The Role of Account Length in Detecting Deception in Written and Orally Produced Autobiographical Accounts using Reality Monitoring

Reality monitoring lie-detection studies, like others that use raw frequency counts as primary data, seem consistently to underestimate the influence of the length of (or number of words in) the account. The decisions as to whether to standardise or not, or what method of standardisation to use, are rarely empirically driven, so it is still unclear as to whether reality monitoring is more effective before or after standardisation for length. Another factor that also has received little attention in the reality monitoring literature is whether statements are produced orally or in written form. To investigate these issues, 42 autobiographical statements, 21 truthful, and 21 deceptive, including 22 oral and 20 written accounts, were analysed before and after word count standardisation. Results showed that reality monitoring criteria only discriminated significantly between truthful and deceptive accounts when no attempt to control for word count was made. Also, oral statements contained more evidence of reality monitoring criteria before standardisation for word count, whereas written statements were denser and contained more evidence of reality monitoring criteria after standardisation. Implications are discussed.

A variety of techniques have been developed by investigators to assess the veracity of statements of suspects and alleged victims in forensic investigations. However, many of these have not survived academic scrutiny or have not been tested with sufficient rigour to be considered reliable; these include Scientific Content Analysis, Investigative Resource Analysis, Verbal Behaviour Analysis, and Lexical Diversity (for reviews see Adams & Jarvis, 2006; Vrij, 2008). Nevertheless, perhaps the best known, more reliable, and most extensively tested techniques of verbal lie detection are criterion-based content analysis (CBCA) and reality monitoring (RM)(Granhag, Strömwall & Landström, 2006; Masip, Sporer, Garrido & Herrero, 2005; Vrij, Edward, Roberts, & Bull, 2000; Vrij, Mann, Kristen, & Fisher, 2007).

At present, the discrimination rates for both CBCA and RM have been considered too low for full integration into the criminal justice system. Hence, as yet, RM has not been used in forensic investigations, and whereas CBCA has been recognised only in criminal court proceedings in Sweden, Germany, the Netherlands, and Switzerland, its application has been limited to testing accounts derived from children-witnesses/victims of sexual abuse (Vrij, 2008).

Nevertheless, comparisons of CBCA and RM techniques have shown a number of advantages of the RM approach (Masip *et al*., 2005; Sporer, 1997; Vrij *et al*., 2000). For example, RM is less complex and potentially more cost-effective to use, and a number of studies have shown that it discriminates better between truthful and deceptive accounts (Granhag, Strömwall, & Landström, 2006; Porter & Yuille, 1996; Sporer, 1997; Strömwall, Bengtsson, Leander, & Granhag, 2004; Vrij, 2000). Perhaps one of the most important reasons why RM might have more discriminatory power is that it is recognised as having a stronger theoretical basis (Masip *et al*., 2005). This happens primarily because the principles of CBCA were derived from child-interviewers' practical experience, the RM framework has its roots in memory theory.

The theoretical basis for RM can be found in Johnson and Ray's (1981) ideas concerning differences between memories of real events and memories of imagined experiences. According to this perspective, memories of real events are obtained through perceptual processes; therefore, they are more likely to contain perceptual details (e.g. sounds, colours, and details of smell), spatiotemporal details (e.g. details regarding the spatial arrangement of people or the time order of events), and affective details (details regarding information about feelings) than imagined or fabricated memories (Sporer, 2004; Vrij, 2000). In contrast, the memories of imagined events are internally derived; hence, they are less likely to contain the aforementioned details and are more likely to contain information regarding cognitive operations (e.g. thoughts, reasoning, and cognitive suppositions of sensory experiences). As noted previously, compared with other verbal statement lie-detection approaches, such as CBCA, RM is considered by many to be relatively easy to use as it has fewer criteria and more clear-cut distinctions (Sporer, 1997; Vrij, 2000); it is thus potentially more cost-effective in terms of interviewer or coder training.

However, 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. Although 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, Akehurst, Soukara & Bull, 2004; Vrij *et al*., 2000), particular problems arise when frequency data are used to measure more discrete 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. This is a particular problem for RM as word count has been found to positively correlate with the entire gamut of RM criteria such that, the longer the accounts, the stronger the presence of the various RM criteria within them

(Memon, Fraser, Colwell, Odinot, & Mastroberardino, 2010). For example, it has been found that criteria such as visual information and cognitive information are more effective in discriminating between truthful and deceptive accounts before standardisation for word count, than after standardisation (Larson & Granhag, 2005; Masip *et al*., 2005).

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, Vrij, & Fisher, 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 *et al*., 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, Strömwall, Landström, 2006; Strömwall *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 (e.g. 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, Hiscock-Anisman, & Memon, 2002; Memon *et al*., 2010; Vrij *et al*., 2004). Indeed, it could be construed as entirely missing the point in terms of the rationale behind RM. RM was originally formulated on the idea that 'memories based in perception have better spatial, temporal, and sensory information'

(Johnson & Raye, 1981, p.82). There is no reference in the seminal RM papers by Johnson and colleagues (Johnson, Hastroudi, & Lindsay, 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, Hastroudi, & Lindsay, 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, little systematic research has been conducted on this issue. In particular, we need to know exactly how the standardisation of accounts for number of words, or length, affects the role of RM criteria (both singularly and in combination) in discriminating lies from truths.

Another potentially under-researched factor that may affect the efficacy of RM in discriminating truths from lies is the mode in which the account is presented, that is whether it is spoken or written. Although some RM studies have used both written statements and oral accounts in their designs (e.g. Granhag *et al*., 2006; Manzanero & Diges, 1996), in these, no attempt was made to compare the effects of the two. Also, results from the few RM studies that have used written statements have presented inconsistent findings. For example, Sporer and Sharman (2006) reported that only realism and temporal information differentiated between truthful and deceptive accounts, whilst Barnier, Sharman, McKay, and Sporer (2005) found that truthful accounts were clearer and contained more affective information than deceptive ones. Moreover, in direct opposition to the predictions of RM theory, Barnier *et al*. (2005) also found that written truthful accounts contained more information regarding cognitive operations than deceptive accounts.

Nevertheless, there are a number of ways in which oral and written statements may differ that may be relevant to lie detection (Beaugrande, 1984; Kroll, 1977). For example, the role of a writer can be construed as very different from that of a speaker. The speaker usually quickly processes ideas into words whilst both utilising and receiving (from the listener) non-verbal, prosodic, and paralinguistic cues, which help the message evolve. The speaker, however, cannot easily look back or re-examine at what he/she has stated; that is, it is difficult for him/her to process the message holistically. The writer, in contrast, normally has more opportunity to look back and make detailed corrections to produce a more thematically

coherent message; however, he/she is not generally in a position to receive instant feedback in any form (Gumperz, Kaltman, & O'Connor, 1984). Also, oral accounts tend to be longer as speaking is much faster than writing (Chafe & Tannen, 1987; Halliday, 1989; Hidi & Hildyard, 1983; Tagg, 2009; Tannen, 1985). However, although the effects of the modality of the message and the medium used to communicate this message have been broadly assessed in the lie-detection context (e.g. Adams & Jarvis, 2006; Carlson & George, 2004; Davis, Markus, & Walters, 2006; Lindholm, 2008), as previously noted, the effects for spoken language have not systematically been compared with those for written language in the RM literature. So, both written transcripts and transcripts based on oral testimonies have been used as if they were functionally identical.

When considering the effects of oral and written accounts in relation to RM criteria, one might expect spoken accounts to display more information relevant to the RM criteria; that is 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, with standardisation of 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. Although 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 and 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). In contrast, 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 this paper is 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; that is 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.

Method

Participants

The participants were an opportunity sample consisting of 21 members of the general public and University of Liverpool students (mean age = 25.80; SD = 7.05); there were eight men and 13 women. There were no exclusion criteria other than participants had to be older than 17 years. All participants volunteered in response to an advertisement posted in the University website. No reimbursement was offered. This research was conducted in accordance with BPS and APA research ethics guidelines and approved by the University of Liverpool Institute of Psychology Health and Society Research Ethics Committee.

Materials

The measures were as follows.

Life experiences inventory

As autobiographical memories have been used as stimulus materials in numerous lie-detection studies (e.g. Ball & O'Callaghan, 2008; Barnier *et al*., 2005; Johnson, Foley, Suengas, & Raye, 1988; Masip *et al*., 2005; Sporer & Sharman, 2006), they were used as stimuli here. An adaptation of the kind of life experiences inventory (LEI) used by Garry, Manning, Loftus, and Sherman (1996) and Paddock *et al*. (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, and 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, describe events taken from books or films, and describe 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’.

Reality monitoring criteria

For RM coding, an RM framework was devised. This consisted of a list of five RM criteria (perceptual information, cognitive operations, temporal information, spatial information, and affective information) and a set of descriptions of their definitions derived from Sporer (2004) and Vrij (2000). The protocol required coders to score numerically the number of occasions where perceptual information, cognitive operations, temporal information, spatial information, and affective information was present. This resulted in a score for each criterion and a total criterion score. The criteria (with examples) were defined to the coders as follows.

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. 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’).
3. Remembered feelings (affect): how well the person remembers feelings (accounts of subjective mental states) from the event (e.g. ‘Joseph was very scared’).
4. 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’).
5. 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’).

Procedure

Participants 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.

Obviously, a problem with the use of materials of this kind is the establishment of ground truth for the 'truthful' accounts. An obvious way round the problem of ground truth would be to expose participants to events controlled by the experimenter. However, this inevitably means the event will have little emotional significance to the individuals concerned and the design will have poor ecological validity. Also, unless participants were being deliberately disruptive and uncooperative, it seems unlikely that they would simply manufacture 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. Consequently, the approach used here has been popular amongst lie-detection researchers (see, e.g. Barnier *et al*., 2005; Johnson *et al*., 1988; Santtila, Roppola, & Niemi, 1999; Sporer & Sharman, 2006) and has received support in perhaps the most complete meta-analytic study of RM research (Masip *et al*., 2005); the main advantage being that participants have some emotional engagement with the experimental process.

Within the constraints of the sample size, deceptive and truthful conditions were also counterbalanced within conditions, so 42 accounts were ultimately recorded and used within a mixed 2 × 2 (modality: written accounts versus oral accounts × truthfulness: real event versus fabricated event) design. Oral accounts were audio recorded and transcribed into written form, and all accounts were timed.

The 42 autobiographical accounts, consisting of 22 oral and 20 written accounts, were then scored by two lie-detection researchers using the RM framework previously described. Both were researchers in the area of lie-detection and familiar with the mechanics and theoretical underpinnings of the RM approach. They were also given an opportunity to familiarise themselves with the RM coding protocol before beginning the scoring process. Both coders were blind as to the truth status of the accounts, or whether they were oral or written, although verbal fillers (e.g. um and uh) were present only in the oral transcripts. These fillers were not removed from the transcripts; however, the coders were not aware of their purpose and function or that they were confined to oral testimonies. 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, there was high and significant inter-coder agreement between the two judges (ICC = 0.90−0.96; r = 0.84−0.96).

Results

Preliminary analysis of accounts

Preliminary analyses using 2 × 2 (modality: written accounts versus oral accounts × truthfulness: real event versus fabricated event) mixed ANOVAs with repeated measures on the second factor 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); $F(1, 19) = 6.94$, $p = 0.016$; $\eta^2_p = 0.27$. 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); $F(1,19) = 7.54$, $p = 0.013$; $\eta^2_p = 0.28$. 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); $F(1, 19) = 8.30$, $p = 0.010$; $\eta^2_p = 0.30$. 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 the written accounts (M = 193.04, SD = 97.79); $F(1, 19) = 12.90$, $p = 0.002$, $\eta^2_{p.} = 0.40$. 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); $F(1, 19) = 13.54$, $p = 0.002$, $\eta^2_p = 0.42$. Speakers' rate of word production was thus higher, producing significantly more words per second (M = 2.99, SD = 0.42) than writers (M = 0.42, SD = 0.10); $F(1, 19) = 395.04$, $p = 0.001$, $\eta^2_p = 0.95$. There were no significant interaction effects.

Reality monitoring results before word count standardisation

The RM frequencies before word standardisation were then analysed using a series of six, 2 × 2 (modality: written accounts versus oral accounts × truthfulness: real event versus fabricated event) 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, mean scores were higher for the truthful accounts for all RM criteria with the exception of cognitive operations. However, significant effects were found only for Total RM

scores $F(1,19) = 18.05$, $p = 0.001$, spatial information $F(1,19) = 17.79$, $p = 0.001$, and temporal information $F(1,19) = 8.32$, $p = 0.010$ (Table 1).

Table 1. RM mean (SD) frequency counts as a function of truthfulness before and after standardisation of word count

- RM, reality monitoring; SD, standard deviation.
- * $p < 0.05$;
- ** $p < 0.01$.

Although mean scores for all of the criteria were higher for oral accounts than written accounts, a significant main effect for modality was only found for cognitive information $F(1,19) = 8.38$, $p = 0.009$, although there were near significant trends for temporal information $F(1,19) = 4.22$, $p = 0.054$, and spatial information $F(1,19) = 3.53$, $p = 0.076$ (Table 2). No interactions between truthfulness and modality were found for any of the analyses.

Table 2. RM mean (SD) frequency counts as a function of modality before and after standardisation of word count

- RM, reality monitoring; SD, standard deviation.
- * $p < 0.01$.

Results after word count standardisation

To standardise word count, the RM raw scores were re-calculated per 100 words of account; that is raw frequencies per 100 words (e.g. of this method, see Larson & Granhag, 2005; Strömwall & Granhag, 2005; Vrij *et al*., 2004).

The frequencies after standardisation were again analysed using a series of six, 2 × 2 (modality: written accounts versus oral accounts × truthfulness: real event versus fabricated event) 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 (Table 1).

However, main effects for modality were found for the perceptual information $F(1,19) = 23.19$, $p = 0.001$; spatial information $F(1,19) = 9.69$, $p = 0.006$;

cognitive information $F(1,19) = 7.41$, $p = 0.014$, temporal information $F(1,19) = 12.12$, $p = 0.002$ and Total RM scores $F(1,19) = 59.09$, $p = 0.001$ (Table 2). In each case, written accounts were richer in RM criteria. Again, no interactions between truthfulness and modality were found for any of the analyses.

Sample-size considerations

Because of the small sample sizes a power analysis was conducted to ensure that the non-significant effects were not simply a feature of sample size. Results showed that to achieve a power value greater than 0.80, the non-significant effect with the largest effect size would need a sample size in excess of 2,500 to be significant at $p < 0.05$. We would argue that if the appropriate level of statistical power can only be achieved with a sample size of this magnitude, the results would effectively have no practical relevance.

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, although significant individual criterion effects were found only for spatial information and temporal information, total RM scores 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 on the prediction of veracity were not moderated by modality. In general, therefore, the present results appear to lend some support to previous findings suggesting that some RM criteria are more able to discriminate between truthful and deceptive accounts if there is no attempt to control for word count (e.g. Masip *et al*., 2005; Larson & 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 use a standardised measure of the raw frequencies (e.g. 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 would be little point standardising scores if to do so would mean rendering RM criteria relatively ineffective in predicting veracity.

Another obvious issue to consider is that, 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? Would it not be simpler just to use word count to predict whether accounts are more likely to be truthful? A brief examination of the relevant effect sizes in the present data suggests otherwise. For example, the effect size (Cohen's *d*) for the difference between truthful and deceptive accounts using total RM scores is 0.62; in contrast, those for word count per se and time spent are only 0.30 and 0.25, respectively. This suggests that the RM criteria may potentially outperform word count per se in predicting veracity. It can also be noted that, notwithstanding the finding that, other things being equal, truthful accounts tend to be shorter than deceptive ones (Zhou, Burgoon, Nunamaker, & Twitchell, 2004), 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, Blandón-Gitlin, Masip, & Sporer, 2012).This finding is in line with research using the Linguistic Inquiry Word Count (LIWC) computer software, which also shows that length 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 length needs to be considered in conjunction with other cues and with reference to the conditions under which the study has been conducted.

Importantly, the 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 veracity 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 word count and word count standardisation on predicting veracity, there was some support for the hypothesis that, regardless of whether accounts are truthful or deceptive, oral accounts tend to be longer and, therefore, produce higher RM 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). Also as predicted, however, the 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; that is written accounts were denser in terms of temporal, spatial, perceptual, and total RM scores. This seems to be in line with the rationale provided earlier arising from the different roles of speakers and writers (Chafe & Tannen, 1987; Halliday, 1989, 2001; Pu, 2006), 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, e.g. Granhag, Strömwall, & Landström, 2006; Manzanero & Diges, 1996). Moreover, if written accounts overall tend to be denser 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 cognition information to predict veracity, either before or after word count standardisation, this again emphasises the importance of not assuming that written and oral accounts 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. This may be particularly significant in that the cognitive information criterion is often considered the weakest RM criterion for predicting veracity (Granhag *et al*., 2006; Masip *et al*., 2005; Vrij *et al*., 2000; Vrij, 2008), and there has been some scepticism surrounding its 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 been often been found to be higher in truthful accounts than in the deceptive ones before word count standardisation.

To conclude, obviously the present study was limited in many respects, and the present findings must be interpreted with considerable caution. The most obvious limitations were the very small sample size and the lack of control over ground truth. Moreover, even though the LEI procedure used here could be construed as more ecologically valid than a laboratory manipulation involving presentation of some kind of scenario of low emotional salience, the motivation for participants to lie was arguably lower than would 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). A major challenge for future research on this topic, therefore, is not only to replicate the findings on a larger sample but also to use materials that are more directly relevant to the forensic context.

However, perhaps the major problem facing RM researchers is that of developing a protocol that might actually be useful for forensic investigators examining individual cases (Masip *et al*., 2005). As noted earlier, perhaps one of the more discouraging features of the present findings is that the RM criteria were not discriminating after standardisation; this suggests that it could potentially be very difficult to develop normative criteria, which could be used in the field to classify individual cases. Given this, perhaps one way forward might be to explore the relative efficacy of 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.

Another possible avenue for inquiry is to examine the interaction between RM criteria and the use of verbal fillers (e.g. um and 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. It should also be emphasised, that 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 with both adults and children, it has yet to be compared with RM (Williams *et al*., 2014).

Nevertheless, in the meantime, at the very least, the present results suggest that when judging the veracity 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 investigated systematically, and measured and applied consistently, if researchers wish to compare and replicate findings within and across studies.

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

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. doi: 10.1002/acp.1139

Beaugrande, R. (1984). *Text production: Toward a science of composition*. Norwood, NJ: Ablex.

Brennan, S. E., & 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.

Carlson, J. R., & George, J. F. (2004). Media appropriateness in the conduct and discovery of deceptive communication: The relative influence of richness and synchronicity. *Group Decision and Negotiation*, 13, 191–210. doi:10.1023/B:GRUP.0000021841.01346.35

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

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

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

Davis, M., Markus, K. A., & Walters, S. B. (2006). Judging the credibility of criminal suspect statements: Does mode of presentation matter? *Journal of Nonverbal Behavior,* 30, 181– 198. doi:10.1007/s10919-006-0016-0

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. In Granhag, P. A. & Strömwall, L. A. (Eds), *The detection of deception in forensic contexts* (pp.15-40). Cambridge: Cambridge University Press

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.

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

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:10.1002/acp.1584

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.doi: 10.1348/135532505X49620

Gumperz, J. J., Kaltman, H., & O'Connor, M. C. (1984). Cohesion in spoken and written discourse: Ethnic style and the transition to literacy. In D. Tannen (Ed.), *Coherence in spoken and written discourse* (pp. 3-19). Norwood, NJ: Ablex.

Hauch, V., Blandón-Gitlin, I., Masip, J., & Sporer, S. L. (2012). Linguistic cues to deception assessed by computer programs: A meta-analysis. *Proceedings of the Workshop on Computational Approaches to Deception Detectio*n (pp. 1-4). 13th Conference of the European Chapter of the Association for Computational Linguistics: April 23.

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

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

Hidi, S., & Hildyard, A. (1983). The comparison of oral and written productions in two discourse types. *Discourse Processes*, 6, 91–105. doi:10.1080/01638538309544557

Johnson, M. K., Foley, M. A., Suengas, A., & 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., Hastroudi, S. & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, 114, 3–29.

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

Kroll, B. (1977). Combining ideas in written and spoken English. In T. L. Bennett (Ed). *Discourse across time and space. Southern California occasional papers in linguistics 5*. Los Angeles: Dept. Linguist, University of South California.

Larson, 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

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.

Lindholm, T. (2008). Who can judge the accuracy of eyewitness statements? A comparison of professionals and lay-persons. *Applied Cognitive Psychology*, 22, 1301–1314. doi: 10.1002/acp.1439

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, 259–257.

Manzanero, A. L. & Diges, M. (1996). 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., Sánchez-San Segundo, M., & Herrero, C. (2012). Deception detection from written accounts. *Scandinavian Journal of Psychology*, 53, 103-111. doi: 10.1111/j.1467-9450.2011.00931.x

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. doi:10.1080/10683160410001726356

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

Pu, M. (2006) Spoken and written narratives: A comparative study. *Journal of Chinese Language and Computing*, 16(1), 37–62.

Nahari, G., Vrij, A., & Fisher, R. P. (2012). Exploiting liars' verbal strategies by examining the verifiability of details. *Legal and Criminological Psychology*. doi:10.1111/j.2044-8333.2012.02069.x

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.

Porter, S., & ten Brinke, L. (2010). The truth about lies: What works in detecting 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. i, 20, 443–458.

Santtila, P., Roppola, H., & Niemi, P. (1999). 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.

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

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

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. doi: 10.1002/(SICI)1099-0720(199710)11:5<373::AID-ACP461>3.0.CO;2-0

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. doi: 10.1002/acp.1234

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. doi: 10.1002/acp.1021

Tagg, C. (2009). A corpus linguistic study of SMS text messaging. Unpublished Thesis. University of Birmingham.

Tannen, D. (1985). Relative focus on involvement in oral and written discourse. In: Olson, D. et al.(Eds.), *Literacy, language and learning* (pp. 124–147). New York: Cambridge University Press

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

Vrij, A. (2000). *Detecting lies and deceit: The psychology of lying and the implications for professional practice*. Chichester, England: Wiley.

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., Roberts, K. P., & Bull, R. (2000). Detecting deceit via analysis of verbal and nonverbal behaviour. *Journal of Nonverbal Behaviour*, 24, 239–263.doi: 10.1023/A:1006610329284

Vrij, A., Mann, S., Fisher, R., Leal, S., Milne, B., & Bull, R. (2008). Increasing cognitive load to facilitate lie detection: The benefit of recalling an event in reverse order. *Law and Human Behavior*, 32, 253–265.doi: 10.1007/s10979-007-9103-y

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. doi: 10.1007/s10979-006-9066-4

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.

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