Distinguishing truthful from invented accounts using reality monitoring criteria

Amina Memon1*, Joanne Fraser1, Kevin Colwell2, Geralda Odinot1 and Serena Mastroberardino1

1University of Aberdeen, Aberdeen, Scotland, UK
2Southern Connecticut State University, New Haven, USA

**Purpose.** Previous research has suggested that true and invented memories can be distinguished between using the reality monitoring criteria. Two different coding schemes were used to examine the correct classification of reports as truthful or deceptive on the basis of individual reality-monitoring (RM) criteria.

**Method.** Drawing upon the RM framework the present experiment examined transcripts of verbal accounts of eyewitnesses to a staged event or made up details about the incident. The statements were elicited during a face-to-face cognitive interview (CI) and were analysed by coders trained in the identification of criteria indicative of self-experienced and invented accounts (referred to here as Version 1) and a similar coding method (referred to here as Version 2).

**Results.** A distinction was made between ‘external memories’ (affective, perceptual, and contextual details) and ‘internal’ memories (cognitive operations). For Version 1, the results indicated a higher number of contextual and external details in the descriptions of experienced events and as is commonly found in the deception literature, truthful accounts were longer. For Version 2, temporal and auditory details were more frequent in true accounts. Contrary to prediction, there were also more references to cognitive operations in true accounts.

**Conclusions.** Any forensic application of RM should consider how external factors (characteristics of the event, motivation to deceive, and questioning style) influence the presence of RM criteria.

In many forensic situations, it is critical to be able to determine whether someone is providing an account of an experienced event or lying. The term lying here refers to an intentional act on the part of the communicator to deceive their audience. Thus, someone who mistakenly reports an event as having taken place is not lying (Vrij, 2000). The focus here is identifying systematic differences between reports of personally experienced events (referred to here as truthful accounts) and intentionally false accounts.
(or invented) accounts. A number of approaches have been used to assess the veracity of an individual’s account ranging from use of physiological measures such as the polygraph to various investigative approaches involving detailed analysis of statements and non-verbal behaviour. Research into verbal indicators of deception has been shaped by the original research on the content-based approach to evaluating the credibility of eyewitness testimony; the best known being criteria-based content analysis (CBCA).

The use of CBCA requires trained evaluators to judge the content of the testimony for the presence or absence of 19 criteria that are presumed to distinguish between truthful and false statements (for a detailed description of the Statement Validity Analysis procedure and research on CBCA see Vrij, 2005a; Vrij & Mann, 2006). The current paper examines, the potential of an alternative content-based approach for assessing statement veracity that is grounded in theory, namely the reality-monitoring (RM) approach.

The basis of the RM approach is that memories of experienced events differ in quality from memories of imagined events. According to Johnson and Raye (1981) perception of external stimuli produce two different types of memories with different characteristics. ‘External’ memories based on perceptual information, sensory processes, and contextual information and ‘Internal’ memories based on reasoning, imagination, and thought processes. The fundamental assumption behind RM is that memories originating from perception should have more perceptual information (visual details, sounds, smells, tastes, and physical sensations related to the event) and more contextual information (details about where and when the event took place). Memories for perceived events also include more information about feelings or apperceptive reactions to the target events (e.g. I thought it was strange) and reports of feelings such as anger, frustration or boredom (Hashtroudi, Johnson, & Chrosniak, 1990; Suengas & Johnson, 1988). Memories based on imagination or fantasies on the other hand, contain more information (at encoding) about cognitive operations. Hashtroudi et al. (1990) define ‘cognitive operations as referring to mental activities that are a part of creating or establishing a target event’ (e.g. thinking an event has occurred before). We compare the effectiveness of two coding schemes based on RM in distinguishing between accounts of an experienced (staged) event and a non-experienced (invented event).

Drawing upon the RM framework, Johnson, Foley, Suengas, and Raye (1988) developed the memory characteristics questionnaire (MCQ). In this questionnaire, people rated their phenomenal experience of both general memorial characteristics (e.g. vividness), more specific attributes (e.g. perceptual detail, temporal information, associated or supporting information), as well as internal processes (e.g. feelings and thoughts). Consistent with the predictions of the RM framework, Johnson et al. (1988) found that real events received higher ratings of perceptual and contextual information than imagined events. This initial finding has been replicated and extended in many other studies (e.g. Henkel, Franklin, & Johnson, 2000; Johnson, Bush, & Mitchell, 1998; Mather, Henkel, & Johnson, 1997). A somewhat different strategy for using the MCQ as a diagnostic tool in eyewitness contexts underlies a method based on interpersonal reality monitoring (Johnson et al., 1998). In this method, the phenomenal characteristics associated with the witness’ memory report are extracted from the content of the report by independent trained evaluators or judges rather than being explicitly provided (rated) by the witness. Several studies have tested the diagnostic capacity of RM criteria derived in this manner in distinguishing accounts based on memory from those based on imagination (e.g. Barnier, Sharman, McKay, & Sporer, 2005; Hernandez-Fernaud & Alonso-Quecuty, 1997; Larsson & Granhag, 2005; see Masip, Sporer, Garrido, & Herrero, 2005; Roberts, Lamb, Zale, & Randall, 1998; Santilla,
Roppola, & Niemi, 1999; Sporer, 1997; Strömwall, Bengtsson, Leander, & Granhag, 2004). The RM approach has been operationalized and studied in different laboratories each with overlapping criteria but with differences in their approach. Some researchers have taken some of the CBCA criteria and included them as part of the RM (see Masip et al., 2005) while others have developed sophisticated rating scales and revised versions of the MCQ (see Sporer, 2004, 2008, for reviews).

Major limitations of the studies based on reality monitoring to date are that they have all used slightly different criteria to test the RM approach and many have failed to provide detailed examples of the criteria, sufficient details about how the coders were trained or inter-coder reliability. Moreover, the internal criteria ‘cognitive operations’ does not always appear to characterize accounts of non-experienced events (see Masip et al., 2005; Sporer, 2004). A few studies have reported more references to Cognitive Operations in true accounts (Blandon-Gitlin, Pezdek, Lindsay, & Hagen, in press, Expt 2; Granhag, Stromwall, & Olsson, 2001) and other studies report no differences (Sporer & Kuepper, 1995; Vrij, Edward, & Bull, 2001; Vrij, Edward, Roberts, & Bull, 2000). The definition of ‘cognitive operations’ appears to differ between researchers and this has contributed to the discrepancy in findings in the literature (see Masip et al., 2005; Sporer, 2004, for a similar point). This makes comparisons across different studies difficult. There is a need to establish and specify in detail a standard set of RM criteria. It is with this general aim that the current research was undertaken. In the current research, we compare two coding schemes, each of which presents a unique set of RM criteria and assess the effectiveness of each in assessing the veracity of eyewitness statements. The first is a scheme that closely mirrors Johnson and Raye’s (1981) original formulation of the RM Framework. The scheme is referred to here as Version 1 (Colwell, Hiscock-Anisman, & Memon, 2002; Colwell, Hiscock-Anisman, Memon, Rachel, & Colwell, 2007; Colwell, Hiscock-Anisman, Memon, Taylor, & Prewett, 2007). The second is taken from the contemporary research of Vrij, Akehurst, Soukara, and Bull (2004); Vrij, Mann, Kirsten, and Fisher (2007); Vrij, Mann, Kirsten, and Fisher (2008); and is referred to here as Version 2 in this paper.

The first scheme based on RM (Version 1, Colwell, Hiscock-Anisman, Memon, & Rachel et al., 2007; Colwell, Hiscock-Anisman, Memon, & Taylor et al., 2007) posits that honestly-reported memories derived from perceptual experiences will be longer and more detailed, as predicted by earlier studies in the deception literature (Cody & O’Hair, 1983; O’Hair, Cody, & McLaughlin, 1981; Rockwell, Buller, & Burgoon, 1997; Zuckerman, & Driver, 1985; see Vrij, 2000, 2005a, for reviews). Several studies using the RM approach have reported that truthful accounts are longer than deceptive ones (Colwell et al., 2002; Strömwall, Granhag, & Jonsson, 2003; Vrij, Akehurst, Soukara, & Bull, 2004) although occasionally studies report no differences in statement length as a function of veracity (e.g. Vrij, et al., 2007). Response length is therefore included as a predictor of veracity in Version 1 together with the different types of details elicited in true accounts. In Version 1, these include references to external details which are derived from senses (e.g. ‘I saw the black sports car with the shiny wheels’), contextual details which pertain to spatial and temporal relationships (e.g. ‘The car was parked next to the garage door’) and affective details or references to mood or emotional state (e.g. ‘I was surprised’). Within this scheme internally derived memories are hypothesized to contain more references to, cognitive operations defined as anything inside the respondent’s head that deals with previous idiosyncratic memory (or commentary) such as ‘that happened to me once’, and inferences whether be they about the self or others e.g. ‘I hadn’t thought about it till now’ (Johnson & Raye, 1981).
The presentation of idiosyncratic information allows for a deceiver to give the appearance of cooperation without actually providing information regarding the event in question. As such, Version 1 predicts an increase in references to internal details in deceptive accounts.

In contrast to the scheme used by Colwell et al. which specifies the more general (increase in response length and detail) as well as specific criteria associated with deception, Vrij et al. (2008) focus primarily on the qualities associated with experienced versus non-experienced events. According to Vrij et al. (2008) memories based on experienced events are likely to contain references to sensory information (details of smell, touch, taste, visual, and auditory details) contextual information, which include details about location as well as spatial and temporal details (e.g. details about the ordering of events). In contrast, non-experienced or invented accounts are hypothesized to contain more references to cognitive operations. The latter have been defined more broadly in recent work (Vrij et al., 2007, 2008) to include suppositions about sensory experiences (such as ‘She seemed quite clever’). Vrij (2008) points out that cognitive operations also refer to descriptions of inferences made by participants in their event descriptions such as ‘She looked smart’ (as in smartly dressed which an inference of visual details) and ‘She was wearing her coat so it must have been cold’ the latter could either be an inference of visual details but possibly also an inference about affect (she was feeling cold). Vrij et al. (2007) compared the RM with CBCA and a software programme that automatically codes RM criteria (Linguistic Inquiry and Word Count). In that study, not only did RM distinguish truth tellers from liars better than CBCA but manual RM coding also resulted in more verbal cues to deception than automatic coding. However, there were two criteria that did not discriminate between liars and truth-tellers in the Vrij et al. (2007) study. The broad CBCA criteria quantity of details did not discriminate (the length of truthful and deceptive statements did not differ). Moreover, the RM criteria, visual details did not yield significant differences between liars and truth-tellers. However, the latter was also the case in the Vrij et al. (2008) study and the authors suggest that because they had told liars what truth-tellers had experienced during the event, they could incorporate visual details into their stories. This is an important point because it suggests where a witness has been coached or provided with a detailed script the predicted differences between liars and truth tellers in visual details may not be found (see Vrij, 2005b). Interestingly, Strömwall and Granhag (2005) found a higher number of visual details in the accounts of children who had invented an account of a magic show while Strömwall et al. (2004) found no significant differences in visual details or cognitive operations in experienced and invented events.

The focus of this research was to test the effectiveness of RM in discriminating a truthful account of a staged event from an invented account of the same event. The study employs two coding schemes; Version 1 which draws upon the contemporary research of Colwell, Hiscock-Anisman, Memon, and Rachel et al. (2007), for an overview and version 2 as specified by Vrij in his recent work (Vrij et al., 2007, 2008). Both of these schemes closely mirror the original framework but employ slightly different definitions of the RM criteria. Data are presented using two versions of these criteria in an attempt to shed light on the hypothesis that internal details (Version 1) and cognitive operations (Version 2) are more likely to appear in deceptive accounts. One of the primary aims of this study was to address the discrepancy in the past literature as to the whether references to cognitive operations can be taken to be indicative of fabrication and deception. The RM truth criteria are operationalized by coding for external, contextual, and affective details in Version 1 (Colwell, Hiscock-Anisman,
Memon, & Rachel et al., 2007; Colwell, Hiscock-Anisman, Memon, & Taylor et al., 2007) and visual, spatial, auditory, and temporal details in Version 2 (Vrij et al., 2007, 2008).

The basic method for eliciting true and deceptive accounts from participants is through an initial instruction for participants to provide a detailed free recall either of an event they have experienced during the course of the study (staged event) or to follow an experimenter provided script. The instructions and procedure for creating a script closely followed the work of Vrij et al. (2004) namely the ‘story telling’ paradigm.

Method
Participants
Sixty college students comprised of 16 males and 44 females (mean age = 20.75, SD = 4.47) 28 in the truthful condition (6 males and 22 females) and 32 in the invented condition (10 males and 22 females) took part in the study. Participants were tested in groups of four and received a small honorarium or course credit for taking part.

Procedure
Participants were led to believe they were taking part in a study on health and mood. They had been told they would be required to fill in several questionnaires as part of the study. They were seated in a test room with the questionnaires. The test room contained three tables, two facing the walls and one in the centre of the room with a computer on it. The side tables were used by the participants while the experimenter used the centre table to rest her laptop and papers. Upon arrival at the laboratory participants completed several personality questionnaires such as the Freiburger Persönlichkeitsinventar (FPI; Fahrenberg, Hampel, & Selg, 2001) which was translated into English. However, since none of the measures yielded any correlations with our dependent measures, the questionnaires will not be discussed further in this paper. After about 5 min, the first experimenter (E1) left the room with an excuse (i.e. she told participants she needed more booklets), telling the participants that she was going to be back in a minute. Participants were asked to continue filling out the questionnaires. At this point, participants were randomly assigned to one of two conditions: truthful or invented.

Truthful condition
Approximately, 5 min after E1 had exited the room, participants in the ‘true’ condition experienced the following event. An actor playing the role of a computer technician entered the room. He asked about the whereabouts of the experimenter informing participants that there had been a report that the computer on the table in the centre of the room was not working and he had to check it. The technician explained that in order to reach the back of the computer, he had to move the table and asked the participants to help him. When, he had all the participants’ attention he picked up the laptop that was on the desk in order to shift the centre table but it slipped from his hands landing on the floor. The technician quickly picked it up off the floor and placed it on the table behind him. E1 re-entered the room and asked what the technician was doing. He explained that he was just there to fix the computer. The experimenter tried to switch the laptop on, but it would not work so she asked the technician if he had moved
the laptop, he mumbled something about having to move it off the table. The experimenter then became visibly angry and upset and blamed the technician because the laptop would not work. This resulted in a heated exchange between the experimenter and the technician. The technician left saying he will have to tell his boss what has happened. The experimenter apologized for the disruption and told participants that the laptop belonged to a colleague. The experimenter went out of the room informing the participants she needed to phone her colleague to ask for her help to turn the laptop on again (second experimenter, E2).

**Invented condition**

In the invented condition, the E1 left the room and returned 5 min later. There was no technician in the scenario for this condition. E1 tried to switch on the laptop as before and then called E2. Again, E1 told participants that she had to call her colleague (E2) in order to ask for her help to turn the laptop on again.

**Interview**

E2 came into the test room and requested that each participant come to her office for the next phase of the study. Participants were escorted one at a time by the second experimenter to a separate room to be interviewed about the laptop incident. Importantly, the experimenter was blind to experimental condition so they did not know if participants had actually witnessed the laptop being damaged or not. Upon arriving at the room, participants were seated some distance from the experimenter and handed a sealed envelope. In the truthful condition, the envelope contained a note explaining that the participants must provide a detailed and convincing account of what had happened. In the invented condition, participants received a bullet point description of the laptop incident (see Appendix) and a brief blurb telling them that the second experimenter did not know that they did not actually witness the event. They too were told they should provide a detailed and convincing account of what happened. Participants in each condition were given 5 min to think about what they were going to say.

Following earlier researchers (e.g. Hernandez-Fernaud & Alonso-Quecuty, 1997; Köhnken, Schimossek, Aschermann, & Höfer, 1995; Strömwall et al., 2004) a modified version of the CI was used to interview participants as this procedure typically results in a detailed recall (Köhnken, Milne, Memon, & Bull, 1999). Both interviewers received 2 days of intensive training in the CI which involved practice role-plays with mock interviewees and extensive feedback. The CI began with a short rapport building which allowed the interviewer to put the person at ease. This was followed by context reinstatement during which participants were asked to mentally construct a picture of the physical context of the event and any thoughts or feelings that they were having at the time. Following the context reinstatement instruction, participants were given a standard free recall instruction to provide a detailed uninterrupted description of what happened including all details pertaining to the event. Once participants had provided a free recall they were asked if they could tell the experimenter more with the following open-ended probes: What happened when the experimenter left the room? Describe the technician; tell me more about what he did. Participants were taken back to the testing room at the end of the interview and they completed a second anxiety measure prior to the end of the experiment. This was done to check that providing testimony (true or false) had not elevated anxiety levels.
Coding of interview transcripts (Version 1)

Four female graduate students received training in a system based on the RM scheme referred to as Version 1 here (based on Colwell, Hiscock-Anisman, Memon, & Rachel et al., 2007; Colwell, Hiscock-Anisman, Memon, & Taylor et al., 2007 and Johnson & Raye, 1981). The training was supervised by an expert in the field (K. C.) and the senior researcher on the project (A. M.). The coders were blind to the conditions and hypotheses of the experiment. The training session began with a general introduction to the background literature on statement validity analysis and reality monitoring. The raters were then provided with a detailed description each of the four coding criteria with examples of the definitions to ensure consistency. External details were defined as information that has been obtained from the senses (e.g. ‘the girl was wearing a pink jumper’, contains four external details). Contextual details were defined as information related to time and spatial relationships (e.g. ‘he put it on the table’, contains one contextual detail). Affective details were defined as emotions of the respondent at the time of the event (e.g. ‘I was really nervous’, contains one affective detail). Internal details were defined as thoughts, cognitive processes, or memory of any event other than the target event (e.g. ‘I didn’t pay attention because I was concentrating on the questionnaire’, contains one internal detail). Information was only scored once so repeated information was not scored. The four practice transcripts were taken from an earlier pilot study that had a similar method to the current experiment. The raters completed each transcript at home and then attended a meeting where they compared their results and feedback was given. During these meetings, each practice transcript was discussed, sentence by sentence, to ensure that the raters were coding the same features/words with the same criterion. Any disagreements were discussed and resolved. The raters received three further practice transcripts; these were compared and discussed as before. A high agreement was achieved for all the coding criteria following the last meeting therefore, the expert felt no more training was needed. The raters were then split into pairs and each rater was given 40-transcripts so that each transcript was coded twice in order to check for reliability among the four raters. Each rater coded each interview for the frequency of the RM criteria (internal, affective, external, and contextual details). Pairwise inter-rater reliability scores (Pearson’s correlations) were as follows: internal $r = .70$, affective $r = .75$, external $r = .67$, and contextual $r = .88$. Spearman correlations revealed a similar pattern: internal $r = .69$, affective $r = .82$, external $r = .60$, and contextual $r = .82$; all $p < .01$.

As is common practice in the literature (e.g. Vrij et al., 2008), the frequency scores were transformed using five-point scales to assess the presence of each RM criterion; 1 = absent and 5 = strongly present. These scales were created in a systematic manner by dividing the standard score range of each variable into five equal units. The mean of each scale was set to three. Standard scores that were more than one-fifth of the range above the mean were assigned a score of four, standard scores that were more than two-fifths of the range above the mean were assigned a standard score of five. A similar conversion was performed for standard scores less than the mean.

Version 2

This scheme was based on the recent work of Vrij et al (2007, 2008) as well as the original RM framework Johnson and Raye (1981). The training was given by the senior researcher working in the Vrij laboratory (S. Mann) and the senior researcher
on the current project (A. M.). The training session began with a general introduction to the background literature on statement validity analysis and reality monitoring. The raters were also provided with a detailed description each of the five coding criteria with examples and the definitions to ensure consistency. Visual details were defined as any action or thing that was seen in the event (e.g. ‘He walked into the room’, contains three visual details). Spatial details were defined as information relating to the location and positioning of items (e.g. ‘he put it on the table’, contains one spatial detail). Temporal details were defined as any information relating to the timing of events (e.g. ‘shortly after the man arrived’, contains one temporal detail). Auditory details were any speech or sound (e.g. ‘she asked my name’, contains one auditory detail). Cognitive operations were defined as suppositions, thoughts, reasoning and attributions of intent, and emotion (e.g. ‘he didn’t make a fuss when he dropped the laptop, he didn’t care, it wasn’t his business’, contains one cognitive operation). Note, in this coding scheme, references to emotion are coded as cognitive operations but are coded as affective in Version 1. Again, information was only scored once so repeated information was not scored. The four practice transcripts were taken from an earlier pilot study that had a similar method to the current experiment. The raters completed each transcript at home and then attended a meeting where they compared their results sentence by sentence and feedback was given. Any disagreements were discussed and resolved. The raters received three further practice transcripts; these were compared and discussed as before. Each rater coded each interview for the frequency of the RM criteria (cognitive operation, spatial, temporal, auditory, and visual details).\(^1\) Pair wise inter-rater reliability scores (Pearson’s correlations) were as follows: cognitive operations \(r = .86\), spatial \(r = .73\), temporal \(r = .90\), auditory \(r = .68\), and visual \(r = .91\); all \(p’s < .01\). Spearman correlations were as follows: cognitive operations \(r = .74\), spatial \(r = .60\), temporal \(r = .71\), auditory \(r = .73\), and visual \(r = .81\); all \(p’s < .01\). Again two of the raters transformed their frequency scores into 5-point scales to assess the presence of each RM criterion; 1 = absent and 5 = strongly present.\(^1\)

**Results**

The aim of this study was to examine, whether each of the individual RM criteria could be used to discriminate truthful from invented accounts. Descriptives (frequency counts) were obtained for each of the RM criteria for Versions 1 and 2 and these are presented in Tables 1 and 2, respectively.

**Version 1**

A MANOVA was carried out with the transformed RM scores for each individual criterion (external, contextual, affective, and internal details) with veracity (true or invented) as

---

\(^1\) Each rater also calculated a total RM score using the following equation: visual + spatial + temporal + auditory + cognitive operation (a similar calculation is used by Vrij et al., 2008). The total based on the transformed data was \(9.37 \text{ and } 7.08\) for the true and invented conditions respectively with a one-way ANOVA yielding a significant difference, \(F(1.59) = 10.05, p = .002\). However, given the unreliability of cognitive operations as an indicator of deception, and the fact that we had more references to cognitions in our true accounts, total RM scores are not used in our main analysis.
the independent variable. The results indicated a significant multivariate effect of veracity, $F(4, 55) = 3.73$, $p = .015$, $\eta^2 = .20$. Univariate tests revealed significantly more contextual details and external details in the true accounts. There were no significant differences in the number of affective or internal details (see Table 1 for the mean number of details, $F$ values, and effect sizes). When response length was controlled for in a MANCOVA, the difference in contextual details remained statistically significant, $F(1, 57) = 9.75$, $p = .01$. There were no differences in external details, $F(1, 57) = 1.59$, $p = .21$, or in affective or internal details ($Fs < 1$). As hypothesized, participants in the True condition produced longer responses (defined by number of words) than those in the Invented condition, $F(4, 54) = 19.82$, $p = .006$, $\eta^2 = .59$.

A discriminant function analysis was used to classify statements as true or invented on the basis of the individual RM criteria (external, contextual, affective, and internal details) together with response length. The discriminant analysis showed a significant effect with $\lambda(5) = .80$, $p = .031$. To address the problem of inflated prediction rates arising from individual error-specific variance, the leave one out method was used. Cross-validated groups cases were correctly classified at 57 and 75% for the true and invented accounts, respectively. The Structure coefficients (effect sizes) for each variable were: .98 (contextual); .58 (external); .28 (affective); .32 (internal), and .49 (response length).

### Table 1.

<table>
<thead>
<tr>
<th></th>
<th>True</th>
<th>Invented</th>
<th>$M$</th>
<th>SD</th>
<th>$M$</th>
<th>SD</th>
<th>$F$</th>
<th>$P$</th>
<th>$r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal</td>
<td>26.82</td>
<td>18.76</td>
<td>20</td>
<td>12.45</td>
<td>1.53</td>
<td>.22</td>
<td>.16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Affective</td>
<td>1.82</td>
<td>2.14</td>
<td>1.64</td>
<td>2.46</td>
<td>1.16</td>
<td>.28</td>
<td>.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>External</td>
<td>160.70</td>
<td>49.42</td>
<td>136.32</td>
<td>44.40</td>
<td>4.90</td>
<td>.03</td>
<td>.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contextual</td>
<td>75.30</td>
<td>33.05</td>
<td>52.50</td>
<td>16.99</td>
<td>13.98</td>
<td>.01</td>
<td>.43</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A MANOVA was carried out with the transformed RM scores as per Version 1. The results indicated a significant multivariate effect of veracity, $F(5, 54) = 4.30$, $p = .002$, $\eta^2 = .285$. Univariate tests revealed effects for temporal and auditory details with

---

2 The data were screened by obtaining Mahalanobis distance from External, Contextual, Internal details, and Response Length. This distance was graphed by case number. No multivariate outliers were evident, so the data were analyzed without dropping any cases. Here, we report Pillai's criteria which is the most robust to violations of assumptions concerning homogeneity of the covariance matrix (Olson, 1976). However, note that the $F$ ratios and $p$ values computed by the four types of MANOVA commonly reported in SPSS were exactly the same. This indicates that the data were not significantly affected by violations of the assumptions of MANOVA. If significant violations had been present, then the $F$ ratio and $p$ values for Pillai's Trace would diverge from the others (Field, 2005). MANOVA and ANOVA are both very robust regarding violations of normality because they are based upon the sampling distribution of the mean, rather than upon the original distribution of the data. The distribution of the mean is only slightly affected by non-normality, and these effects further diminish with increasing case numbers (Grim & Yarnold, 2000). Differences in response length were controlled for by including number of words as a covariate in the analysis.

3 Note that the discriminant function analysis was performed on the frequency scores and transformed scores with no differences in the results. Response length is included as a predictor in Version 1 for the reasons outlined in the introduction.
more of each type of detail in the True accounts. There were no differences in spatial details or visual details as a function of veracity. Contrary to prediction, there were more cognitive operations in the true accounts (see Table 2 for means, \(F\) values and effect sizes). Controlling for response length, the differences in cognitive operations fell short of statistical significance, \(F(1, 57) = 3.07, p = .08\). As before there were more references to auditory, \(F(1, 57) = 6.28, p = .01\) and temporal details, \(F(1, 57) = 15.30, p = .001\) in the true accounts but no significant differences in visual and spatial details (\(Fs < 1\)) as a function of veracity. Again, participants in the true condition produced longer responses than those in the invented condition, \(F(5, 53) = 12.69, p = .001, \eta^2 = .54\).

A discriminant function analysis was used to classify statements as honest or deceptive on the basis of each of the RM criteria in Version 2. The analysis revealed a significant effect with \(\lambda(5) = .71, p = .002\). Correct classification accuracy (cross-validated) was 61\% for true cases and 78\% for invented accounts. The Structure Coefficients for each variable were: .66 (auditory); .92 (temporal); .29 (visual); .26 (spatial); and .52 (cognitive operation).

### Table 2. Mean number of details (frequency counts) for each individual RM criteria for Version 1 in the truthful and invented conditions and univariate ANOVA results (\(df\)s = 1, 58) effect sizes (r)

<table>
<thead>
<tr>
<th></th>
<th>True</th>
<th></th>
<th>Invented</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Visual</td>
<td>94.88</td>
<td>28.04</td>
<td>82.16</td>
<td>26.49</td>
</tr>
<tr>
<td>Auditory</td>
<td>18.13</td>
<td>8.50</td>
<td>11.46</td>
<td>5.14</td>
</tr>
<tr>
<td>Spatial</td>
<td>17.33</td>
<td>5.57</td>
<td>15.52</td>
<td>5.79</td>
</tr>
<tr>
<td>Temporal</td>
<td>26.40</td>
<td>10.45</td>
<td>17.25</td>
<td>8.16</td>
</tr>
<tr>
<td>Cognitive</td>
<td>15.40</td>
<td>7.90</td>
<td>11.52</td>
<td>8.44</td>
</tr>
</tbody>
</table>

**Comparison of Versions 1 and 2**

Table 3 presents a full correlation matrix of all RM measures and length of report. Of the 110 pairwise correlations, 94 were significant (\(p < .01\)) indicating substantial overlap between the two sets of criteria and the individual criteria. As one would expect with two systems derived from the same basis, there was substantial overlap among the two versions of RM. The external detail category was collinear with the visual detail category. Similarly, the contextual detail category was collinear with the temporal detail category. Finally, the internal detail category was collinear with the cognitive operations detail category. This collinearity indicates that these categories are describing almost the same things. Also, response length was highly correlated or collinear with multiple detail categories, indicating that the number of words in a statement gives a strong indicator of the amount of unique and useful information in that statement.

This multi-collinearity across the two systems is good in that it indicates that the two systems each assess much the same thing. However, this multi-collinearity also precludes putting both of these systems in the same multivariate statistical technique for comparison of the overall amount of variance described by each system. Therefore, in order to assess whether one system was superior to the other in its ability to describe the observed differences due to veracity, the multiple correlation derived from each
system was converted into a $z$ score. This $z$ score conversion allowed for direct comparison of the two techniques, to determine whether one better explained the effect of veracity. This yielded a non-significant result ($z_{\text{Diff}} = 0.471, \ p > .05$). This indicates that the two schemes are approximately equal in their potential to describe differences in statements due to the effect of veracity despite subtle but important differences in the definition of the internal (Version 1) or cognitive (Version 2) criteria. Each of the schemes suffers from similar shortcomings in their applicability and these will now discussed in the next section.

**Discussion**

The present study was conducted to examine the potential of a small set of RM criteria in aiding the discrimination of true from invented accounts provided by eyewitnesses. For Version 1, external and contextual details as specified in the RM scheme adopted by Colwell and colleagues and derived from the original Johnson and Raye (1981) framework, were more common in truthful accounts. However, internal criteria did not differ as a function of veracity in Version 1. For Version 2, based on the contemporary research of Vrij and colleagues our data indicated that two of the RM criteria namely auditory and temporal details appeared more frequently in true as compared to invented accounts while there were no significant differences in visual or spatial details. Unlike Vrij et al. (2007, 2008) we did not find more references to cognitive operations in invented accounts. On the contrary, there were more references to cognitive operations in the truthful accounts although the differences were not statistically significant when response length was controlled for. The differences in the cognitive criteria aside, the Colwell and Vrij schemes were comparable and the pairwise correlations reported in Table 3 indicate there is substantial overlap between the two sets of criteria. The correct classification accuracy rates for Version 1 and 2 were 57 and 61%, respectively, for true accounts and 75 and 78%, respectively, for deceptive accounts. The rates are somewhat lower than some of the earlier studies. In a review of 10 studies using either RM criteria or a RM summative score as predictor variables, Sporer (2004) reports classification rates (based on multiple discriminant analyses) ranged from 61 to 91% for truthful accounts and 61–81% for deceptive accounts. As pointed out by Sporer (2004), these rates are comparable to those obtained with the CBCA approach. However, what do these classification rates really tell us about the usefulness of the RM approach in view of the fact that all the criteria (with the exception of affective details in Version 1 and visual and spatial details in Version 2) were essentially ‘truth’ criteria? In order to address this question, it is essential to evaluate critically the methodology used in laboratory studies that have tested the efficacy of the RM criteria and to consider to what extent the differences are a function of the nature of the event, the type of situation being studied (e.g. creating a false account of an experienced event or creating an entirely fictitious event with or without intent to deceive) as well as the way in which the criteria was measured.

As indicated in the introduction, a lack of standardization in the RM criteria has made it difficult to compare findings across different set of studies. Most problematic is the single criteria that is deemed to be indicative of deception namely cognitive operations. In the original RM framework, cognitive operations referred to internal memories based on reasoning and thought processes (Johnson & Raye, 1981). More recently, the definition has been broadened to include subjective references concerning visual details as in the work of Vrij et al. (2008) and also confounded with references to emotions and
Table 3. Correlations of Versions 1 and 2 RM criteria and response length (N = 60)

<table>
<thead>
<tr>
<th>V &amp; L</th>
<th>W &amp; V</th>
<th>T &amp; V</th>
<th>S &amp; V</th>
<th>A &amp; V</th>
<th>C &amp; V</th>
<th>I &amp; V</th>
<th>A &amp; E</th>
<th>E &amp; C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

* Correlation is significant at the .05 level (two-tailed).
** Correlation is significant at the .01 level (two-tailed).
*** Correlation is significant at the .001 level (two-tailed).
feelings (Colwell, Hiscock-Anisman, Memon, & Rachel et al., 2007). The two schemes, we compare in this study differ on a number of factors. Focusing on cognitive operations, the definition adopted in Version 2 is very broad and includes thoughts, reasoning processes, attributions of intent, emotion, inferences of others, and suppositions of visual/sensory information, e.g. ‘she was tall’. Version 1, however, would categorize emotions as an affective detail and inferences of visual details such as ‘he looked about 21’ as an external detail. Furthermore, the definition of internal details in Version 1 is also difficult to pin down including references to an individual’s thoughts processes and meta-cognitive judgements about self and others (I don’t remember, I think so, That happened to me once, and I guess he must be some guy who works here). Note there was a tendency for more internal details to be reported in the true accounts here but considerably variability among participants. As argued by Sporer (2004) idiosyncratic information and cognitive operations also reflect supporting memories and rehearsal processes and should be considered as memory cues based on self-experienced events rather than indicators of invented or deceptive accounts. This may explain why we found more references to cognitive operations for true accounts with Version 2 and previous studies have obtained inconsistent results when this criterion is applied.

Contemporary researchers have expressed concerns with the operationalization of the cognitive operation criterion as an indicator of deception (Sporer, 2004; Granhag, personal communication, July 2008). Sporer (2008) has made the point that laboratory procedures, such as the ‘story telling paradigm’ used here, have resulted in a measurement of references to cognitive operations at time of retrieval or memory report. This conflicts with the original theory of RM, which refers to internal processes at the time of encoding. Another methodological issue also highlighted by Sporer (2004) is that many studies use video events as the experienced event and if the participants had not experienced the event first-hand themselves then they are unlikely to refer to their thoughts and feelings or indeed their sensory experiences. Instead, they are more likely to make reference to the thoughts and the perceived emotional reactions of the characters in the video or make inferences about the behaviour of actors. In the current study, the participants actually experienced the event but made few references to the way that they felt throughout the incident. This may be why the affect criteria (Version 1) did not discriminate between truthful and invented accounts. However, participants did make inferences about the affect of the technician and visual inferences about the technician, experimenter, and the other participants. These would have been coded as cognitive operations under Version 2, which might explain why there were more references to cognitive operations in the true accounts. Another factor that may account for more references to cognitive operations is that the CI instruction to report everything and reinstate context could have encouraged participants who had experienced the event to make more inferences about visual details, sensory experiences, and their impressions of people. After all, it is not usual for the CI to increase the reporting of suppositions or speculations (e.g. Stein & Memon, 2006).

It has recently been suggested that cognitive load induced by certain types of questioning could be an important variable. Vrij et al. (2008) noted that cognitive operations were more common in untruthful accounts under conditions of cognitive load (reverse order recall component of the CI). It is possible that the different components of the CI vary in how much cognitive load they exert on a participant. Future studies should examine more closely how questioning technique and cognitive load could influence what witnesses report (see Colwell, Hiscock-Anisman,
Memon, & Rachel et al., 2007; Colwell, Hiscock-Anisman, Memon, & Taylor et al., 2007, 2008; Vrij et al., 2007).

Several other factors may moderate the efficacy of the RM approach in distinguishing veracity. References to perceptual and affective details will be influenced by the nature of the staged incident, the nature of script that the participant is provided with in the deception condition, motivation, and undoubtedly the amount of time a participant has to prepare the statement (see Alonso-Quecuty, 1992). Thus as pointed out by Vrij (2008) if a witness is provided with a detailed script in the deceptive condition, then they are likely to elicit a large number of visual details. The latter could account for why the criteria ‘visual details’ is not always useful (see also Vrij et al., 2007). In this regard, the storytelling paradigm could have the same effect as coaching and simply serve to increase the presence of all RM criteria hence reducing discriminability of scores (Vrij, Akehurst, Soukara, & Bull, 2002). A related point is that because false accounts are partly constructed by combining information from true events, when participants create full (as opposed to partial) false accounts, the discriminative power of RM is reduced. This hypothesis is based on recent research using CBCA for discriminating between accounts of true events and suggested events (Blandon-Gitlin et al., 2008).

As for differences in response length in the truthful and invented accounts, this is not surprising assuming that honest people reconstruct their memory during the interview while liars just repeat their lie script (Granhaq & Strömwall, 1999; Granhaq, Strömwall, & Jonsson, 2003). Similarly, Colwell, Hiscock-Anisman, Memon, and Taylor et al. (2007) have argued that deceivers deliberately provide shorter simpler statements in order to reduce the chances that they will be caught. This argument is supported by information manipulation theory (McCornack, Levine, Solowczuk, & Torres, 1992), which holds that deceivers must balance the need to produce enough information to satisfy the interviewer with the need to control and hide any potentially incriminating facts. To do this, deceivers mentally prepare and rehearse a lie script, and they base their answers during interviewing on this lie script rather than episodic memory (Porter & Yuille, 1996; Colwell et al., 2002). This leads to deception being shorter, more scripted, and more carefully phrased than honest reporting. It should, however, be pointed out that the deceptive accounts are not always longer than truthful accounts (see Sporer & Schwandt, 2006) and this may be because numerous factors can inflate response length such as the whether or not the witness has been coached, the ‘lie’ script that a witness has been provided, a witness’s emotional state and motivation to conceal a lie. With respect to the last point, it should be noted that in the study reported here participants were simply asked to make up an account of an experience they had not witnessed. We were not asking our participants to hide any information. It may be that when we ask participants to add false details or to deliberately mislead an interviewer about an experienced event participants exert more control over the specific details they provide. These factors have not as yet been sufficiently addressed in contemporary research on reality monitoring and deception.

In terms of the forensic potential of the two coding schemes, we have studied here, no one scheme is currently suitable for application in a forensic setting. Investigators may find it beneficial early on in an investigation if they can assess the veracity of a statement. However, we would caution against this unless investigators are fully aware of the circumstances under which false positives and false negatives could occur when assessing statement veracity. For example, some types of interrogative practices can result in a false confession from an innocent suspect (see Lassiter, 2004; Hill & Memon, 2007, for a review of the relevant literature). Such practices are not uncommon in police
interviews of suspects (Soukara, Bull, Vrij, Turner, & Cherryman, in press). Even in a witness interview, the style of questioning can itself influence what a witness reports. Therefore analysis of statements should only be undertaken where there is detailed information available on the questioning style and the questions asked. However, herein lies another potential problem which concerns perceptions of various questioning practices. For example, the use of the CI and other investigative practices is not widespread (Wells, Memon, & Penrod, 2006) and police officers may not feel they are adequately trained or sufficiently equipped to conduct a CI (Dando, Wilcock, & Milne, in press) let alone try and use that interview to assess statement veracity! Ideally, future research, and practice will address the development of careful interviewing to facilitate the detection of deception with RM criteria clearly defined so that they are sensitive to control of information and impression management as well as differences due to memory.

**Acknowledgement**

This research was supported by a European Framework six grant to the first author. We would like to thank the following people for assistance with coding: Fraz Chaudhry, Nikola Bergis, Julie Gawrylowicz, Lorina Hornstein, Susan Dixon, and Samantha Mann. We would also like to thank Aldert Vrij and Sigi Sporer for comments on an earlier draft of this manuscript.

**References**


Appendix: Script used in Experiments 1 and 2

You arrive for your experiment.
The experimenter asks your name and you fill in the consent form.
You start the first questionnaire.
The experimenter leaves the room to get more photocopies of the questionnaire for
the next experiment.
A technician knocks on the door asking for the experimenter.
The technician says that he is here to fix the computer.
The technician asks you to help him move the table that the computer is on.
The technician picks up a laptop from the table and drops it.
The technician picks the laptop up from the floor and puts it on a table behind him.
The technician continues to fix the computer.
The experimenter returns and asks what is happening.
She asks you to continue with the questionnaires.
The experimenter tries to switch on the laptop but it does not work.
The experimenter asks the technician if he has moved the laptop.
There is a heated argument between Experimenter and technician.
The experimenter tells you that the laptop is broken and that she has to tell the
owner.
The owner of the laptop comes into the room and says that she will have to have an
account of what happened so that she can get her laptop fixed by the insurance
company.