appear to indicate that the presence of others neither inhibits nor facilitates the ability to distinguish truthful from deceptive accounts using RM; however, when unstandardised RM scores are used it does affect the overall size of the RM criteria scores. If shown to be reliable, this latter finding is potentially very important as it suggests that, like oral and written accounts, accounts generated in the presence of different numbers of people

should never be treated as equivalent either within, or across studies; i.e. this is potentially a highly significant confound.

PART III

Discussion and Conclusions

Chapter 12

General discussion and conclusions

To reiterate, the key objectives of this thesis were to test whether the RM approach has any value overall in distinguishing between truthful and deceptive accounts, and to investigate the circumstances under which it might give optimal results; i.e. to assess what factors moderate its efficacy in this respect. As yet, only a few investigators have identified the assessment of potential moderators as a key priority in RM and lie-detection research (i.e. DePaulo et al., 2003; Masip et al., 2005), hence little work has been conducted on this specific aspect. In this thesis, therefore, a number of potentially important, but under-researched moderators were targeted to achieve a better understanding of the contexts in which RM can be used most successfully. This gave rise to the following subsidiary aims.

1. To compare the relative efficacy of rating scales and raw frequency RM scoring systems.
2. To examine possible second-language effects on the efficacy of the RM approach.
3. To test the relative efficacy of using spoken and written statements with the RM approach.
4. To test the effects of standardisation for word-count and duration of accounts on the efficacy of the RM approach..

5. To test more generally the usefulness of word-count, duration and speech rate as cues to deception (i.e. as influences on and possible adjuncts to or elaborations of, RM).
6. To assess whether demand characteristics (blind coding) may influence the coding and efficacy of RM criteria.
7. To test the influence on the efficacy of the RM approach of the number of people present in the room when accounts are given.

The purpose of the final chapter of this thesis is to summarize and discuss the present findings in the light of these aims. Possible implications of the findings are also discussed for RM procedures and verbal lie-detection generally, but with particular reference to comparisons with CBCA, which bears a number of conceptual and methodological similarities with RM. The chapter finishes with a consideration of limitations of the studies described in this thesis, and a consideration of some possible implications of the findings for future research and practice using the RM approach.

12.1. Summary of findings for RM criteria

12.1.1. *Findings for Total RM scores*

Overall, in most of studies presented there was evidence that Total RM scores, as determined by the procedures used here (both ratings and raw frequencies), successfully discriminated between truthful and deceptive accounts. The exceptions were for the rather anomalous Video 1, in Study 1, and Study 3, involving experimentally blind raters. However, perhaps the most salient finding in this thesis, was the association between RM scores and the word-count and duration of the

accounts. Specifically, RM was a more effective diagnostic tool before the accounts were standardised for length; indeed, Total RM scores failed to distinguish between truthful and deceptive accounts after standardisation. Moreover, the diagnostic strength of Total RM scores (in terms of the respective effect sizes) was reduced after standardisation for duration, although RM scores still discriminated significantly between truthful and deceptive accounts. As previously noted, this should perhaps not be considered surprising given that it is widely accepted that longer statements tend to sharpen the differences between liars and truth-tellers, as, by definition, words are the carriers of verbal cues to deception (Leal, Vrij, Warmelink, Vernham & Fisher, 2015; Vrij et al, 2007). What is perhaps surprising, however, is that investigators have not consistently recognised the significance of this when applying RM criteria.

It can be noted also that standardisation for word-count and duration had a direct impact on the between-subjects main effects of the moderators tested. For example, modality and presence of others were two key variables whose impact on Total RM scores varied depending on whether or not the accounts were standardised. In particular, written accounts were richer in terms of RM raw frequencies than spoken accounts after standardisation, but not before. These differences can be attributed to the respective *lexical-density* differences between the two types of account. So as written language is richer in content words per clause (or per any other lexical unit) than spoken language (Halliday, 1989; 2001; Tannen & Chafe, 1982; 1986), it is also richer in RM criteria when the scores are standardised for account length. Also, the main effects of presence of others on RM scores before standardisation were diminished after standardising for word-count and length. That is, presence of others affected total RM scores before standardisation but less so

after standardisation for both word-count and duration. In particular, unstandardised accounts produced in a room with one person were richer in terms of total RM scores. By way of explanation, it was suggested in Chapter 11 that because story tellers displayed a faster speech rate when one person was present in the room, possibly because speaking out load is more natural in the presence of another, this allowed them to generate more real and fabricated information, thus increasing total RM scores. However, standardising for word-count and duration minimized the differences between conditions in terms of this effect.

12.1.2 Individual RM criteria

This section overviews the findings for each of the RM criteria assessed.

12.1.2.1 *Spatial and temporal information*

In line with past findings (Masip et al., 2005; Vri, 2008a; 2015), frequency measures of spatial and temporal information were found to be two of the most consistently effective diagnostic RM criteria in this thesis (see for example, Studies 4 and 6). However, their effectiveness was greater with unstandardized raw frequencies; indeed, when account-length was taken into account through standardisation, neither criterion could reliably distinguish between truthful and deceptive accounts; however, temporal information could still discriminate truthful from deceptive accounts when standardisation for the duration of the account was applied.

Nevertheless, although both spatial and temporal information criteria were reasonably strong diagnostically, spatial information appeared to be the most consistent in this respect; it also showed the larger effect sizes of the two (see section 12.4 and Table 12.2 later in this Chapter for a comparative overview).

Temporal information scores overall seemed more affected by the moderators under investigation, in much the same way as total RM scores. Thus, irrespective of their truth status or veracity, compared to written accounts, spoken accounts were richer in terms of both Total RM scores and temporal information (determined by both ratings and raw frequencies) before standardisation. In contrast, written accounts were denser in terms of both Total RM scores and temporal information scores after standardisation. Accounts produced in a room with one person were also richer in Total RM and temporal information than accounts produced in a room with two or no persons in the room. These results suggest that these moderating influences on Total RM scores may have come about particularly as a result of their impact on temporal information.

It is perhaps worth noting here also that the CBCA criterion of contextual information, which comprises temporal and spatial information, is reportedly one of the most useful CBCA criteria, displaying medium to high effect sizes (Amado, Arce & Farina, 2015; Vrij, 2008a; 2015); indeed, both spatial and temporal information have been used with some success as lie-detection criteria by themselves (Warmelink, Vrij, Mann & Granhag, 2013). This may be significant given that contextual information of this kind could potentially be used not only to detect lies about past actions and events, but also about intentions (Warmelink et al., 2013); i.e. it could be used in security settings in assessing intentions for future actions (e.g. airport checks, immigration offices, etc).

This raises the issue of whether temporal and spatial information could be subsumed under a single heading such as *contextual embedding* as in the CBCA. In fact, the two criteria were operationalised as one in the original RM studies (Johnson et al, 1993; Johnson & Suengas, 1989) and have been elsewhere (see, for example,

Masip et al., 2005; Colwell et al, 2007). However, whilst there may be advantages in terms of semantic and methodological parsimony in doing this, the present results suggest that the two may be both conceptually and empirically distinct; for example, spatial information appears to be more effective diagnostically, and temporal information more susceptible to certain moderating effects.

12.1.2.2. *Vividness*

The vividness rating criterion significantly distinguished between truthful and deceptive accounts, on two of the four occasions in which it was measured (i.e. in Video 2 from Study 1 and, more reliably, in Study 2. These findings may be important given that vividness has often received little attention in previous research and has even occasionally been avoided or dismissed as a RM criterion (Alonso-Quecuty, 1992; 1995; Manzanero & Diges, 1996; Stromwall & Granhag, 2005). Also, when it has been applied, results have been mixed (see, for example, Granhag et al., 2006; Masip et al., 2005; Porter et al., 1999; Sporer & Sharman, 2006; Vrij, 2008a).

A number of factors may impact on the use of the vividness criterion to produce inconsistent findings. For example, in Study 2, vividness was diagnostic of deception in both spoken and written accounts, however, it was diagnostically stronger in spoken accounts. This is consistent with the idea that written statements can allow liars to give their accounts without the extra anxiety that may accompany interpersonal interaction, and hence may inhibit the distinction between truths and lies (Colwell et al., 2007); hence modality may be an influential factor.

Another factor that could potentially affect the diagnostic capacity of vividness concerns the realism of the stimulus materials. As mentioned earlier, a

number of researchers have suggested that the realistic materials, and particularly those associated with high stakes lies, are likely to enhance the distinction between truthful and deceptive accounts (Ball & O'Callaghan, 2008; Barnier et al., 2005; Johnson et al, 1988; Masip et al., 2005; Sporer & Sharman, 2006), and this may be particularly important when applying the vividness criterion (Colwell et al., 2007). However, increasing the motivation to lie effectively in laboratory settings, and doing this ethically, is quite a methodological challenge. For example, one possible way of doing this might be to introduce instrumental incentives such as financial rewards or course credits to SPs for producing deceptive accounts that are difficult to detect. However, DePaulo et al. (2003) found that such incentives have almost no effect at the diagnostic value of lie-detection cues. Moreover, as vividness was diagnostic in the second Video in Study 1, as well as in the second study when the materials were more realistic autobiographical accounts about important personal events, it is not obvious that realism was particularly influential here. Nevertheless, in future, it would make sense to make materials as realistic as possible, and this might include increasing incentives for escaping detection.

On first consideration, it could also be argued that the delay between the occurrences of the events to be described in the account and when the account is given could also be particularly influential with the vividness criterion (Masip et al., 2005; Vrij, 2008a). For instance, one might think that truthful accounts would be most vivid if given immediately after the event, thus allowing a greater differentiation between truthful and deceptive accounts. However, there is some evidence that accounts may seem more vivid when recalled a week after they happened than when recalled immediately (Masip et al., 2005). Moreover, Manzanero and Diges (1996) have argued that delay and retrospective preparation

may also allow imagined accounts to become more vivid over-time. The present studies did not investigate this variable systematically, however, there were differences in delay between Studies 1 and 2, but as mentioned previously, vividness was diagnostic to some extent in both. The most that can be said at present, therefore, is that the effects of delay on the diagnostic value of vividness have yet to be established.

Of course, these factors could apply to a number of other RM criteria besides vividness. For example, if an account is vivid it would also presumably be more likely to be rich in contextual and perceptual information; indeed, Colwell et al. (2007) have conceptualised vividness as a criterion which is found in accounts that are rich in detail involving sensory experience and time, place, people and objects. In a similar way, Vrij (2015) has argued that vividness can be construed as sharing similarities with CBCA's most successful criterion, *quantity of details*; i.e. richness of details regarding time, place, people and objects. Nevertheless, in the absence of empirical evidence from a thorough cluster or factor analysis, the present results suggest as the results for vividness do not exactly mirror those for the other criteria, there may be merit in considering them to be conceptually distinct.

12.1.2.3. *Perceptual information*

The results regarding perceptual information were again mixed. The results from the ratings of Video 2 in Study 1 and also frequency scores in Studies 5 and 6 indicated that truthful accounts were, as predicted by RM theory, richer in perceptual information than deceptive accounts, but only before standardisation for length. However, in Studies 2, 3 and 4 truthful accounts did not contain significantly more (or less) perceptual information than deceptive accounts, either in terms of ratings or

frequencies. It can be noted that previous researchers have reported similarly mixed findings (Alonso-Quecuty, 1992; Granhag et al., 2001; Masip et al., 2005; Manzanero & Digest, 1996).

As with the other criteria, one issue that may be influential here is how the criterion is defined. As it is currently defined, both here and elsewhere, perceptual information incorporates multiple subcategories (e.g. sounds, visual information, odours, etc.), and researchers have yet to agree on a uniform or standard definition (Colwell et al, 2007). Hence perceptual information has variously been conceptualised as visual and audio information (Granhag et al, 2001; Vrij et al., 2000; Vrij, et al, 2001), visual, audio and smell(s) (Granhag et al., 2006) and about objects and people (Roberts & Lamb, 2010). Indeed, often the coding strategy is not reported at all in published research (e.g. Manzanero & Diges, 1995) and only rarely are specific examples provided (Vrij, 2015). If results are to be more consistent, therefore, particularly across studies, this criterion, like all of the RM criteria, needs to be supported by clearer instructions regarding coding and measurement.

12.1.2.4. *Affective information*

Affective information did not significantly discriminate accurately (i.e. in the predicted direction) between truthful and deceptive accounts until it was tested using unstandardised raw frequencies in accounts produced in a room with two extra people present (Study 6). Indeed, in some of the other studies there was a trend for affective information to be richer in deceptive accounts; and significantly so for both ratings and frequencies in Video 1 in Study 1.

Other results indicated that, irrespective of veracity, accounts produced in a room with one person present were richer in affective information than accounts

produced in a room with two persons on no persons. There was also some evidence from Studies 1 and 5 that second-language accounts were richer in affective information both before and after standardisation. This rather odd set of results in some ways typifies the conjecture that exists concerning the nature and role of affective information in RM. For instance, it has been argued by some that the application of the affective information criterion may suffer from the same problems as measures such the polygraph, in that they can be confounded by general indicators of other affective reactions (stress, surprise, etc.) that may exist in a situation, or in response to a question, independently of whether the account giver is lying, and in the case of RM, independent of what might have, or not have, been experienced at the time (Masip et al., 2005; Santtila et al., 1998). Possibly in line with this, the present results suggest that affective information might be as much influenced by the number of people present in a room when the account was given as by whether the person has actually lied. If the presence of one other person, in particular, provokes an emotional reaction to being evaluated, this might overshadow any differences due to lying or telling the truth. This might also account for why second-language accounts were richer in affective information, given that previous research has shown that second-language speakers display higher psychophysiological responses than first-language speakers. For example, Caldwell-Harris and Ayçiçeği-Dinn (2009) found that that skin conductance responses, a key indicator of emotional states, were higher when both false and true statements were expressed in the speaker's second-language (again possibly because of evaluation apprehension). In fact, one could argue that contrary to the predictions of RM, if affective information is a feature of increased anxiety or arousal, there is a case for predicting that it might actually be greater when the person is lying (as in Video 1 in

Study 1); the rationale being that lying is emotionally arousing and this triggers or primes the inclusion of affective material in fabricated accounts. And, indeed, in line with most of the present findings, there are a number of reports that, in opposition to the original RM theory, invented stories can show a significant tendency to contain more affective information than the genuinely experienced stories (Gnisci et al, 2010; Logue, Book, Frosina, Huizinga & Amos, 2015; Masip et al., 2005). For example, using children, Santtila et al. (1998) found that deceptive accounts of imagined negative past experiences were richer in affective information than actual past experiences. However, this is not consistently the case. In fact, what might be happening here is that, in a lie-detection situation, affective information is influenced by two opposing factors; i.e. an increased situational emotional response to lying which increases cognitive load and decreases overall memory for events when one is lying, but also, because of the affective priming that the situation engenders, an increased tendency to incorporate memory for emotional features into accounts. If the former is emphasised, it is more possible to come up with an explanation for why affective information increased generally with one person present in the room in Study 6. Thus, as suggested previously, it could be the case that because the accounts were spoken, the account-givers felt more comfortable or more natural, allowing them to generate more real and fabricated information, as indicated also in an increase in Total RM scores. This trend might then have been overturned by evaluation apprehension with two people present, resulting in a sharper differentiation between truthful and deceptive accounts. It should be emphasised, of course, that there is not a simple correspondence between feeling an emotion and verbally expressing an emotion; for example, a variety of situational constraints may limit the extent to which the latter can be said to represent the

former (see Fussell, 2014). The possibility also exists, therefore, that there might also have been an interplay between factors influencing the priming and generation of affective material internally, and situational (including possible social) factors influencing its expression. Whatever the case, however, all this would potentially make affective information a rather weak and unreliable diagnostic criterion.

12.1.2.5. *Cognitive Information*

It will be remembered that, according to RM theory, cognitive information is the only criterion that should be higher in deceptive accounts. In the present studies, there was a preponderance of non-significant trends in the predicted direction. Significant trends, however, were found for ratings of Video 2 in Study 1 and frequency scores after standardisation for word-count and duration in Study 6. In fact, in Study 6, this was the only RM criterion that discriminated between truthful and deceptive accounts after standardising for both duration and word-count, though it did not discriminate before standardisation; in this respect, it contradicts Granhag et al.'s (2001) finding that cognitive information was diagnostically stronger before standardizing for word-count.

With regard to moderators, in terms of frequencies, spoken accounts had higher cognitive information frequency scores both before and after standardisation, though when subjective ratings were applied, spoken accounts received lower cognitive information subjective ratings than written statements. Also, in terms of frequencies, accounts (spoken) produced in a room with two persons in the room contained less cognitive information than accounts produced in a room with one person present. These results seem problematic given that if it is assumed that speaking in the presence of others, increases cognitive demand (because of

evaluation apprehension), one might expect more cognitive information in all of these situations. However, it could be the case that, like affective information, the cognitive information criterion may be influenced by two opposing factors in situations of cognitive demand; i.e. first, an increase in cognitive information because the cognitively demanding situation primes cognitive processes at the expense of, or as a substitute for other forms of information; and second, a decrease in memory for, or the ability to produce examples of cognitive operations that might actually have occurred in a real situation. In fact, there might even be a curvilinear effect on the lines of the classic Yerkes-Dodson principle (Yerkes & Dodson, 1908); i.e. there is an increase in the generation of cognitive material from low to medium levels of cognitive demand and arousal, but at very high levels of arousal and cognitive demand the ability to produce any information, including examples of cognitive operations, is impaired. Whatever the case, the results arising from the operation of this criterion are potentially difficult to interpret. Indeed, concerns regarding the value of the cognitive information criterion have been expressed by a number of researchers (see, for example, Masip et al., 2005; Sporer & Sharman, 2006; Vrij et al, 2004). Given the present findings, one might suggest that, like affective information, this criterion is rather weak.

12.1.2.6. *Realism and reconstructability*

Realism and reconstructability, however, were perhaps diagnostically the weakest criteria in that they did not significantly discriminate between truthful and deceptive accounts in any of the studies. There were some non-significant trends in the predicted direction for both criteria in Studies 1 (for both videos) and 3, but, on the other hand, there were trends in the opposite direction in Study 2 (i.e. deceptive

accounts received marginally higher scores than truthful accounts). Reconstructability also displayed the lowest intercoder rating agreement scores, suggesting that training individuals to uniformly conceptualise it to rate statements may be challenging.

Once again these findings seem to fit with trends in the literature. Hence, whilst there is some limited evidence in the literature that realism in particular may have some diagnostic value (Masip et al., 2006; Sporer & Sharman, 2006), others have found no support for these criteria in this respect (see see Santtila et al., 1998). These criteria are also again particularly problematic in terms of the vagueness and subjectivity of their definitions which can lead to low intercoder agreement (Logue et al., 2015; Masip et al., 2005). As with most RM criteria, such difficulties have been exacerbated by the fact that, when researchers have applied the criteria, very limited information is given about the definitions used, or the way coders are informed or trained in their use (see for example, Manzanero & Diges, 1995). In fact, there is even potentially an issue as whether they should even be included at all amongst the standard range of RM criteria. Significantly, neither realism nor reconstructability were included in the list of criteria offered in the original seminal paper on RM by Johnson and Raye (1981), and numerous RM studies have not included them (see, for example, Barnier et al, 2005; Logue et al., 2015; Vrij & Nahari, 2014; Vrij et al., 2000; Vrij et al., 2004; Vrij, 2008a).

12.2. Summary of effects of moderators

In the previous sections, mention was made of some of the effects of the various selected moderators on the RM scores; in this section, for additional clarity, the moderators are considered as topics in their own right.

12.2.1. *RM scores and modality; i.e. spoken vs written accounts*

Importantly, overall, there was no evidence that modality, in terms spoken vs written accounts, affected the capacity to detect lies using RM. Nevertheless, as predicted, spoken accounts received higher raw frequency scores than written accounts before standardisation, and lower RM scores after standardisation. Similarly, as predicted, RM ratings for spoken accounts were also higher than the RM ratings for written accounts. The rationale for this (see Chapter 4, and section 12.1.1 in this chapter) is that as spoken accounts are faster in terms of word production than those produced by writers, the former tend to be longer and thereby contain more RM information before standardisation. On the other hand, as written accounts have higher lexical density than spoken accounts, they tend to be richer in RM information after standardisation for length (Halliday, 1989; 2001; Tannen & Chafe, 1982; 1986). However, contrary to predictions, written accounts contained higher ratings for cognitive information than spoken accounts. Possible explanations for the latter finding have also been given in the previous section.

12.2.2. *RM scores and language proficiency*

Language proficiency received only very limited consideration in this thesis, and was abandoned after the first study, but in as far as the findings went, they indicated no consistent pattern. Thus in Study 1, for Video 1, deceptive first–language accounts were more likely to be judged as truthful and truthful first-language accounts as deceptive. As mentioned previously, this was particularly unexpected since past research has shown that the ability to detect lies is generally poorer when second-language accounts are investigated and stronger when first-language

accounts are assessed (Broadhurst & Cheng, 2005; DaSilva & Leach, 2013; Leach & DaSilva, 2013). However, these results were not replicated with Video 2; moreover, with Video 2, second-language statements received significantly higher raw scores for affective information than first-language statements, and first-language statements received significantly higher raw scores for perceptual information than second-language statements, regardless of whether the statements were true or not. These effects were also moderated by standardisation (Study 5), but not in any meaningful pattern.

Given the methodological inadequacies of Study 1, it would seem presumptuous to make much of these findings, other than to conclude that language proficiency may potentially be a confounding factor in RM research and practice. This may be particularly significant considering the lack of reference to stimuli-participants' language proficiency in experimental research in lie-detection, which often uses students (who may be international and vary in terms of language proficiency).

12.2.3. *Rating scales vs frequency counts*

Comparisons between raw frequency and rating scale measures of RM produced mixed results which, in the case of frequency scores, interacted with standardisation for word-count. Thus raw frequency measures were generally weaker lie-detection tools than rating scales, after the former were standardised for word-count, but they showed higher diagnostic strength (as evidenced by effect sizes) before standardisation. Considering that the ratings in particular were applied by relatively inexperienced and untrained judges it cannot be assumed that more highly trained individuals would have produced the same results; nevertheless, the fact that effects

for simple RM ratings by inexperienced judges were found at all, suggests that ratings may potentially provide a quick and ready diagnostic indication of what might be worth following up. They at least appear to surpass inaccurate global estimates of whether or not an account is truthful or deceptive (see Study 2, in particular).

12.2.4. *Effects of demand characteristics/blind coding on the application of RM criteria*

The results of Study 3 gave no support for the idea that being blind to the purpose of the study facilitates the accurate application of the RM criteria ratings; on the contrary, significant effects were only found when judges were not blind to the procedures (Study 2); i.e. there was no evidence that the blind RPs actually performed better than the non-blind RPs. Consequently, as suggested in Chapter 8, given that global subjective estimates of veracity were ineffective in discriminating between truthful and deceptive accounts, and assuming the results are valid, either the ratings of non-blind judges were not influenced by explicit or implicit global judgments about the truthfulness of the stimulus information, or, if they were, this was perhaps offset by the influence of their relative expertise in the area, bearing in mind that blind participants had no psychology background. Whatever the case, there seems to be no advantage to keeping judges blind to the nature and purpose of the procedures (which obviously makes RM ratings easier to use in practice).

12.2.5. *RM scores and presence of others*

The key findings regarding the effects of the presence of others on RM scores (Study 6) can be summarised as follows. It was generally predicted that RM objective criteria would discriminate better between truthful and deceptive accounts with others present, as this would increase cognitive load. However, in general, with the exception of affective information discussed in the previous section (which possibly because of evaluation apprehension may have a special status in this respect), the presence of others had little effect on the ability to discriminate truthful from deceptive accounts. Nevertheless, overall, scores for the various non-RM indicators, and RM indicators before standardisation, tended to be highest when one interviewer was in the room. For the RM indicators, however, this trend appeared to disappear after both forms of standardisation. In Chapter 11 it was suggested that this might have occurred because, with these particular tasks and this situation, any adverse effects of increasing cognitive load that might be expected to interfere with executive processing might have been counterbalanced by a social facilitation effect for speech production; i.e. speaking to another may actually be construed as a more comfortable and less demanding task than talking to one's-self, or speaking alone into a microphone. However, when an extra observer is present, social inhibition might set in. It would be interesting to see, therefore, whether the same results would occur with written accounts.

In Chapter 11, it was also noted that, if shown to be reliable, this latter finding suggests that, like oral and written accounts, accounts generated in the presence of different numbers of people should never be treated as equivalent either within, or across studies; i.e. this is potentially an important confound in RM research.

12.3. Other potential and related cues to deception

12.3.1. *Global subjective veracity assessments*

As emphasised in the preceding sections, subjective global truthfulness judgments (i.e. simple overall assessments of whether an account is truthful or not) as used in Studies 1 and 2 did not significantly discriminate between truthful and deceptive accounts; the only exception was for Video 1 in Study 1, where first-language deceptive accounts actually received higher truthfulness ratings than truthful statements. This supports previous research showing that subjective global judgments of veracity are generally diagnostically weaker than RM scores (Vrij, 2008a).

12.3.2. *Word-count, duration and speech rate*

In Studies 2 and 6 word-count, duration and speech rate were effective in discriminating between truthful and deceptive accounts, and were more or less equal in this respect. Hence, as predicted, truthful accounts were longer than deceptive accounts in both duration and word-count and the number of words produced per second was significantly greater for truth-tellers than for liars. Moreover, in Study 2, this finding was consistent for both writers and speakers. It was noted in Chapter 7 that in some previous studies speech rate has not discriminated between truthful and deceptive accounts (Vrij et al., 2000; 2004). However, in these studies, the definition of speech rate included the time-interval between the interviewer's question and the answer. This was not the case in the present studies, as in those of Vrij et al. (2008), where the definition of speech rate did not include the latency period, and was diagnostically stronger in distinguishing between truthful and deceptive accounts.

However, as noted previously, these measures were affected to some degree by the number of people present when the account was given; hence when one interviewer was present, the accounts were longer, both in terms of their duration in time and word-count than when no or two interviewers were present. Also speech rate was fastest when the account was given in a room with one person than when in a room with no persons or two persons present. Although these results may have possible explanations in terms of factors such as evaluation apprehension and cognitive load, the presence of others did not actually affect the ability to detect deception when applying these measures; nevertheless, they suggest again that the presence of others may create a possible confound in research in this area, if not systematically considered.

Given the association between word-count and RM scores reported here and in previous research (see, for example, Memon, et al, 2010), one could argue that word-count, in particular, could be used as an additional, or proxy measure of RM. However, as has been emphasised on a number occasions in this thesis, although word-count differences between truthful and deceptive accounts are sometimes included in preliminary analyses, perhaps surprisingly, the number of words contained in an account has never been identified or used as an official RM cue. In contrast, the number of words in an account has frequently been used to define the CBCA criterion *quantity of details* (Lamb, Stenberg, Esplin, Hershkowitz, Orbach & Hovav, 1997; Santtila, Roppola, Runtti & Niemi, 2000) which is considered one of the most effective diagnostic criteria in CBCA (Amado et al., 2015; Vrij, 2005). The issue of the use of word-count in its own right will be considered again shortly.

12.4. Some comparative considerations

Taken as a whole, perhaps the key findings of this thesis can be summarized as follows: the RM approach can be used to distinguish statistically between truthful and deceptive accounts to some extent; however, its effectiveness in this respect is dependent on the conditions under which it is applied. Thus the diagnostic efficacy of the RM approach is potentially highly influenced by its susceptibility to a number of different moderators. In other words, in discussions about whether (or not) researchers and practitioners should use the RM for veracity assessments, consideration should be given to the circumstances under which it will be used. Although this might appear to be glaringly obvious and predictable conclusion, it seems, nevertheless, to be the case that researchers have paid little if any attention to its implications.

Perhaps most important of the relevant findings in this respect, is that when frequency counts are used, the approach weakens if word-count is standardised. This finding suggests that future researchers and practitioners in this area should adopt a more uniform approach to this issue; so instead of deciding autonomously whether and how to standardise on an ad hoc basis, factors such as word-count and duration should be considered formally and systematically. Ignoring such factors may not only potentially jeopardise the diagnostic value of RM scores, it may also influence the impact of other RM moderators. Therefore, if there is a key message in terms of the effects of moderators, it is that if one wishes to list or classify the different moderators that influence verbal deception judgments using the RM approach, then starting with word-count might be a good idea, since all comparisons will inevitably involve accounts of various lengths. In other words, word-count can possibly be seen

as the primary moderator in as much as upon which all the other moderators function to affect RM scores.

However, given its significance in this respect, this again raises the issue of whether word-count should be handled as a moderator or as a cue to deception in its own right; i.e. in the same way that word-count is used to define the CBCA credibility-criterion *quantity of details* (Lamb et al, 1997; Santtila et al., 2000). Indeed, given word-count per se seems to significantly predict veracity, if we are not going to standardise scores for word-count, is there actually any point bothering with the RM criteria at all? A number of findings suggest that it would be wise to exercise caution before becoming too enthusiastic about the virtues of word-count as a cue to deception. For instance, notwithstanding the general trend that, other things being equal, truthful accounts tend to be longer than deceptive accounts, there have been studies in which truthful accounts were not longer than deceptive accounts (Sporer, 1997; Vrij, Mann, Kristen & Fisher, 2007), or were in fact shorter (Sporer & Sharman, 2006); this was also shown for Video 1 material in Study 1 here. Furthermore, a recent meta-analysis of linguistic cues accessed by computer programs has questioned whether word-count per se can generally be considered a reliable cue to deceptive behaviour (Hauch et al., 2012). This finding is in line with other research using the Linguistic Inquiry Word-count (LIWC) computer software which also shows that word-count per se is not a reliable cue to deception (Masip, Bethencourt, Lucas, Sánchez-San Segundo, & Herrero, 2012; Williams,Talwar, Lindsay, Bala, & Lee, 2014). In other words, to predict veracity with any degree of accuracy, the variable of word-count needs to be considered in conjunction with other RM cues, and with reference to the conditions under which the study is conducted.

This in turn raises the issue of which of the measures studied in this thesis best discriminates between truthful and untruthful accounts. To investigate this, ideally one would conduct a Binary Logistic Regression (BLR) or Discriminant Function Analysis (DFA) with truthful and deceptive accounts as the dependent variable, and the other measures as predictors, to determine the relative abilities of the ratings and objective measures in discriminating between truthful and deceptive accounts. However, BLR and DFA on these lines was not feasible in most of the present studies because, as a number of researchers have noted, at present, there is no generally accepted method for conducting repeated measures BLR or DFA, and no readily available software to conduct such an analysis (see Reinerman-Jones, 2011). One cannot, for example, simply assign two scores to each participant from each level of a within-subjects variable and plug them into standard software, as this would clearly violate the assumption of independence of cases (which is serious). It can also be noted here, that if one attempted to perform a similar analysis on the data from Study 1 to assess the relative contribution of each of the various RM criteria this would be invalid due to insufficient cases per variable (see, for example, Harrell, 1984; Tabachnick, & Fidell, 2001).

It may be informative, however, to look at how the rating and more objective criteria fare in discriminating between truthful and deceptive accounts in terms of their relative effect sizes. In accordance with convention for ANOVA, in the present studies the effect sizes were calculated in terms of $\eta^2_{p.}$ According to convention also, η^2_p sizes of .01, .06 and .14, represent small medium and large effect sizes, respectively (Cohen, 1988; Hattie, 2009). With this mind, Table 12.1 shows the relative effect sizes for significant differences between truthful and deceptive

accounts for Total RM ratings and the various objective measures, in descending rank order for Studies 2, 4 and 6; all were in the predicted direction.

In terms of the effect size rankings, raw RM frequencies before standardisation were diagnostically the strongest cue (evidenced in two studies) followed by time spent (duration of accounts) and subjective RM ratings. Word-count, speech rate and duration also predicted veracity with large effect sizes. Notably, word-count was further down the hierarchy, though still with a relatively large effect size. This would suggest that there is indeed merit in considering RM criteria in addition to (or rather than simply) word-count, in attempting to distinguish between truthful and deceptive reports.

A similar comparative analysis can be done for the various RM criteria. The relative effect sizes for the significant results in Studies 2, 4 and 6 are displayed hierarchically in Table 12.2; again all were in the predicted direction.

Table 12.1 Effect sizes for differences between truthful and deceptive accounts for Total RM ratings and raw frequencies, for Studies 2, 4 and 6.

Measure	η^2_p
Unstandardised Total RM Frequencies (Study 6)	.50
Unstandardised Total RM Frequencies (Study 4)	.49
Time Spent (Study 6)	.33
Total RM subjective ratings-not blind (Study 2)	.32
Overall Length (Study 6)	.30
Words per second (Study 2)	.30
Time spent (Study 2)	.28
Overall Length (Study 2)	.27
Standardised Total RM Frequencies* (Study 6)	.26
Words per second (Study 6)	.14

* standardised for account duration

As previously reported, unstandardised (for word-count) scores were diagnostically stronger than standardised scores. Also, temporal and spatial frequency information, along with ratings for vividness, were diagnostically the stronger cues. It can be noted, however, that if we ignore the rather anomalous Study 1, vividness was used in only one study, whereas the other cues were used in all three studies. The effect sizes for all criteria were large, but exceptionally so for spatial frequency information. Interestingly also, standardisation for duration did not

seem to have less of an effect than standardisation for word-count, though this form of standardisation has received little if any attention in the literature.

Table 12.2. Effect sizes for differences between truthful and deceptive accounts for RM criteria ratings (R), standardised (S) and unstandardised (U) raw frequencies, for studies 2, 4 and 6.

Measure	η^2_p
(U) Spatial information raw frequencies (Study 4)	.47
(U) Spatial information raw frequencies (Study 6)	.38
(U) Temporal information raw frequencies (Study 4)	.31
(U) Temporal information raw frequencies (Study 4)	.31
(S) Cognitive information raw frequencies* (Study 6)	.31
(R) Vividness ratings (Study 2)	.26
(S) Cognitive information raw frequencies** (Study 6)	.26
(S) Temporal information raw frequencies** (Study 6)	.24
(U) Perceptual information raw frequencies (Study 6)	.16

* standardised for word-count; ** standardised for account duration

12.5. Theoretical importance of findings: practical implications and recommendations

Assuming the present findings have any validity, what implications might they have for improving diagnostic value of RM procedures? In considering this question, one should perhaps be mindful of how any potential changes may affect the original research upon which the RM theory was built. It can be noted here that there have

already been numerous modifications to the original RM technique since its inception; these include generally redefining and adding to the criteria (Alonsso-Quecuty, 1993), breaking down contextual information into spatial and temporal information (Barnier et al, 2005; Elntib et al., 2014; Sporer, 1997; Sporer & Hamilton, 1996; Sporer & Sharman, 2006) and conceptualising perceptual information in a variety of different ways (Granhag et al, 2001, 2006; Roberts & Lamb, 2010; Vrij et al., 2000, 2001). In fact, in the original theory (Johnson & Suengas, 1981), and the first studies that ever used RM in a lie detection context (Alonso-Quecuty, 1992; 1995), only two criteria were applied, namely internal (i.e. cognitive information) and external (i.e. perceptual, spatial and temporal) information. Bearing this in mind, it may be worth considering first the merits of reducing or rationalising the RM criteria, and concentrating on those that appear to be most effective in discriminating between truthful and deceptive accounts.

12.5.1. *Using fewer RM criteria: implications for RM theory and procedures*

To reiterate, the present findings suggested that, on balance, the most useful individual RM criteria were unstandardised raw frequencies of temporal and spatial information and global subjective ratings of vividness. In contrast, reconstructability and realism were not useful at all while perceptual, and affective and cognitive information produced mixed results, though, in terms of trends, the latter tended to be stronger diagnostically (see Study 6). Given this, one might propose that if one were to offer a more efficient and streamlined RM procedure it would consist of a rather heterogeneous mixture of unstandardised frequency ratings of temporal, spatial and perceptual information, alongside a subjective vividness rating and

possibly a cognitive (operations) information frequency rating following standardisation for duration.

Arguably, excluding the criteria of reconstructability and realism would not be contrary to the original theory which made no reference to these criteria (Johnson & Raye, 1981). Moreover, as mentioned previously, reconstructability and affective information, in particular, have been shown to be weak diagnostically in previous research (Masip et al., 2005). However, at this stage, notwithstanding issues concerning the robustness and generality of the present results, proposing an abbreviated version of this kind is problematic for a number of reasons. Most important of these is the fact that the best diagnostic predictors were in fact *Total* frequency and rating scores; i.e. although certain criteria appeared to be diagnostically stronger than others, the best predictors were the sum totals of the criteria, which included those criteria which, by themselves, seemed to be weak. This was particularly true of the rating scale measures. Given this, it would not necessarily follow that some kind of hetereogeneous mini RM procedure would outperform a full RM scale. Rather, the present results suggest that what is really necessary is a full-blown componential analysis that systematically examines the effects of different combinations of criteria and their associated totals. This exercise was beyond the scope of the present thesis, but, in any case, it would have been methodologically unacceptable to have examined retrospectively the diagnostic value of various 'cherry picked' combinations of criteria.

12.5.2. *The theoretical implications of standardisation*

The results also indicated that, using frequencies, the RM approach was generally most effective before standardisation. In fact, it could only still discriminate between

truthful and deceptive accounts after standardising for duration. Reasons for, and possible implications of, these findings have been presented in Chapter 11, but for clarity, it may be worth reprising some of the possible implications here. For example, if, as has been argued, account length and quantity of details are of themselves reliable cues to deception (Colwell et al., 2002; Memon et al., 2010; Vrij et al., 2004; Vrij, 2015), it seems rather counterproductive to correct for word-count. Perhaps most important, however, RM theory was originally formulated on the fundamental idea that genuine memories will contain more external and less internal characteristics than imagined memories (Johnson & Raye, 1981). As such, there is no reference in the RM research to the density of various kinds of information within accounts (Johnson et al, 1993; Johnson & Raye, 1981); but rather the theory proposes that truthful accounts will overall contain 'more perceptual, spatial and temporal, semantic and affective information and less information about cognitive operations' (Johnson et al., 1993, p.4). Arguably, therefore, standardising RM scores per 100 words, or according to some other formula, is not justified by the original theory. Moreover, by the same token, it could be argued that, notwithstanding the predictions for cognitive information, one would expect accounts possessing 'more perceptual, spatial and temporal, semantic and affective information' to be longer than those not possessing such information; hence there is no obvious reason for excluding word-count per se as an additional RM criterion, as it could be considered to be an indirect or proxy measure of RM (as indicated by its status in the effect size hierarchy in Table 12.1).

12.5.3. *Implications of effects of moderators; modality, written/spoken, presence of others and demand characteristics.*

The implications of the findings concerning the other moderators can be summarized as follows.

In general, there was no significant or consistent evidence that first or second-language accounts, written or spoken statements, or accounts produced in a room with one, two or no persons in the room, affected the ability to discriminate between truthful and deceptive accounts. Nevertheless, all of these moderators produced a variety of significant main effects with regard to RM scores, both in totality and at the individual criterion level. Without going over all of the individual results again, there is an obvious point to be re-emphasised here; i.e. when conducting RM assessments, researchers and practitioners need to systematically control for, or take into account, the conditions under which the accounts were generated. This can be illustrated with just one of many examples. For instance, the results from Study 4 showed a significant effect of modality on cognitive information both before and after word-count standardisation, such that spoken accounts contained more cognitive information. Especially given the general failure of cognitive information to predict veracity, this again emphasises the importance of not assuming that written and oral accounts will be equivalent in terms of their effects on RM scores. For example, a simple comparison of cognitive information from an oral and a written account might give the spurious impression that the oral account is more likely to be deceptive. With regard to RM scores generally, similar problems could arise from comparing accounts produced with different numbers of individuals present (which can frequently happen in forensic investigations), and making comparisons between individuals who have different levels of language proficiency,

as both variables may affect RM scores independently of whether they are derived from truthful or deceptive accounts.

Also, assuming they are reliable and valid, the implications of the results of Study 3 with regard to demand characteristics are that, with regard to subjective ratings at least, there seems to be no advantage to keeping judges blind to the nature and purpose of the procedures; in fact, to do so may be counterproductive.

12.6. Some observations on the relationship between RM and CBCA

The present results may also have implications for a possible alignment between RM and its main 'competitor', CBCA. As mentioned on a number of occasions, one of the key differences between CBCA and RM is that CBCA uses quantity of details, which is associated directly with word-count, foremost as a lie-detection criterion, whereas RM researchers occasionally use the number of words contained in an account as a moderator, and never as a RM criterion. Nevertheless, some CBCA researchers have viewed word-count as a problem and have evolved a variety of strategies for dealing with it; these have included dismissing raw frequency CBCA measures altogether on the grounds that they are confounded by word-count (Granhag et al., 2006), or transforming the raw frequency scores for quantity of details into ratings (Stromwall et al., 2004) or dichotomous scales, split by the median score (Leal et al., 2015; Vrij et al, 2000). Somewhat paradoxically, therefore, it seems that, in CBCA quantity of details has effectively been viewed as both a cue and a moderator. However, one implication of the present results is that there may be advantages for both approaches to use word-count as a cue (indirect or otherwise) in its own right.

Indeed, in general, there appear to be a number of possible grounds for an alignment between CBCA and RM. For example, as mentioned previously, Vrij (2015) has argued that the CBCA criterion of *contextual information* shares a degree of conceptual overlap with RM spatial and temporal information. Moreover, RM's affective information criterion may, to some extent, encompass or overlap with CBCA's *subjective states* criterion (Vrij et al., 2015); although 'subjective states' includes both feelings and thoughts. Also, according to Colwell et al. (2007), RM's vividness criterion bares similarity with CBCA's *quantity of details*, while RM's realism is similar to CBCA's *logical structure*, although only the former takes plausibility into account (Vrij, 2015). These similarities have been empirically supported by correlational and factorial analyses (Sporer, 2004), but as such, they would likely to be subject to the effects of moderators in the same way as RM measures were in the present thesis; i.e. it might be worthwhile conducting a similar investigation of these moderators on CBCA scores.

12.7. Methodological limitations and implications for future research

Nevertheless, clearly, any conclusions drawn from the present results must be treated cautiously, as all of the studies were subject to limitations, many of which have already been mentioned. However, some the more obvious problems are highlighted again here.

The most obvious limitation of this thesis was the small sample sizes used, particularly in Studies 1 and 5. Small sample sizes inevitably reduce the generalizability of the results and reduce the analytical options in terms of more extensive multivariate analyses as there are insufficient cases per variable. On the other hand, small sample sizes also reduce the degrees of freedom and make it more

difficult to achieve a significant result (limiting Type 1 errors). Also, the use of small sample sizes in lie-detection research is not unusual, particularly a small number of raters (Harpster et al., 2009; Koper & Sahlman, 1991; Mann et al., 2002; ten Brinke & Porter, 2012; Villar et al., 2011; Vrij & Mann, 2001). A fundamental assumption here is that any real-life veracity assessment tool must be accurate even on a case to case basis to be of any real use in the criminal justice system. The fact that there may be significant mean differences between large populations may have little practical significance (considerations regarding the production of norms will be considered in the concluding section). From this viewpoint, therefore, what really counts is not one or two studies involving hundreds of participants, but consistent trends across a number of smaller studies involving a few individuals. What perhaps is needed at this stage, therefore, is a check of the robustness of the present findings by repeating the studies with similar small sample sizes, to see whether the results are consistent.

Although design issues were not investigated systematically, some observations can be made in the light of the present results. In many respects, the results of Study 1 could be considered a salutary lesson in the drawbacks of using between-subjects designs (where different individuals produce truthful and deceptive accounts). Of course, the results might have been better in Study 1 if a larger sample had been employed, and the materials had been more realistic etc., but it remains the case that between-subjects designs are inherently vulnerable to a variety of confounding individual difference effects, especially word-count, which make it more or less impossible to draw conclusions from anything but a very large sample (and even this remains to be established), which might produce results of little practical relevance. Hence the implication here seems to be that, both in research

and practice with regard to stimulus participants at least, within-subjects designs are likely to be the more informative.

The general benefits of using repeated-measures are well documented. For example, as alluded to above, they allow the researcher to control for some effects of individual differences. Another benefit of repeated-measures designs is that they require fewer participants than independent groups. However, as in the present studies, this kind of design (as well as sample size) can limit the analyses that can be conducted. As noted earlier, and more specifically here, multivariate BLR and DFA were not feasible because there is no generally accepted method for conducting repeated measures BLR or DFA and no readily available software to conduct such an analysis. This meant that, other than by comparing effect sizes, it was not possible to conduct a thorough and systematic assessment of the relative diagnostic strength of the various criteria and other forms of measurement. Indeed, a systematic componential assessment of RM, that specifies in detail the criteria applied and how they are measured, is long overdue.

This latter point emphasises another very obvious limitation of the present findings; the particular procedures employed. Because one of the main aims of the thesis was to investigate the possibility of producing an RM investigative tool with a definitional framework that could be used by relatively inexperienced raters/coders, and standardised across studies, the results are necessarily limited to the particular procedures and instructions used in this respect. Perhaps if the definitions had been more detailed or elaborated in some other ways, they might have been more effective. In defence of the procedures used, however, it could be argued that, unlike in many previous studies, at least an attempt was made to be transparent about the exact definitions used and instructions given. Moreover, it does not necessarily

follow that a more elaborate set of definitions and instructions would have fared any better. As research in the areas of personality and clinical diagnosis has shown, complex definitions and instructions do not necessarily lead to more effective prediction (Mischel, 1968); these are empirical matters that only further research can inform.

This in turn relates to another possible key factor in RM research, the background and training expertise of the coders. In general, results indicated that relatively non-experienced coders were successful to some degree in discriminating truthful from deceptive accounts using the RM; though not when they were completely inexperienced and experimentally blind. However, the assessments were more consistent when more experienced coders were used (as in Studies 4 and 6). However, this variable was not formally investigated, and, as a result, could be considered a possible confound; most particularly because those making the global subjective ratings had generally less relevant experience than those coding the frequencies. This raises the possibility that the diagnostic value of the ratings might have been increased if the raters had received more training. Similar concerns regarding the effects of insufficient training have been also expressed in relation to the CBCA criteria (Vrij, 2005). Interestingly, some recent research suggests that lie-detection training starts having a significant impact on trainees only when delivered in sufficient detail such that sessions last longer than 20 minutes (Hauch, Sporer, Micheal & Meissner, 2014). The training did not last this long in any of the studies in this thesis, though the coders in Studies 4 and 6 had previously spent considerable time conducting their own research on verbal lie-detection. If generalizable, therefore, the present results could be considered to endorse the view that some

degree of training in or experience of RM is preferable (as distinct from no training at all), but this in itself could be a confound if not applied in a consistent manner.

Another problem that has been touched upon concerns the ecological validity of the stimulus materials. With this in mind, the stimuli used in Study 1 (accounts of violent videos watched) were subsequently changed to allow the utilisation of more realistic stimuli (autobiographical accounts of personal significance) that potentially allow more emotional engagement with materials. As mentioned in Chapter 5, the LEI and similar approaches utilising autobiographical accounts have been popular amongst lie-detection researchers (see, for example, Barnier et al., 2005; Johnson, et al., 1988; Santilla et al, 1998; Sporer & Sharman, 2006) and have received support in perhaps the most complete meta-analytic study of RM research (Masip et al, 2005). However, even though the LEI procedure used here has been used in a number of previous studies, and notwithstanding the issues surrounding ground-truth discussed in Chapter 5, the motivation for participants to lie was lower than might usually occur in a real-life high stakes context. This may be important in that high-stakes situations, in which the motivation to lie is strong, may produce more reliable cues to deception (DePaulo & Morris, 2004; Porter, & ten Brinke, 2010; Wright-Whelan, Wagstaff & Wheatcroft, 2015a; 2015b; 2014). An obvious implication of this, therefore, is that future research into RM should attempt to use more high-stakes stimulus materials.

12.8. Conclusions

According to Vrij (2015), both RM and CBCA fulfil the four out of the five criteria used by the United States Supreme Court for admitting expert evidence in American federal courts; namely, they provide testable scientific research hypotheses, their

propositions have been tested, their error rate is known and they are both supported by relevant research. However, if the results of the present thesis are anything to go by, arguably, for RM at least, these assumptions may be considered somewhat overly optimistic, or at least they need qualifying. For example, given that RM criteria scores are subject to the kinds of moderating effects identified in this thesis, and many of these have not previously been assessed, it is not clear that the 'error rates' for RM are known.

It is also problematic that there appears to be no clear consensus as to how the various criteria are to be defined and measured, a problem that is frequently exacerbated by a lack of detail and general transparency in describing the procedures used. One is reminded here of problems that have occurred in relation to the practical application of the Cognitive Interview in forensic investigations, where, despite promising signs in the laboratory, police practitioners have faced huge difficulties deciding exactly how they should define and apply the various mnemonic techniques; indeed, often the problems have appeared so intractable that they have abandoned using the technique (Kebbell & Wagstaff, 1999; Kebbell, Milne & Wagstaff, 1999). And this relates to what is perhaps the major difficulty facing RM researchers, that of developing a protocol that might actually be useful for forensic investigators examining individual cases.

In this latter respect, perhaps the most discouraging feature of the present findings was that the RM criteria were not generally discriminating after standardisation. Given the large individual differences in the lengths of accounts given in non-deceptive situations in response to the same questions or stimulus materials, this suggests that it could potentially be very difficult to develop normative criteria which could be used in the field to classify individual cases. Of

course, there may some ways forward; for example, researchers have variously suggested other ways of standardising word-count, such as cues per 50 words (Vrij et al., 2000), the transformation of raw frequencies into a 5-point-rating scale (Memon et al., 2010; Vrij et al., 2008) and controlling for the duration of the account (Gnisci et al., 2010); or even combinations of these. Nevertheless, despite some promising findings for RM, therefore, both in this thesis and in the literature generally, we appear to be a long way off from devising a practical standardised RM procedure that can be applied to individual cases in everyday forensic practice.

Another possible avenue for inquiry, alluded to in Chapter 4, in the detection of deception in verbal accounts, is to examine the interaction between RM criteria and the use of verbal fillers (e.g. um, uh) which are found, as here, in oral accounts. Some investigators have argued that, in contrast with equivalent micro-level non-verbal behaviours (e.g. muscle micro movements), these kinds of paralinguistic cues may have been underestimated in lie-detection research (Brennan & Williams, 1995; Linell, 1982; 1998). The functions of such paralinguistic cues have variously been described as both accidental and intentional (Corley & Stewart, 2008), biophysical (e.g. essential in breathing and articulation), psychological (e.g. reflecting stress and anxiety), communicative (signalling new information to the speaker), emotional and linguistic (dividing the discourse into clauses/themes) (Esposito, Stejskal, Smekal & Bourbakis, 2007). Also, clinically they have been described as indicators of characteristics such as emotional instability (Mahl, 1959), and, psycholinguistically, as a sign of limited preparedness (Maclay & Osgood, 1959). These features suggest that they may have potential as cues for lie-detection, especially if combined with other cues. However, as yet, we have little comparative data on the relative efficacy of RM and alternative more complex computerised word-count deception detection

techniques. For example, the LIWC software provides 72 linguistic dimensions of speech, which can be further grouped into larger linguistic categories (e.g. Linguistic Processes, Psychological Processes, Personal Concerns, and Spoken Categories). Although some success has been reported using LIWC for lie-detection independently (with both adults and children, see for example Williams et al., 2014) or in conjunction with the RM framework (Bond & Lee, 2005; Newman et al, 2003), research that directly compares manual (as opposed to computerised) RM coding and LIWC is still very rare (Williams et al., 2014). Moreover, despite the merits of using computerised coding via LIWC (i.e. it can be faster, and more consistent and cost-effective), some have argued that human interpretation of the cues/criteria is still needed (Vrij et al., 2007)

Nevertheless, in the meantime, at the very least, the present results suggest that when judging the veracity of accounts using RM criteria, the scoring and other moderating variables identified in this thesis should be investigated systematically, and measured and applied consistently, if researchers wish to compare and replicate findings within and across studies.

References

Adams, S.H. & Jarvis, J.P. (2006). Indicators of veracity and deception: An analysis of written statements made to police. *The International Journal of Speech, Language and the Law, 13* (1), 1-22. doi:10.1558/sll.2006.13.1.1

Aiello, J. R., & Douthitt, E. A. (2001). Social facilitation from Triplett to electronic performance monitoring. *Group Dynamics: Theory, Research, and Practice, 5*(3), 163.

Allen, J.J., Mäthger, L.M., Barbosa, A., Hanlon, R.T. (2009) Cuttlefish use visual cues to control three-dimensional skin papillae for camouflage. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioural Physiology, 195* (6), 547-555.

Alonso-Quecuty, M. L. (1992). Deception detection and reality monitoring: a new answer to an old question (In F. Lo¨ sel, D. Bender and T. Bliesener (Eds.), *Psychology and Law: International Perspectives* (pp. 328-332). Berlin: Walter de Gruyter.

Alonso-Quecuty, M. L. (1995). Detecting fact from fallacy in child and adult witness accounts. In G. Davies, S. Lloyd-Bostock, M. McMurran and C. Wilson (Eds.), Psychology, *Law and Criminal Justice. International Developments in Research and Practice* (pp. 74-80). Berlin: Walter de Gruyter.

Amado, B. G., Arce, R., & Fariña, F. (2015). Undeutsch hypothesis and Criteria-based Content Analysis: A meta-analytic review. *The European Journal of Psychology Applied to Legal Context, 7*(1), 1-10.

Baldwin, J. (1993). Police Interview Techniques Establishing Truth or Proof?. *British Journal of Criminology, 33*(3), 325-352.

Ball, C.T. & O'Callaghan, J. (2008). Judging the accuracy of children's recall: A statement- level analysis. *Journal of Experimental Psychology: Applied*, *4*, 331-345.doi:10.1037/1076-898X.7.4.331

Barnier, A. J., Sharman, S. J., McKay, L., & Sporer, S. L. (2005). Discriminating adults' genuine, imagined, and deceptive accounts of positive and negative childhood events. *Applied Cognitive Psychology, 19*, 985–1001.

Baron, R. S. (1986). Distraction/conflict theory: Progress and problems. In L. Berkowitz (Ed.), *Advances in experimental social psychology (*Vol. 19, pp. 1 40). Orlando, FL: Academic Press.

Beaman, K. (1984) 'Coordination and Subordination Revisited: Syntactic Complexity in Spoken and Written Narrative Discourse', in D. Tannen (ed.), *Coherence in Spoken and Written Discourse*, Norwood, N.J.: Ablex.

Beaugrande, R. (1984). *Text Production: Toward a Science of Composition*. Norwood, NJ: Ablex

Bembibre, J., & Higueras, L. (2012). Comparative analysis of true or false statements with the source monitoring model and the cognitive interview: special features of the false accusation of innocent people. *Psychology, Crime & Law, 18*(10), 913-928.

Ben-Shakhar, G., & Elaad, E. (2003). The validity of psychophysiological detection of information with the Guilty Knowledge Test: a meta-analytic review. *Journal of Applied Psychology, 88*(1), 131-151

Ben-Shakhar, G., & Furedy, J. J. (1990). *Theories and applications in the detection of deception: A psychophysiological and international perspective. N*ew York: Springer-Verlag.

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgements. *Personality and Social Psychology Review, 10,* 214-234.

Bond, C. F., Jr., Kahler, K. N., & Paolicelli, L. M. (1985). The miscommunication of deception: An adaptive perspective. *Journal of Experimental Social Psychology, 21*, 331–345.

Bond, G. D., & Lee, A. Y. (2005). Language of lies in prison: Linguistic classification of prisoners' truthful and deceptive natural language. *Applied Cognitive Psychology, 19*, 313-329.

Bond, C. F., Jr., & Robinson, M. (1988). The evolution of deception. *Journal of Nonverbal Behaviour, 12,* 295–307.

Bond, C. F., Jr., Omar, A., Mahmoud, A., & Bonser, R. N. (1990). Lie-detection across cultures. *Journal of Nonverbal Behavior, 14*, 189–204.

Bogaard, G., Meijer, E. H., & Vrij, A. (2014). Using an Example Statement Increases Information but Does Not Increase Accuracy of CBCA, RM, and SCAN. *Journal of Investigative Psychology and Offender Profiling*, *11*(2), 151-163.

British Psychological Society. (2004). *A review of the current scientific status and fields of application of polygraphic deception-detection*. Final Report from the BPS working Party. Leicester, England: British Psychological Society.

Brennan, S. E. and Williams, M. (1995). The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language , 34,* 383-398.

Broadhurst, R.G., Cheng, K.H.W. (2005). The detection of deception: the effects of first and second-language on lie-detection ability. *Journal of Psychiatry, Psychology and Law,12,* 107–118.

Bross, I.D.J. (1971). Critical Levels, Statistical Language and Scientific Inference. In Godambe V.P. and Sprott, I. (eds.) *Foundations of Statistical Inference*. Toronto: Holt, Rinehart & Winston of Canada, Ltd.

Burgoon, J. K., & Buller D. B. (1994). Interpersonal deception: III.Effects of deceit on perceived communication and nonverbal behaviour dynamics. *Journal of Nonverbal Behavior, 18,* 155–184.

Caldwell-Harris, C.L. & Ayçiçeği-Dinn, A. (2009). Emotion and lying in a non-native language. *International Journal of Psychophysiology, 71,* 193–204.

Chafe. W. (1982). Integration and involvement in speaking, writing, and oral literature. In Deborah Tannen., (ed). *Spoken and Written Language: Exploring Orality and Literacy.* Norwood, NJ: Ablex.

Chafe, W., & Tannen, D. (1987). The relation between written and spoken language. *Annual Review of Anthropology, 16,* 383–407.

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

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304 – 1312. doi:10.1037/0003-066X.45.12.1304

Colwell, K., Hiscock-Anisman, C. K., & Memon, A. (2002). Interviewing techniques and the assessment of statement credibility. *Applied Cognitive Psychology*, 16, 287–300. doi: 10.1002/acp.788

Colwell, L.H., Miller, H.A., Miller, R.S., & Lyons, P (2006) US police officers' knowledge regarding behaviors indicative of deception: Implications for

eradicating erroneous beliefs through training, *Psychology, Crime & Law, 12,* 5, 489-503, doi:10.1080/10683160500254839

Corley, M., & Stewart, O. W. (2008). Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass*, *2*(4), 589-602.doi:10.1111/j.1749-818X.2008.00068.x

Cotlrell, N. B. (1972). Social facilitation. In C. G. McClintock (Ed.), *Experimental Social Osychology* (pp. 185- 236). New York: Holt, Rinehart & Winston

DaSilva, C. S., & Leach, A. M. (2013). Detecting deception in second-language speakers. *Legal and Criminological Psychology*, *18*(1), 115-127.

Davies, G. M., Westcott, H. L., & Horan, N. (2000). The impact of questioning style on the content of investigative interviews with suspected child sexual abuse victims. *Psychology, Crime and Law*, *6*(2), 81-97.

DePaulo, B. M., Lindsay, J. L., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin, 129*, 74–118.

DePaulo, B. M., & Morris, W. L. (2004). Discerning lies from truths: Behavioural cues to deception and the indirect pathway of intuition. *The detection of deception in forensic contexts*, 15-40.

Dilmon, R., (2009). Between thinking and speaking—Linguistic tools for detecting a fabrication . *Journal of Pragmatics, 41,* 1152–1170

Driscoll, L.N. (1994). A validity assessment of written statements from suspects in criminal investigations using the SCAN technique. *Police studies, 17,* 77-88.

Driskell, T. (2013). *Investigative Interviewing: A team-level approach* (unpublished Doctoral dissertation, University of Central Florida Orlando, Florida).

Driskell, T., Blickensderfer, E. L., & Salas, E. (2013). Is three a crowd? Examining rapport in investigative interviews. *Group Dynamics: Theory, Research, and Practice*, *17*(1), 1.

Ekman, P. (2001). *Telling lies: Clues to deceit in the marketplace, politics, and marriage.* New York: Norton.

Ekman, P., Friesen, W. V., & Simons, R. C. (1985). Is the startle reaction an emotion? *Journal of Personality and Social Psychology, 49,* 1416– 1426.

Elaad, E. (1990). Detection of guilty knowledge in real-life criminal investigations. *Journal of Applied Psychology, 75,* 521–529.

Elaad, E., Ginton, A., & Jungman, N. (1992). Detection measures in real-life criminal Guilty Knowledge Tests. *Journal of Applied Psychology, 77,* 757-767.

Elntib, S., Wagstaff, G. F., & Wheatcroft, J. M. (2014). The role of account length in detecting deception in written and orally produced autobiographical accounts using reality monitoring. *Journal of Investigative Psychology and Offender Profiling*, *12,* 185-198.

Esposito, A., Stejskal, V., Smekal, Z., & Bourbakis, N. (2007). The significance of empty speech pauses: Cognitive and algorithmic issues. In F. Mele, G. Ramella, S. Santillo, & F. Ventriglia (Eds.), *Advances in brain, vision, and artificial intelligence. Lecture notes in computer science* (Vol. 4729, pp. 542–554). Heidelberg: Springer.

Evans, J. R., & Michael, S. W. (2014). Detecting deception in non-native English speakers. *Applied Cognitive Psychology*, *28*(2), 226-237.

Fussell, S.R. (Ed.) (2014). *The Verbal Communication of Emotions: Interdisciplinary Perspectives.* London: Rouledge.

Garry, M., Manning, C. G., Loftus, E. F., & Sherman, S. J. (1996). Imagination inflation: Imagining a childhood event inflates confidence that it occurred. *Psychonomic Bulletin and Review, 3*, 208-214. doi:10.3758/BF03212420

Glass, G. V., & Stanley, J. C. (1970). *Statistical Methods in Education and Psychology*. Englewood Cliffs, NJ: Prentice-Hall.

Gnisci, A., Caso, L., & Vrij, A. (2010). Have you made up your story? The effect of suspicion and liars' strategies on reality monitoring. *Applied Cognitive Psychology, 24*, 762–773. doi:10.1002acp.1584

Goffman, E. (1974). *Frame analysis: An essay on the organization of experience*. New York: Harper & Row.

Gordon, N. J. and W. L. Fleisher. (2002). *Effective Interviewing and Interrogation Techniques.* San Diego, CA: Academic Press

Granhag, P.A. & Stromwall, L.A. (2004). Research on deception detection. In P.A. Granhag & L.A. Stromwall. *The detection of deception in forensic settings* (pp.3-14). Cambridge : Cambridge University Press.

Granhag, P.A., Strömwall, L.A., & Landström, S. (2006). Children recalling an event repeatedly: Effects on RM and CBCA scores. *Legal and Criminological Psychology, 11*, 81–98.

Granhag, P. A., Stromwall, L. and Olsson, C. (2001). Fact or fiction? Adults' ability to assess children's veracity. Paper presented at the 11th European Conference on Psychology and Law, Lisbon, Portugal, June 2001.

Greenhalgh, A. & Branigan, S. (2008). Martin. Available from Elk films, 128 Georgian Village, Castleknock, Dublin 15, Ireland

Halliday, M.A.K.(1989).*Spoken and Written Language*. Oxford: Oxford University Press.

Halliday, M.A.K. (2001). *Analysing English in a global context.* Routledge in association with Macquarie University and The Open University

Hancock, J. T., Curry, L., Goorha, S., & Woodworth, M. T. (2008). On lying and being lied to: A linguistic analysis of deception. *Discourse Processes, 45*, 1–23.

Hauch, V., Blandón-Gitlin, I., Masip, J., & Sporer, S. L. (2014). Are computers effective lie detectors? A meta-analysis of linguistic cues to deception. *Personality and Social Psychology Review*, 1088868314556539.

Hauch, V., Sporer, S. L., Michael, S. W., & Meissner, C. A. (2014). Does training improve the detection of deception? A meta-analysis. *Communication Research*, 0093650214534974.

Harpster, T., Adams, S. H. and Jarvis, J. P. (2009). Analyzing 911 homicide calls for indicators of guilt or innocence). *Homicide Studies, 13* (1), 69-93

Harrell, F. E. , Lee, K., Califf, R.M., Pryor, D.B. & Rosati, R.A. (1984). Regression modelling strategies for improved prognostic prediction. *Statistics in Medicine*, 3, 143-152.

Hartwig, M., Granhag, P.A., Stromwall, L., & Doering, N. (2010). Impression and information management: On the strategic self-regulation of innocent and guilty suspects [Special issue]. *The Open Criminology Journal, 3*, 10–16.

Hattie, J. (2009). *Visible Learning*. London: Routledge.

Horgan, J. J., & Horgan, J. J. (1979). *Criminal investigation*. Gregg Division, McGraw-Hill.

Honts, C. R. (1994). Assessing children's credibility: Scientific and legal issues in 1994. North Dakota, *Law Review, 70,* 879–903

Honts, C. R. (2004). The psychophysiological detection of deception. In P. Granhag & L. Stromwall (Eds.), *Detection of deception in forensic contexts*. London: Cambridge.

Horowitz, S. W., Lamb, M. E., Esplin, P. W., Boychuk, T. D., Krispin, O., & Reiter-Lavery, L. (1997). Reliability of criteria-based content analysis of child witness statements. *Legal and Criminological Psychology, 2,* 11-21. doi: 10.1111/j.2044-8333.1997. tb00329.x

Horvath, F., Jayne, B. & Buckley, J. (1994). Differentiation of truthful and deceptive criminal suspects in behaviour analysis interviews. *Journal of Forensic Sciences, 39,* 793–807.

Howell, D.C. (1992). *Statistical Methods for Psychology*. Belmont, CA: Duxbury.

Inbau, F. E., Reid, J. E., Buckley, J. P., & Jayne, B. C. (2001). *Criminal interrogation and confessions* (4th ed.). Gaithersburg, MD: Aspen.

Johnson, M. K. and Raye, C. L. (1981). Reality monitoring. *Psychological Review, 88, 67,* 85.

Johnson, M. K., Foley, M. A., Suengas, A. and Raye, C. L. (1988). Phenomenal characteristics of memories for perceived and imagined autobiographical events. *Journal of Experimental Psychology: General , 117,* 371-376.

Johnson, M. K., Hashtroudi, S. & Lindsay, D. S. (1993). *Source monitoring. Psychological Bulletin , 114,* 3-29.

Johnson, M. K. and Suengas, A. (1989). Reality monitoring judgments of other people's memories. *Bulletin of the Psychonomic Society, 27,* 107-110.

Jurgens, A., El-Sayed, A. & Suckling, D. (2009). Do carnivorous plants use volatiles for attracting prey insects? *Functional Ecology, 23,* 875–887.

Kassin, S. M., & Fong, C. T. (1999). "I'm innocent!": Effects of training on judgments of truth and deception in the interrogation room. *Law and Human Behavior, 23*, 499–516.

Kincaid, H.V., & Bright, M. (1957). The tandem interview: A trial of the two-interviewer team. *Public Opinion Quarterly 21,* 304-312.

Kirk, R. E. (1968). *Experimental design: Procedures for the behavioral sciences*. Belmont, CA: Brooks/Cole Publishing Company.

Kebbell, M. R., Milne, R., & Wagstaff, G. F. (1999). The cognitive interview: A survey of its forensic effectiveness. *Psychology, Crime and Law, 5*(1-2), 101-115.

Kebbell, M. R., & Wagstaff, G. F. (1999). The effectiveness of the cognitive interview. In D. Canter &L. Alison (Eds.), *Interviewing and deception* (pp. 25–39). Aldershot: Ashgate Publishing.

Köhnken, G. (2004). Statement validity analysis and the 'detection of the truth'. In P. A. Granhag & L. A. Strömwall (Eds.), *Deception detection in forensic contexts* (pp. 41–63). Cambridge, England: Cambridge University Press.

Koper, R. J., & Sahlman, J. M. (1991). The behavioural correlates of real-world deceptive communication. Paper presented at the Annual Meeting of the International Communication Association. Retrieved from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_ storage_01/0000019b/80/24/18/ae.pdf

Kostelnik, J.O. & Reppucci, N.D. (2009). Reid training and sensitivity to developmental maturity in interrogation: results from a national survey of police. *Behavioural Sciences and The Law, 23*(3), 361-379.

Kroll, B . (1977). *Combining ideas in written and spoken English.* In Bennett, T. L. (ed). *Discourse A cross Time and Space. Southern California Occasional Papers in Linguistics 5.* Los Angeles: Dept. Linguist, University of South California.

Lamb, M. E., Sternberg, K. J., Esplin, P. W., Hershkowitz, I., Orbach, Y., & Hovav, M. (1997). Criterion-based content analysis: A field validation study. *Child Abuse & Neglect, 21*(3), 255-264.

Larsson, A. S. & Granhag, P. A. (2005). Interviewing children with the cognitive interview: Assessing the reliability of statements based on observed and imagined events. *Scandinavian Journal of Psychology, 46*, 49–57. doi:10.1111/j.1467-9450.2005.00434.x

Leach, A. M., & Da Silva, C. S. (2013). Language Proficiency and Police Officers' Lie-detection Performance. *Journal of Police and Criminal Psychology*,*28*(1), 48-53.

Leal, S., Vrij, A., Warmelink, L., Vernham, Z., & Fisher, R. P. (2015). You cannot hide your telephone lies: Providing a model statement as an aid to detect deception in insurance telephone calls. *Legal and Criminological Psychology, 20*(1), 129-146.

Lindskold, S., &Walters, P. S. (1983). Categories for acceptability of lies. *The Journal of Social Psychology, 120,* 129-136.

Linell, P. (1982). The Concept of Phonological Form and the Activities of Speech Production and Speech Perception. *Journal of Phonetics, 10*: 37-72.

Linell, P. (1998). Discourse across boundaries: on recontextualisation and the blending of voices in professional discourse. *Text, 18(2),* 143-157.

Logue, M., Book, A. S., Frosina, P., Huizinga, T., & Amos, S. (2015). Using Reality Monitoring to Improve Deception Detection in the Context of the Cognitive Interview for Suspects. *Law and Human Behaviour, 39* (4), 360-367.

Lykken, D. T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology, 43*, 385–388.

Lykken, D. T. (1960). The validity of the guilty knowledge technique: The effects of faking. *Journal of Applied Psychology, 44*, 258–262.

MacDonald, J. M., & Michaud, D. L. (1992). *Criminal interrogation*. Denver, CO: Apache Press.

Maclay, H., & C. E. Osgood. (1959). Hesitation phenomena in spontaneous speech. *Word, 15,* 19-44.

Mahl, G. (1959) Measuring the Patient's Anxiety during Interviews from "Expressive" Aspects of His Speech. *Transactions of the New York Academy of Science, 2,* 21, 259-257.

Mann, S. & Vrij, A. (2006). Police officers' judgement of veracity, tenseness, cognitive load and attempted behavioural control in real-life police interviews. *Psychology, crime & Law, 12* (3), 307-319.

Mann, S., Vrij, A. & Bull, R. (2002). Suspects, lies and videotape. An analysis of authentic, high-stakes liars. *Law and Human Behaviour, 26,* 365-376.

Mann, S., Vrij, A., & Bull, R. (2004). Detecting true lies: Police officers' ability to detect deceit. *Journal of Applied Psychology, 89,* 137–149.

Mann, S., Vrij, A. & Bull, R. (2006). Looking through the eyes of an accurate lie detector. *The Journal of Credibility Assessment and Witness Psychology, 7* (1), 1-16.

Manzanero, A.L. & Diges, M. (1995). Effects of preparation on internal and external memories. In G.Davies, S.M.A. Lloyd-Bostock, M. McMurran and C.Wilson (Eds.): *Psychology, law and criminal justice. International developments in research and practice (pp.56-63).* Berlín: W. De Gruyter & Co.

Masip, J., Bethencourt, M., Lucas, G., Segundo, M. S. S., & Herrero, C. (2012). Deception detection from written accounts. *Scandinavian Journal of Psychology, 53*(2), 103-111.

Masip, J., Sporer, S., Garrido, E., & Herrero, C. (2005). The detection of deception with the reality monitoring approach: A review of the empirical evidence. *Psychology, Crime, & Law, 11*, 99–122.

McQuaid, S. M., Woodworth, M., Hutton, E. L., Porter, S., & ten Brinke, L. (2015). Automated insights: verbal cues to deception in real-life high-stakes lies. *Psychology, Crime & Law, 21* (7), 617-631.

Mischel, W. (1968). *Personality and Assessment*. London: Wiley.

Miller, G. R., deTurck, M. A., & Kalbfleisch, P. J. (1983). Selfmonitoring, rehearsal, and deceptive communication. *Human Communication Research, 10*, 97–117.

Memon, A., Fraser, J.,Colwell, K., Odinot, G., & Mastroberardino, S. (2010). Distinguishing truthful from invented accounts using reality monitoring criteria. *Legal and Criminological Psychology*, 15, 177–194. doi: 10.1348/135532508X401382

Nahari, G., & Vrij, A. (2014). Are you as good as me at telling a story? Individual differences in interpersonal reality monitoring. *Psychology, Crime & Law*, *20*(6), 573-583.

Nahari, G., Vrij, A., & Fisher, R. P. (2012). Does the truth come out in the writing? Scan as a lie-detection tool. *Law and human behaviour*, *36*(1), 68.

Nakayama, M. (2002). Practical use of the concealed information test for criminal investigation in Japan. In M. Kleiner (Ed.), *Handbook of polygraph testing* (pp. 49–86). London, UK: Academic Press.

Nakanishi, R. & Imai-Matsumura, K. (2008). Facial skin temperature decreases in infants with joyful expression, *Infant Behaviour and Development, 31*, 137–44.

National Research Council (2003). *The polygraph and lie-detection.* Committee to Review the Scientific Evidence on the Polygraph. Washington, DC: The National Academic Press.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin, 29*, 665-675.

Nozawa, A. & Tacano, M. (2009). Correlation analysis on alpha attenuation and nasal skin temperature, *Journal of Statistical Mechanics, 17*, 1-10.

Orne, M.T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, 17, 776-783.

Otgaar, H., Candel, I., Memon, A., & Almerigogna, J. (2010). Differentiating between children's true and false memories using reality monitoring criteria. *Psychology, Crime & Law*, *16*(7), 555-566.

Oxford English Dictionary (8th ed.). (1990). Oxford, UK: Clarendon Press.

Paddock, J. R., Noel, M., Terronova, S., Eber,H. W.,Manning, C. G., & Loftus, E. F.(1999).Imagination inflation and the perils of guided visualization. *Journal of Psychology, 133*, 581-595.

Pekar, S. & J. Kral. (2002). Mimicry complex in two central European zodariid spiders (Araneae: Zodariidae): how Zodarion deceives ants. *Biological Journal of the Linnean Society*, 75, 517–532.

Porter, S. & ten Brinke, L. (2010). The truth about lies: What works to detect high stakes-deception. *Legal and Criminological Psychology* , *15*, 57–75.

Porter, S., & Yuille, J. C. (1996). The language of deceit: an investigation of the verbal clues to deception in the interrogation context. *Law and Human Behaviour, 20*, 443–459.

Pu, M. M. (2006) Spoken and Written Narratives: A Comparative Study. *Journal of Chinese Language and Computing 16(1)*, 37-62.

Rabon, D. (1992), *Interviewing and Interrogation*. Durham, NC: Carolina Academic Press.

Raskin, D. C. and Esplin, P. W. (1991). Assessment of children's statements of sexual abuse. In J. Doris (Ed.), *The Suggestibility of Children's Recollections* (pp. 153-164). Washington, DC: American Psychological Association.

Redlich, A. D. (2004). Mental illness, police interrogations, and the potential for false confession. *Psychiatric Services, 55*, 19–21.

Reinerman-Jones, L., Taylor, G., Cosenzo, K., & Lackey, S. (2011). Analysis of multiple physiological sensor data. *Foundations of Augmented Cognition.*

Directing the Future of Adaptive Systems Lecture Notes in Computer Science, 6780, 112-119.

Roberts, K., & Lamb, M. (2010). Reality-monitoring characteristics in confirmed and doubtful allegations of child sexual abuse. *Applied Cognitive Psychology*, *24*, 10491079. doi: 10.1002/acp.1600.

Santtila, P., Roppola, H. and Niemi, P. (1998). Assessing the truthfulness of witness statements made by children (aged 7/8, 10/11, and 13/14) employing scales derived from Johnson and Raye's model of Reality Monitoring. *Expert Evidence, 6,* 273-289.

Santtila, P., Roppola, H., Runtti, M., & Niemi, P. (2000). Assessment of child witness statements using Criteria-Based Content Analysis (CBCA): The effects of age, verbal ability, and interviewer's emotional style. *Psychology, Crime and Law*, *6*(3), 159-179.

Sapir, A. (1987/2000). *The LSI course on scientific content analysis* (SCAN). Phoenix, AZ: Laboratory for Scientific Interrogation.

Schmider,E., Ziegler,M., Danay, E., Beyer,L., & Bühner, M. (2010). Is it really robust? Reinvestigating the Robustness of ANOVA against violations of the normal distribution assumption. *Methodology*, 6(4), 147–151.

Shavelson, R. J. (1996). *Statistical reasoning for the behavioral sciences* (3rd ed.). Boston: Allyn and Bacon

Simmel, G. (1964). *The sociology of Georg Simmel* (K. H. Wolf, Ed. and Trans.). Glencoe, IL: Free Press.

Smith, N. (2001). *Reading between the lines: An evaluation of the Scientific Content Analysis technique (SCAN)*. Home Office, Policing and Reducing Crime Unit, Research, Development and Statistics Directorate.

Sporer, S. L. (1997). The less travelled road to truth: verbal cues in deception detection in accounts of fabricated and self-experienced events. *Applied Cognitive Psychology, 11,* 373-397.

Sporer, S.L. (2004). Reality monitoring and detection of deception. In P.A. Granhag, & L.A. Stromwall (Eds.), *Deception detection in forensic contexts* (pp. 64–102). Cambridge, England: Cambridge University Press.

Sporer, S. L. and Hamilton, S. C. (1996). *Should I believe this? Reality monitoring of invented and self-experienced events from early and late teenage years*. Poster presented at the NATO Advanced Study Institute. Port de Bourgenay, France, June 1996

Sporer, S. L., & Sharman, S. J. (2006). Should I believe this? Reality monitoring of accounts of self-experienced and invented recent and distant autobiographical events. *Applied Cognitive Psychology, 20,* 837–854.

Steller, M., & Kohnken, G. (1989). Criteria-Based Content Analysis. In D. C. Raskin (Ed.),*Psychological methods in criminal investigation and evidence* (pp. 217– 245). New York: Springer-Verlag.

Stiff, J. B., & Miller, G. R. (1986). "Come to think of it":Interrogative probes, deceptive communication, and deception detection. *Human Communication Research, 12,* 339–357.

Streeter, L. A., Krauss, R. M., Geller, V., Olsen, C., & Apple, W. (1977). Pitch changes during attempted deception. *Journal of Personality and Social Psychology, 35,* 345–350.

Strömwall, L.A., Bengtsson, L., Leander. L.,,& Granhag, P.A..(2004). Assessing children's statements: the impact of a repeated experience on CBCA and RM ratings. *Applied Cognitive. Psychology, 18,* 653–668.

Strömwall, L. A., & Granhag, P. A. (2005). Children's repeated lies and truths: Effects on adults'judgements and reality monitoring scores. *Psychiatry, Psychology and Law, 12,* 345–356. doi: 10.1002/acp.1288

Stuart-Fox, D.M. & Moussalli A. (2008)**.** Selection for social signalling drives the evolution of chameleon colour change. *Public Library of Science Biology, 6,* 22–29.

Szechtman, H., Woody, E., Bowers, K. S., & Nahmias, C. (1998). Where the imaginal appears real: a positron emission tomography study of auditory hallucinations. *Proceedings of the National Academy of Sciences*, *95*(4), 1956-1960.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using Multivariate Analysis*. Boston, MA: Allyn & Bacon.

Tanaka, H., Ide, H., & Nagashuma, Y. (2000). An attempt of feeling analysis by the nasal temperature change model. In *Systems, Man, and Cybernetics, 2000 IEEE International Conference on* (Vol. 2, pp. 1265-1270). IEEE.

ten Brinke, L., & Porter, S. (2012). Cry me a river: Identifying the behavioral consequences of extremely high-stakes interpersonal deception. *Law and Human Behavior,* 36(6), 469-477. doi: 10.1037/h0093929

Tickle-Degnen, L., & Rosenthal, R. (1990). The nature of rapport and its nonverbal correlates. *Psychological Inquiry, 1,* 285–293. doi:10.1207/ s15327965pli0104_1

Underwood, R. H. (1993). False witness: a lawyer's history of the law of perjury. *Ariz. Journal of International & Comparative Law*, *10*, 215-228.

Undeutsch, U. (1984). Courtroom evaluation of eyewitness testimony. International *Review of Applied Psychology, 33,* 51–67.

Utz, S. (2005). Types of deception and underlying motivation: what people think. *Social Science Computer Review*, *23*(1), 49-56.

Vernham, Z., Vrij, A., Mann, S., Leal, S., & Hillman, J. (2014). Collective interviewing: Eliciting cues to deceit using a turn-taking approach. *Psychology, Public Policy and Law*. DOI: org/10.1037/law0000015

Villar, G., Arciuli, J., & Mallard, D. (2012). Use of "um" in the deceptive speech of a convicted murderer. *Applied Psycholinguistics, 33* (1), 83-95. doi: 10.1017/S0142716411000117

Vrij, A. (1993). Credibility judgments of detectives: The impact of nonverbal behavior, social skills, and physical characteristics on impression formation. *Journal of Social Psychology, 133,* 601–610.

Vrij, A. (1995). Behavioral correlates of deception in a simulated police interview. *Journal of Psychology, 129,* 15–28.

Vrij, A. (2000). *Detecting lies and deceit: The psychology of lying and the implications for professional practice*. New York: John Wiley & Sons.

Vrij, A. (2005). Criteria-based content analysis: A qualitative review of the first 37 studies. *Psychology, Public Policy, and Law, 11,* 3–41.

Vrij, A. (2008a). *Detecting lies and deceit: Pitfalls and opportunities*. Chichester: Wiley

Vrij, A. (2008b). Nonverbal dominance versus verbal accuracy in lie-detection: A plea to change police practice. *Criminal Justice and Behaviour, 35,* 1323-1336.

Vrij, A. (2015). Verbal Lie-detection Tools: Statement Validity Analysis, Reality Monitoring and Scientific Content Analysis. *Detecting Deception: Current Challenges and Cognitive Approaches*. Chichester: Wiley

Vrij, A., & Akehurst, L. (1998). Verbal communication and credibility: Statement Validity Assessment. In A. Memon, A. Vrij, & R. Bull, *Psychology and law: Truthfulness, accuracy and credibility* (pp. 3–31). Maidenhead, Great Britain: McGraw-Hill.

Vrij, A., Akehurst, L., Soukara, S., & Bull, R. (2004). Let me inform you how to tell a convincing story: CBCA and reality monitoring scores as a function of age, coaching, and deception. *Canadian Journal of Behavioural Science, 36,* 113–126. doi:10.1037/h0087222

Vrij, A., Edward, K. and Bull, R. (2001). Stereotypical verbal and nonverbal responses while deceiving others. *Personality and Social Psychology Bulletin , 27*, 899- 909.

Vrij, A., Edward, K., Roberts, K. P., & Bull, R. (2000). Detecting deceit via analysis of verbal and nonverbal behaviour. *Journal of Nonverbal Behaviour, 24*, 239–263.

Vrij, A., Evans, H., Akehurst, L., & Mann, S. (2004). Rapid judgements in assessing verbal and nonverbal cues: Their potential for deception researchers and lie-detection. *Applied Cognitive Psychology, 18*, 283–296.

Vrij, A., & Granhag, P. A. (2007). Interviewing to detect deception. In S. A. Christianson (Ed.), *Offenders' memories of violent crimes* (pp. 279–304). Chichester, UK: Wiley.

Vrij, A., Granhag, P. A., & Porter, S. (2010). Pitfalls and opportunities in nonverbal and verbal lie-detection. Psychological, Science *in the Public Interest,11(3),* 89-121.

Vrij, A., & Heaven, S. (1999). Vocal and verbal indicators of deception as a function of lie complexity. *Psychology, Crime, & Law, 5*, 203–215.

Vrij, A., Leal, S., Granhag, P.A., Mann, S, Fisher, R.P., Hillman, J. & Sperry, K. (2009). Outsmarting the liar. The benefit of asking unanticipated questions, Law *and Human Behaviour.*

Vrij, A. & Mann, S. (2001a). Telling and detecting lies in a high-stake situation: the case of a convicted murderer. *Applied Cognitive Psychology, 15*, 187-203

Vrij, A. & Mann, S. (2001b). Who killed my relative? Police officers' ability to detect real-life, high-stakes lies. *Psychology, Crime and Law, 7*, 119-132

Vrij, A., Mann, S., & Fisher, R. (2006). An empirical test of the Behaviour Analysis Interview. *Law and Human Behaviour, 30*, 329–345.

Vrij, A., Mann, S. A., Fisher, R. P., Leal, S., Milne, R., & Bull, R. (2008). Increasing cognitive load to facilitate lie-detection: the benefit of recalling an event in reverse order. *Law and human behavior*, *32*(3), 253.

Vrij, A., Mann, S., Kristen, S., & Fisher, R. P. (2007). Cues to deception and ability to detect lies as a function of police interview styles. *Law and Human Behaviour, 31*, 499–518.

Vrij, A., Semin, G. R., & Bull, R. (1996). Insight into behaviour displayed during deception. *Human Communication Research, 22*, 544–562.

Wagstaff, G.F. (1981). *Hypnosis, Compliance, and Belief*, Brighton: Harvester/ New York: St Martin's Press.

Wagstaff, G. F., Wheatcroft, J., Cole, J. C., Brunas-Wagstaff, J., Blackmore, V., & Pilkington, A. (2008). Some cognitive and neuropsychological aspects of social inhibition and facilitation. *European Journal of Cognitive Psychology*,*20*(4), 828-846.

Walczyk, J. J., Griffith, D. A., Yates, R., Visconte, S. R., Simoneaux, B., & Harris, L. L. (2012). Lie-detection by Inducing Cognitive Load Eye Movements and Other Cues to the False Answers of "Witnesses" to Crimes. *Criminal Justice and Behavior*, *39*(7),887-909.

Warmelink, L., Vrij, A., Mann, S., & Granhag, P. A. (2013). Spatial and temporal details in intentions: A cue to detecting deception. *Applied Cognitive Psychology*, *27*(1), 101-106.

Weinstock, R. & Thompson, C. (2009). Commentary: ethics-related implications and neurobiological correlates of false confessions in juveniles. *Journal of the American Academy of Psychiatry and the Law.37*, 344-348.

Williams, S. M., Talwar, V., Lindsay, R. C. L., Bala, N., & Lee, K. (2014). Is the truth in your words? Distinguishing children's deceptive and truthful statements. *Journal of Criminology*, *2014*, 1-9.

Wright-Whelan, C Wagstaff, G., & Wheatcroft, J. (2014): Subjective Cues to Deception/Honesty in a High Stakes Situation: An Exploratory Approach, The *Journal of Psychology: Interdisciplinary and Applied*, DOI: 10.1080/00223980.2014.911140

Wright Whelan, C., Wagstaff, G., & Wheatcroft, J. M. (2015a). High stakes lies: police and non-police accuracy in detecting deception. *Psychology, Crime & Law*, *21*(2), 127- 138.

Wright Whelan, C., Wagstaff, G. F., & Wheatcroft, J. M. (2015b). Subjective Cues to Deception/Honesty in a High Stakes Situation: An Exploratory Approach. *The Journal of psychology*, *149*(5), 517-534.

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit formation. *Journal of Comparative Neurology and Psychology, 18*, 459-482.

Zajonc, R. B. (1965). Social facilitation. *Science, 149*, 269-27'4.

Zajonc, R. B. (1980). Compresence. In P. B. Paulus (Ed.), *Psychology of group influence* (pp. 35-60). Hillsdale, NJ: Erlbaum.

Zhou, L., Burgoon, J. K., Nunamaker, J. F., & Twitchell, D. (2004). Automating linguistics based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group Decision and Negotiation, 13*(1), 81-106.

Zhou, L., & Lutterbie, S. (2005). Deception across cultures: Bottom-up and top-down approaches. In *Intelligence and Security Informatics* (pp. 465-470). Springer Berlin Heidelberg.

Zhou, L. & Zhang, D. (2008). Following linguistic footprints: Automatic deception detection in online communication. *Communications of the ACM, 51* (9), 119-122.

Zuckerman, M., DePaulo, B.M., & Rosenthal, R. (1981). Verbal and nonverbal communication of deception. In U Berkowitz (Ed.) *Advances in experimental social psychology* (Vol 14, pp. 1-59). New York: Academic Press.

Zuckerman, M., Driver, R., & Koestner, R. (1982). Discrepancy as a cue to actual and perceived deception. *Journal of Nonverbal Behavior*, 7, 95–100.

Zuckerman, M., Kernis, M. R., Driver, R., & Koestner, R. (1984). Segmentation of behavior: Effects of actual deception and expected deception. *Journal of Personality and Social Psychology, 46,* 1173–1182.