This is the peer reviewed version of the following article: Akehurst, L., Easton, S., Fuller, E., Drane, G., Kuzmin, K. and Litchfield, S. (2015), An evaluation of a new tool to aid judgements of credibility in the medico-legal setting. Legal and Criminological Psychology. doi: 10.1111/lcrp.12079, which has been published in final form at <http://onlinelibrary.wiley.com/doi/10.1111/lcrp.12079/full>. This article may be used for non-commercial purposes in accordance with [Wiley Terms and Conditions for Self-Archiving](#).

Preliminary evaluation of a new tool to aid judgements of credibility in the medico-legal setting

Lucy Akehurst*, Simon Easton, Emily Fuller, Grace Drane, Katherine Kuzmin and Sarah Litchfield

Department of Psychology, University of Portsmouth, UK

**Address for correspondence:* Dr Lucy Akehurst, Department of Psychology, University of Portsmouth, King Henry Building, King Henry I Street, Portsmouth, PO1 2DY. UK.

Email: lucy.akehurst@port.ac.uk

Author note: This is a post-print version of the paper NOT the copy of record. The published version may contain minor grammatical or typographical changes.

Preliminary evaluation of a new tool to aid judgements of credibility in the medico-legal setting

Abstract

Purpose. Clinical psychologists and other health professionals are often requested to act as expert witnesses in Court. They are required to assess, and report upon, the reliability of the accounts of physical and psychological symptoms made by their clients. This study investigated the effectiveness of a checklist drawing upon relevant literature on lying and malingering to aid the detection of exaggeration of physical symptoms.

Method. Sixty-four participants were cast as interviewers and assigned to either a 'checklist' or 'no checklist' condition. Another 64 volunteers were assigned to either a 'truth teller' or 'malingerer' role and, after undergoing a cold pressor procedure, were interviewed about their experience. The interviewers with a checklist drawn from the literature were asked to rate the presence of 28 checklist items on 5-point Likert scales and to indicate whether or not they believed their interviewee was truthful or exaggerating his/her symptoms. The interviewers without the checklist were asked to simply indicate whether their interviewee was truthful or exaggerating.

Results. Evaluators who were not given the checklist did not classify their interviewees at a level significantly better than chance. Those using the checklist achieved an overall hit rate of 70%. Signal detection analysis supported the finding that those with the checklist showed greater discriminability. Nine checklist items significantly discriminated between truth tellers and malingerers. Furthermore the total checklist score was significantly higher for exaggerated accounts than for truthful accounts.

Conclusions. Results suggest that a checklist based on the literatures into lying and malingering warrants further investigation. Such a tool would be useful as an aid for expert witnesses called to provide informed opinion on the likelihood that a claimant is exaggerating, malingering or otherwise misrepresenting difficulties.

Keywords: deception, credibility, medico-legal reports, expert witnesses, malingering

Preliminary evaluation of a new tool to aid judgements of credibility in the medico-legal setting

Clinical psychologists, orthopaedic consultants and other health care professionals can be instructed to assist a Court by providing, in the role of an expert witness, a detailed report on matters such as the cause and course of someone's distress or difficulty after an accident. Expert witnesses are required, when submitting reports in the medico-legal context, to confirm that the opinion they provide "can be expressed finally and without qualification" (Practice Direction supplement CPR Part 35 for Experts and Assessors, UK, 2010, p.10). The process by which an expert might reach such a position of satisfaction is, however, not clear. Experts preparing a medico-legal report will draw upon their professional expertise, providing references to the academic literature as appropriate. However, while experts may be more or less familiar with the academic literature pertaining to the assessment of malingering (e.g. Drob, Meehan & Waxman, 2009; Hall & Hall, 2006; Kramer & Gagliardi, 2009; Resnick, 1995; Rogers, 1997; 2008), fewer may be familiar with theory and research relating to lying in an investigative interview setting.

The established psychological literature on 'lying' relates to the evaluation of an individual's presentation at interview, and can inform interviewers as they consider the validity of someone's reported experiences (e.g. Vrij, 2008; Vrij, Akehurst, Brown & Mann, 2006; Vrij, Granhag & Porter, 2010). While a number of credibility assessment tools have been investigated for the forensic setting, it is proposed here that the underlying principles could serve as the basis for development of a checklist, including informative behavioural cues, to be used by health professionals undertaking assessment through face-to-face interviews for medico-legal reports.

Tools for the detection of lying

Bond and DePaulo (2006) and Vrij (2008), in comprehensive reviews, concluded that lay persons are able to accurately differentiate between liars and truth tellers between 45% and 60% of the time (where 50% can be expected by chance alone). The evidence suggests that, in the first instance, the process of detection itself appears to be a difficult task (Vrij, 2008), and studies have indicated that professional interviewers such as police officers, social workers, and customs officials tend to share the same inaccurate beliefs about how body language changes when people lie (Vrij, Akehurst & Knight, 2006; Vrij, Akehurst, Soukara & Bull,

2004a; Vrij, Edward & Bull, 2001). It is assumed here that physicians, clinical psychologists and others conducting medico-legal assessments for the Courts are unlikely to be any better able to accurately differentiate between liars and truth tellers because of these false beliefs held about nonverbal cues of deceit.

Both verbal and nonverbal behavioural cues can be used to make veracity decisions. However, in the medico-legal setting, a focus on language seems more feasible as nonverbal behaviours can be difficult to detect in real time for those without training (Vrij, 2008, Porter, ten Brinke & Wallace, 2012). As such, no nonverbal cues were included in the checklist under investigation. There are a limited number of paralinguistic cues that have been found to successfully discriminate between truth tellers and liars. As these cues are easier to perceive for an interviewer (e.g. length of pauses and time taken to answer a question (response latency)) they were included in the checklist (see Appendix 1). Pauses and response latencies are expected to be longer for liars compared to truth tellers as liars need to think for longer before giving and during their answers (see DePaulo et al., 2003; Vrij, 2008 for reviews).

The difficulties facing an expert as they endeavour to detect dishonesty via nonverbal cues has led investigators to explore the possibility that what is said might more reliably serve to identify attempted deception. Statement Validity Assessment (SVA) is a tool for judging the credibility of statements based on the content of what an interviewee says. SVA incorporates Criteria-Based Content Analysis (CBCA), which comprises a list of criteria identified as being more frequently included in truthful accounts of actual events (Steller & Köhnken, 1989). It has been proposed that CBCA fosters a more systematic assessment of the veracity of statements by directing the interviewer's attention to specific aspects of someone's description of an event or experience. Blandón-Gitlin, Pezdek, Lindsay and Hagen (2009) have suggested that "Criteria-Based Content Analysis (CBCA) is the most widely used veracity assessment technique for discriminating between accounts of true and fabricated events" (p. 901).

The 19 criteria used in CBCA are organized into four categories (see Table 1). The first category identifies what are described as 'general characteristics', and draws upon the premise that an account of an event that someone has actually experienced will tend to be logical and coherent (Criterion 1), include some digressions or shifts in focus (Criterion 2), and contain a substantial amount of detail (Criterion 3). The second and third categories draw the interviewers attention to the specific contents of the statement (Criteria 4–13) and to motivation-related contents (Criteria 14–18), wherein it is assumed that, for example, a truthful

interviewee will be more likely to report accurately the content of speech or conversations (Criterion 6), and would be more likely to express doubt about his or her own memory of events (Criterion 15). The final category concerns details characteristic of a particular crime that a lay person may question but an expert might recognise as a detail commonly included in the truthful accounts of witnesses to that particular crime. If a substantial number of criteria are present, then CBCA predicts that it is more likely that an account will be truthful. CBCA draws substantially on the assumptions that lying is cognitively more demanding than truth-telling, and that those who lie will be more actively concerned with the impression they are making on interviewers than will truth tellers.

Given that CBCA was not developed for the medico-legal setting some criteria are more directly relevant than others. For example, the criteria 'description of interactions' and 'reproduction of conversation' are not relevant for the reporting of symptoms, they require the recall of a particular event. However, the criterion 'unstructured production' refers to the degree to which an interviewee scatters information throughout their description (rather than presenting details in a structured, coherent and chronological order) and may well be present in a truthful account of symptoms and therefore relevant to the medico-legal setting. Furthermore, criteria such as 'superfluous details' may helpfully draw the medico-legal assessor's attention to consideration of whether an individual's description of an accident or illness includes information that is not essentially relevant. Taking relevance for the medico-legal setting (the reporting of symptoms rather than an event/crime) into account, ten of the CBCA criteria were included in the checklist under investigation (see Table 1 and Appendix 1). Illustrative examples of each CBCA criterion, relating to the medico-legal setting, were included in the checklist to provide extra clarity for evaluators.

Statement Validity Assessment also identifies, in its Validity Checklist (see Table 2), other aspects which may be useful for an interviewer to consider in the medico-legal setting. Furthermore, attention to these elements of the interview can offer a basis for evaluation of the quality of the interview itself (Steller & Köhnken, 1989). A poor interview may lead to misinterpretation of an interviewee's behaviour and their presentation of information. The fact that interviewees may have presented their 'story' repeatedly, and/or may feel annoyed by the impersonal approach of an 'expert' can lead to presentation and behaviour at interview that might easily be misconstrued (e.g. an abrupt interviewee manner and rushed answers may be erroneously considered deceptive). Just as the CBCA criteria may require revision or

adaptation for the medico-legal context, so might the elements of the Validity Checklist of SVA.

The Validity Checklist items are intended to assist interviewers by reminding them to consider some key issues within the context of the interview setting. Truthful individuals in compensation settings may be anxious, fearing inappropriate disbelief by interviewers, or rejection of valid information by interviewers, and some will experience anxiety in the presence of what they perceive to be authority figures. Errors made in the interviewee's reporting of facts may be associated with the extended duration of the litigation process, wherein someone may genuinely be unable to accurately recall detailed information about events which may have occurred five years or more ago. As such, the Validity Checklist items provide a potentially useful list to help interviewers explicitly assess all such aspects more consistently. That said, in a lab-based, medico-legal setting many of these items are not applicable. For the checklist used in this study two items were used (see Table 2 and Appendix 1).

Another potentially useful assessment tool, Reality Monitoring (RM) has been proposed as an aid to the examination of the quality and credibility of information provided in a statement (Johnson & Raye, 1981). RM draws upon the premise that memory characteristics of experienced events differ qualitatively and quantitatively from characteristics of fabricated events. It is suggested that memories originating from real experience are more likely to include perceptual information (visual details, sounds, smells, tastes and physical feelings related to the event), contextual information (information regarding when and where the event happened), and affective information (details about emotional reactions to the event), than accounts based on fabrication. RM suggests that fabricated accounts would involve more information requiring cognitive operations, such as details about thoughts, reasoning, and inferences of events (e.g. "I must have had my coat on as it was very cold that night") than truthful accounts (see Table 3). Masip, Sporer, Garrido and Herrero (2005), in a review of the evidence relating to RM, suggested that empirical results were relatively poor, noting contradictory findings, which they associated potentially with differences in the operationalization and procedures used in various studies. In conclusion, however, they suggested that RM did tend to discriminate above chance level. Vrij (2008) reports, in a review of 10 studies that looked at the effectiveness of RM, an average accuracy rate for truths of 71.7% and for lies of 66.1% (with 68.8% for truths and lies combined). Vrij (2008)

suggests that RM criteria might complement CBCA criteria in that the latter takes cognitive and motivational factors into account whereas the former has a more theoretical basis in memory research. Indeed, three studies (Sporer, 1997; Strömwall, Bengtsson, Leander & Granhag, 2004; Vrij, Akehurst, Soukara & Bull, 2004b) found that the combined accuracy of RM and CBCA criteria was higher than for the CBCA criteria alone.

Again, taking relevance for the medico-legal setting (the reporting of symptoms rather than an event/crime) into account, five of the RM criteria were included in the checklist for the current study and, as for the CBCA criteria, relevant illustrative examples were given for each item (see Table 3 and Appendix 1).

The detection of malingering

The academic literature on lying has largely been based on the examination of active, deliberate and clear cut lying in a laboratory setting – when what is said is wholly true or wholly untrue. In medico-legal settings, however, the detection of partial truth is more likely to be of relevance. When endeavouring to deceive an assessor in relation to their experience of ill-health, pain or restriction, a claimant in the litigation setting might invent some symptoms, exaggerate others, and/or blame a pre-existing ailment on the accident under consideration. Thus, Resnick (1995) distinguished between: pure malingering (feigning non-existent disease); partial malingering (exaggeration of existing symptoms); and false imputation (falsely ascribing real symptoms to unrelated cause).

Resnick (1997) identified a number of aspects that should be considered by the interviewer as he seeks to evaluate the validity of claims made at interview. These included motivation to exaggerate, irregular employment and job dissatisfaction, previous claims for injuries, and psychological test results. It was not possible to include these items in the checklist for the current study due to the laboratory-based nature of the experiment. However Resnick (1997) also suggested that a lack of co-operation at interview and falsely ascribing real symptoms to fabricated event (e.g. suggesting that an already aching neck and back are due to whiplash after a car accident) could be indicators of deception. These two criteria were used in the current checklist (see Appendix 1). However, the issues identified by Resnick (1997) also underline the need within the medico-legal setting to consider the possibility that deception in some circumstances will not necessarily be the product of deliberate and conscious motivation. The possibility that someone may receive large sums in compensation if

their symptoms and difficulties persist may distort and prolong their experience of those difficulties through the very process of being repeatedly asked about the problems. The relevance of awareness of the broader context was emphasised by Ferrari, Kwan, Russell, Pearce and Schrader (1999), who suggested that prognoses in medico-legal settings can be substantially affected by a range of other aspects, including: blame, expectations, labelling and social factors. From these aspects three items were included in the current checklist; blaming others and believing others behaved recklessly (as indicators of lying) and an expectation that the symptoms may only last a short time or show significant improvement (as an indicator of truthfulness) (see Appendix 1).

Rogers (1997) suggested that a medico-legal assessor might usefully explore the nature and pattern of reporting of symptoms and difficulties at interview. Rogers (1997) suggests the assessor should pay particular attention to the prevalence of: indiscriminate symptom endorsement (confirming presence of all symptoms asked about); improbable symptoms (unlikely difficulties in the context); presence of improbable combinations of symptoms (e.g. contradictory patterns of symptoms); and presence of symptoms of improbably extreme severity. These factors make up the final four checklist items, all indicators of lying, for the current investigation (see Appendix 1).

Interviewing and the medico-legal expert

The literature on detecting lying indicates that expert witnesses in the medico-legal setting could make their task easier if they encourage their clients to talk. The structured evaluation of material using SVA and RM also serves to add confidence to the conclusions reached. The absence of such a structured approach in most medico-legal settings may reduce reliability of the opinion expressed, and so lead a Court to feel less confident about the conclusions reached. It could therefore be the case that a structured process could help experts elicit, attend to, and then assess the aspects of an interview that will help them support the level of confidence they place in their opinions. An aide-memoire might be helpful, and reduce the expert's cognitive load. More importantly, perhaps, it might also be that a more structured approach will offer the objective advantages potentially associated with CBCA, thereby improving the assessor's ability to identify lying and malingering.

This approach however differs from other established methods such as the Structured Interview of Reported Symptoms – 2nd Edition (SIRS-2; Rogers, Sewell & Gillard, 2010) and

the Miller Forensic Assessment of Symptoms Test (M-Fast; Miller, 2001, 2004). It has been suggested that undue reliance on such devices may be criticised in the medico-legal setting (Frederick, 2012). The original SIRS, (Rogers, Bagby, & Dickens, 1992) used eight scales in relation to the assessment of individuals as honest or feigning, and five scales in relation to response styles. Scales included those entitled: Rare Symptoms, Symptom Combinations, Improbable and Absurd Symptoms, Blatant Symptoms, Subtle Symptoms, Selectivity of Symptoms, Severity of Symptoms and Reported vs. Observed Symptoms. The M-Fast (Miller, 2001) was based on strategies identified by Rogers (1997), and used themed scales to address aspects such as unusual hallucinations, extreme symptomatology, rare combination, negative image and suggestibility (Miller, 2004).

However in the UK medico-legal setting such devices are rarely used, perhaps because the relatively extreme symptoms would appear, on the basis of the authors' experience, less relevant to the more typical range of difficulties encountered and reported. Whilst there is overlap with some of the aspects previously mentioned, the SIRS- 2 and M-Fast approach is distinct from that of the aide-memoire style of checklist envisaged by the current authors, the preference for the latter springing from criticisms of the reliability and validity of such devices in the medico-legal setting (Faust, 2012).

Development of a checklist for use by medico-legal experts

In summary, for the current investigation, a broad selection of possible verbal cues to lying/malingering and truth-telling was collated (based on the literature and techniques outlined above). Aspects deemed inappropriate to the medico-legal setting were excluded. Where appropriate, items listed in the investigative interviewing literature were rephrased to make them relevant to the medico-legal setting (see Appendix 1). A list of 28 criteria thereby emerged. On considering how the presence of each criterion might be assessed at interview, a graded approach, rather than an 'all-or-nothing' assessment was selected, noting that the interviewer's opinion will tend to be subjective rather than objective. The resulting checklist tool was evaluated in terms of its usefulness in discriminating truthful interviewees from those who were malingering. It was hypothesised that interviewers given the checklist tool would be better able to discriminate between truthful and exaggerated accounts of physical symptoms than those not given the checklist (Hypothesis 1). It was further hypothesised that the total checklist score would be significantly higher for exaggerated compared to truthful accounts (Hypothesis 2) and the 28 checklist items would significantly discriminate between truthful

and exaggerated accounts in the directions specified above (Hypothesis 3). Although the literature suggests that the 28 checklist items should discriminate between truth and exaggeration, the authors were interested to see which items most influenced interviewers when they made their judgements.

Method

Participants

Participants were recruited via advertisements to students and staff at a University in the UK. A small payment was made to each attendee. 128 participants took part in the investigation, 64 in the role of interviewer and 64 in the role of interviewee. Of the 64 interviewers, 32 (14 males and 18 females) were given a checklist to help them assess credibility, these participants ages ranged from 18 years to 54 years ($M = 22.69$ years, $SD = 8.14$ years). The other 32 interviewers (11 males and 21 females) were not given a checklist, their ages ranged from 18 years to 47 years ($M = 22.56$ years, $SD = 6.08$ years). Of the interviewees, 32 (8 males and 24 females) were asked to tell the truth about their symptoms following a cold pressor procedure, their ages ranged from 18 years to 38 years ($M = 21.97$ years, $SD = 4.23$ years), and 32 (15 males and 17 females) were asked to exaggerate their symptoms when asked at interview, their ages ranged from 19 years to 61 years ($M = 26.03$ years, $SD = 9.71$ years).

Design

After undergoing a cold pressor procedure truth tellers and malingerers were interviewed about their experience. They each took part in two interviews, telling the truth both times or exaggerating both times. All interviewers were given the same interview protocol. The interviewers with the checklist were asked to rate the presence of 28 checklist items on 5-point Likert scales ranging from [1] absent to [5] strongly present indicating whether or not they believed their interviewee was truthful or exaggerating his/her symptoms. The interviewers without the checklist were asked to simply indicate whether their interviewee was truthful or exaggerating. All interviewers were additionally asked to rate their confidence in their judgement.

Materials

The checklist. A checklist was developed in which a higher score indicated greater likelihood or risk of lying and/or malingering. The items included in the checklist were taken from some 49 items generated from the relevant literatures. These included all 19 CBCA criteria and all 11 Validity Checklist items from SVA (see Steller & Köhnken, 1989 and Tables 1 and 2), all 8 RM criteria (see Vrij, 2008 and Table 3), 2 paralinguistic cues that had been reported in the literature to discriminate between truth tellers and liars (DePaulo et al., 2003; Vrij, 2008) and the items outlined in the malingering literature (2 from Resnick, 1997; 3 from Ferrari et al., 1999; 4 from Rogers, 1997). The researchers reviewed the identified items, drawing upon experience of working as an expert witness and 21 items were excluded because they were not relevant to the medico-legal setting or were not relevant to the lab setting (see Tables 1,2 and 3). Thus, assessors were asked to consider 28 items.

For each question the participants were asked, “Please circle the number that best reflects your view”. There was a scale for each question where, dependent on whether the literature suggested the item should appear more often in truthful or exaggerated accounts, sometimes 1= not at all present and 5 =very much present and for other items 1 = very much present and 5 = not at all present. In short, the higher the score for each item the more likely it was that the interviewee was exaggerating. For example, the literature suggests that if an interviewee blames someone else or considers the other party was reckless they are more likely to be lying/exaggerating so for this item the scale ran from 1 = not at all present to 5 = very much present. Conversely, research suggests that the CBCA criteria are more likely to be present in truthful accounts so for these items (e.g. unusual details, reproduction of conversation, quantity of detail) the scale ran from 5 = not at all present to 1 = very much present. The interviewers were asked to add up their 28 item scores to give a total and were instructed that the higher the total for the checklist, the more likely that their interviewee was exaggerating his/her symptoms. The full checklist can be seen in Appendix 1, which also identifies the source of selected items.

Procedure

Participants were recruited in groups of four. Two were assigned to the role of interviewer and two to the role of interviewee. All interviewees were asked to take part in a cold pressor procedure that would constitute the event that they would later be interviewed

about. One interviewee in each group was assigned to be a truth teller, the other a malingerer. Each interviewer conducted two interviews, one with the truth teller and one with the malingerer in his group. However, interviewers were told that it was equally likely that they could see any combination of truth tellers or malingerers (i.e. two truth tellers, two malingerers or one of each). The order in which truth tellers and malingerers were presented to interviewers was counterbalanced.

Cold pressor procedure. A cold pressor apparatus was used consisting of an insulated container, measuring 30 x 40 x 30 cm. The entire container was filled with water and ice and was equipped with a moveable armrest used to immerse each participant's arm. Water temperature was maintained at 2° to 4°C. Those undertaking the cold pressor procedure were asked to read and complete a medical risk questionnaire, provide biographical details, and then, if no concerns emerged, complete an informed consent form. A heart rate monitor was fitted and monitored throughout the cold pressor task as part of the recommended risk management procedure. Participants were assured that the procedure would not result in physical injury. There were no cases of participant withdrawal. Participants were instructed to place their arm on the moveable armrest of the cold pressor apparatus, to lower their arm into the ice water, and to keep their arm immersed for as long as they felt they could tolerate with a maximum period of immersion of 3 minutes. If participants lasted for 3 minutes they were instructed by the experimenter to remove their arm¹. All participants were asked to complete the Short Form McGill Pain Questionnaire (SF-MPQ; Melzack, 1987), once in relation to level of discomfort during the ice plunge and a second time in relation to discomfort following removal of their hand from the water. That is, they were asked to rate, on 4-point scales (0 = none, 1 = mild, 2 = moderate, 3 = severe) throbbing, shooting, stabbing, sharpness, cramping, gnawing, burning, aching, heaviness, tenderness, splitting, tiring, sickening, fearfulness and punishing in terms of their severity. They were then given directions to tell the truth or exaggerate their experience to those interviewing them.

Truth tellers. Those in the truth-telling condition were told to "Tell the truth about your experience. Try to include lots of truthful detail about your physical and psychological symptoms to make a really credible impression. You might imagine that, should the interviewer believe you, you would be awarded monetary compensation". The participants were then given 5 minutes to prepare their accounts.

¹ Participants kept their arms in the cold water for between 64 and 180 seconds ($M = 115$ seconds, $SD = 6.04$ seconds).

Malingers. Malingers were given the following instruction, “We would like you to convince the interviewer that the discomfort you experienced was worse than it actually was. You are also asked to pretend that the symptoms are ongoing and substantial. Try to include lots of detail about your physical and psychological symptoms to make a really credible impression. You might imagine that, should the interviewer believe you, you would be awarded monetary compensation”. These participants were then given a list of symptoms for consideration when preparing to exaggerate symptoms². They were given 5 minutes to prepare their accounts.

Each interview was video-taped, unless participants declined to be filmed. After their interviews, each ice-plunge participant was asked to respond on a 7-point scale to a number of questions about their experience of the interviews (see Results section).

Interviewers. Those assigned to be interviewers were split into two groups, only one of which used the specialist checklist.

All interviewers were given the following instructions: “Our interviewees have undergone a cold pressor procedure which involves them holding their forearm and hand in very cold water. We have asked some people to exaggerate the discomfort they have experienced, while others will be simply telling the truth. We want you to decide whether your interviewee is telling the truth or exaggerating. Here are some interview questions that you can use to help you”.

The interviewers who were given the checklist were also told: “Here also is a checklist that the research team have devised. This checklist is made up of many criteria which are thought to be indicators of truthfulness or lying. Before you make your final credibility judgement, work through this checklist and keep your responses in mind when you make your final decision about your interviewee. Ring a number for each item and then add up to get a total score. The higher the total score the more likely that your interviewee is exaggerating”.

All interviewers were provided with a protocol for the interview as follows:

² Truth tellers and malingers were treated differently to mimic real life. A list of symptoms was provided to malingers to replicate the preparation that it is likely that those in real life would undertake before attending a meeting during which they intend to lie about their symptoms. Those malingers may resort to websites and other medical literature to help them to concoct a plausible account of non-existent or exaggerated symptoms. Truth tellers, on the other hand, tend to believe in the illusion of transparency i.e. that the truth will ‘shine through’ and so will simply be confident to report their true symptoms with little preparation pre-interview.

“Please could you tell me everything you can about that experience, even little things that you do not think are important? Can you tell me about the symptoms that you experienced whilst you had your hand in the cold water? Can you tell me about the symptoms you experienced once you had taken your hand out of the cold water? Are you still experiencing any symptoms? If yes, could you tell me about the symptoms that you are still experiencing? Now in your own words please just tell me once again about your experience of putting your hand in the ice water and any subsequent discomfort up to the time of this interview. Please include as much detail as you can.”

After each interview, interviewers were asked to complete a questionnaire about their impressions of the interviewees. Before completing the questionnaire, those in the checklist condition worked through each of the 28 items and calculated a total checklist score (see Materials section above). The questionnaire required that interviewers indicated whether or not they thought their interviewee was telling the truth or exaggerating his/her symptoms (a dichotomous rating to allow for accuracy rates to be calculated). They were also asked to rate on 7-point Likert scales: Credibility (1 = wholly truthful account to 7 = wholly untruthful account and 1 = wholly exaggerated to 7 = not at all exaggerated); confidence in their opinion (1 = very sure to 7 = not at all sure); and the difficulty of the task (1 = extremely difficult to 7 = not at all difficult).

Those who had used the checklist were also asked; “Do you think the checklist helped you make a decision about the credibility of your interviewee?” (1 = very much so to 7 = not at all) and “Did the total score on your checklist reflect or contradict your instinct about the credibility of your interviewee?” (1 = complete agreement to 7 = complete contradiction).

Results

Checks on interviewee ratings

All interviewees were asked, after the cold plunge but before their interview, to complete the SF-MPQ once regarding their symptoms during the ice plunge and then again regarding their symptoms after the ice plunge. They were asked to rate 15 symptoms (see above) in terms of their severity, on 4-point scales (0 = none, 1 = mild, 2 = moderate, 3 = severe). Therefore scores could range from 0 to 45. The higher the SF-MPQ score, the more overall discomfort a participant was feeling. A one-way MANOVA with status of interviewee (truth teller or malingerer) as the independent variable and total scores on the SF-MPQs for

during and after the ice plunge were the dependent variables. There was no significant difference between truth tellers ($M = 26.56$, $SD = 6.19$) and malingerers ($M = 28.61$, $SD = 8.99$) for their SF-MPQ total for perceived discomfort during the ice plunge, $F(1, 124) = 2.20$, $p = .14$, $d = .27$. Furthermore, there was no significant difference between truth tellers ($M = 21.97$, $SD = 4.62$) and malingerers ($M = 22.06$, $SD = 5.39$) for their SF-MPQ total for discomfort after the ice plunge, $F(1, 124) = .011$, $p = .92$, $d = .02$.

After they had been interviewed the interviewees were asked to respond on 7-point Likert scales (1 = not at all to 7 = completely) questions about their experience of the interviews. They were asked: “To what extent do you think the interviewer believed your account?”; “To what extent did you have to think hard to convince the interviewer you were telling the truth?”; “To what extent did you try to control your body language during the interview?”; “To what extent did you have to control what you said during the interview?” and “How nervous did you feel during the interview?”. A 2 (truth teller or malingerer) x 2 (interviewed with or without a checklist) MANOVA was conducted with the scores for these five questions as the dependent variables. The MANOVA revealed a significant multivariate main effect for whether or not the interviewee was a truth teller or an exaggerator, Wilks’ $\lambda = .61$, $F(5, 120) = 15.40$, $p < .001$, $\eta^2 = .97$. Significant univariate effects for status of interviewee were also obtained: The malingerers indicated that they needed to think significantly harder ($M = 4.50$, $SD = .21$) than the truth tellers ($M = 3.33$, $SD = .21$), $F(1, 124) = 16.03$, $p < .001$, $\eta^2 = .11$, $d = 5.57$; the malingerers indicated that they attempted to control their behaviour ($M = 4.28$, $SD = .22$) and the content of what they said ($M = 5.14$, $SD = .20$) significantly more than the truth tellers ($M = 2.81$, $SD = .22$ and $M = 2.73$, $SD = .20$, respectively), $F(1, 124) = 23.33$, $p < .001$, $\eta^2 = .16$, $d = 6.68$ and $F(1, 124) = 75.68$, $p < .001$, $\eta^2 = .379$, $d = 12.05$; and the malingerers indicated that they were significantly more nervous ($M = 3.77$, $SD = .22$) than the truth tellers ($M = 2.77$, $SD = .22$), $F(1, 124) = 10.39$, $p = .002$, $\eta^2 = .077$, $d = 4.55$. There was no significant difference between the truth tellers ($M = 4.95$, $SD = .18$) and the malingerers ($M = 4.55$, $SD = .18$) in terms of how much they felt they had been believed by their interviewers $F(1, 124) = 2.51$, $p = .12$, $\eta^2 = .02$, $d = 2.22$.

At a multivariate level there was no significant main effect for whether or not the interviewee was interviewed with a checklist nor were there any univariate effects. There were no significant interaction effects.

Checks on interviewer ratings

After each interview, all interviewers were asked to complete a questionnaire about the truthfulness of interviewees and their confidence in their judgements. A 2 (truthful or malingerer) x 2 (interviewed with or without checklist) MANOVA was conducted with the scores for these questions as the dependent variables. The MANOVA revealed a significant multivariate main effect for truthfulness of interviewee, Wilks' $\lambda = .82$, $F(7, 118) = 3.80$, $p < .001$, $\eta^2 = .184$. One significant univariate effect showed that the truth tellers were rated as significantly less likely to be exaggerating ($M = 5.06$, $SD = 1.68$) in comparison to those who were malingering ($M = 3.74$, $SD = 1.81$), $F(1, 124) = 18.61$, $p < .001$, $\eta^2 = .130$, $d = .76$. There were no significant univariate effects for confidence ratings and no significant interaction effects. There were no multivariate or univariate main effects for use of checklist and no significant interaction effects.

Those who had used the checklist were also asked; "Do you think the checklist helped you make a decision about the credibility of your interviewee?" (1 = very much so to 7 = not at all) and "Did the total score on your checklist reflect or contradict your instinct about the credibility of your interviewee?" (1 = complete agreement to 7 = complete contradiction). A further one way MANOVA using only the data from those who used the checklist was conducted with the status of the interviewee (truth teller or malingerer) as the independent variable and the responses for these two questions as the dependent variables. No significant differences were found. With regard to ratings regarding the helpfulness of the checklist, for those who interviewed truth tellers the mean rating was 2.94 ($SD = 1.39$) and for those who interviewed malingerers the mean rating was 2.97 ($SD = 1.46$), $p = .97$. With regard to ratings of whether the scores on the checklist agreed with instinct or not, for those who interviewed truth tellers the mean ratings was 3.28 ($SD = 1.33$) and for those who interviewed malingerers the mean rating was 3.47 ($SD = 1.22$), $p = .56$.

Overall judgement accuracy

To allow comparison of results with those of similar studies, judgement accuracy was computed. The evaluators who were not given the checklist to help them with their task achieved an overall hit rate of 58%, with 59% correct classification of truthful interviewees and 56% correct classification of those who were exaggerating their symptoms. A series of two-tailed binomial tests indicated that these evaluators did not classify their interviewees at a level significantly better than chance (50%).

The evaluators who were given the checklist to help them with their task achieved an overall hit rate of 70% with 75% correct classification of truthful interviewees and 66% correct classification of those who were exaggerating their symptoms. A series of binomial tests indicated that these evaluators did classify their interviewees at a level significantly better than chance ($p < 0.01$).

Signal detection analysis

Meissner and Kassin (2002) have suggested that Signal Detection Theory (SDT) can be used to analyse accuracy of credibility assessments. Using SDT, the performance of detectors can be assessed with “two conceptually and computationally independent parameters” (p.471). As such, the performance of our evaluators was calculated in terms of discrimination accuracy d' and response bias B''_D . Table 4 shows the total frequency of hits, misses, false alarms and correct rejections for interviewers with and without the checklist. A hit is a malingerer who is accurately classified as exaggerating symptoms; a miss is a malingerer who is labelled a truth teller; a false alarm is a truth teller who is labelled as exaggerating his/her symptoms; and a correct rejection is a truth teller who is accurately classified. To ascertain whether those with or without the checklist showed greater discrimination, d' was calculated for both groups. This is a pure measure of sensitivity, it measures the signal and the noise means in standard deviation units. It corrects for response bias and guessing. A value of 0 indicates evaluators showed an inability to distinguish exaggerators from truth tellers. Values above 0 indicate an ability to distinguish malingerers from truth tellers and negative values indicate response confusion. For those evaluators given the checklist $d' = 1.08$ and for those without the checklist $d' = .39$. Two one-sample t-tests demonstrated that evaluators with the checklist discriminated malingerers from truth tellers significantly greater than a value of 0 (a score indicative of being unable to discriminate), $t(31) = 2.881$, $p = .007$; whereas evaluators without the checklist were not able to significantly discriminate malingerers from truth tellers, $t(31) = 1.000$, $p = .325$. An independent samples t-test showed that there was a significant difference between those with a checklist and those without in terms of overall discrimination, $t(62) = -2.34$, $p = .023$. These data demonstrate that those with the checklist showed greater discriminability than those who did not receive it thus supporting Hypothesis 1.

Finally, we calculated B''_D as a measure of response bias (i.e. the tendency to respond ‘truth’ or ‘exaggeration’). A value of 1 means participants did not favour a particular response, >1 means participants showed a bias towards a truth judgement and <1 means that

participants showed a bias towards a judgement that the interviewee was exaggerating. For those evaluators with no checklist $B''_D = 1.02$ and for those with a checklist $B''_D = 1.16$. Two one-sample t-tests demonstrated that for those evaluators who did not use the checklist, response bias was not significantly greater than a value of 1 ($p = .142$) and nor was response bias significantly greater than a value of 1 for those who did use the checklist ($p = .061$).

Discriminating objective truths from exaggeration

To determine whether total checklist scores and scores for the individual items on the checklist were predictive of objective truth status, a discriminant analysis was conducted for the raters who had the checklist with actual truth status of interviewees as the dependent grouping variable and the total checklist scores and individual criteria scores as the predictor variables. This yielded a significant discriminant function (Wilks' $\lambda = 0.64$, $X^2(11) = 25.14$; $p < .001$) and an overall classification rate of 81% (78% for truthful interviewees and 84% for exaggerating interviewees).

The discriminant analysis revealed that reporting improbable levels of severity of symptoms obtained the highest pooled within-group correlation ($r = .58$). The other checklist items found to discriminate significantly between truthful and exaggerated accounts were confirming the existence of many symptoms ($r = .56$), reporting symptoms that don't readily fit those expected ($r = .53$), the interviewee reporting that the symptoms should only last a short time ($r = .45$), the interviewee reporting that someone else behaved recklessly ($r = .45$), the checklist total ($r = .41$), the interviewee reporting their own mental state ($r = .39$), reports of unusual details ($r = .37$), the individual blaming someone else ($r = .36$), and reports of unlikely or contradictory symptoms ($r = .34$) (see Table 5).

With regard to the d scores, Cohen (1977) suggested that a d score of around .50 can be considered a medium effect and one around .80 can be considered a strong effect. Therefore, in the present study, it can be seen that at least the top three criteria in Table 5 show strong effects and the next seven show medium effects. As can be seen from Table 5 the significant criteria discriminated between truthful and exaggerated statements in the directions we would have predicted. That is, the total score was higher for exaggerated accounts than for truthful accounts thus supporting Hypothesis 2. Evaluators were more likely to indicate the presence of improbable levels of severity of symptoms, the existence of many symptoms, symptoms that don't readily fit those expected, blaming someone else and suggesting that someone else

behaved recklessly, and unlikely or contradictory symptoms significantly more for exaggerated accounts than truthful accounts. On the other hand, evaluators were more likely to indicate the presence of minimising the symptoms (saying they would last a short time or improve significantly), unusual details and accounts of mental state more often in truthful accounts than exaggerated accounts. These findings partially support Hypothesis 3, nine of the 28 criteria in the checklist significantly discriminated between truthful and exaggerated accounts in the predicted direction.

Discriminating judgements

A further discriminant analysis was conducted to determine whether total checklist scores and scores for the individual items on the checklist were predictive of the judgements that were made by the evaluators. Judgements of 'truthful' and 'exaggerated' (regardless of actual credibility) were the dependent grouping variable and the total checklist scores and individual criteria scores were the predictor variables. This yielded a significant discriminant function (Wilks' $\lambda = 0.47$, $X^2(14) = 41.83$; $p < .001$) and an overall classification rate of 83% (89% for truthful interviewees and 76% for exaggerating interviewees). The discriminant analysis revealed that reporting symptoms that don't readily fit those expected obtained the highest pooled within-group correlation ($r = .59$). The other checklist items found to discriminate significantly between judgements of 'truth' or 'exaggeration' were reports of unusual details ($r = .48$), plausibility of answers ($r = .48$), reporting improbable levels of severity of symptoms ($r = .46$), reporting unlikely or contradictory symptoms ($r = .45$), the checklist total ($r = .43$), confirming the existence of many symptoms ($r = .42$), the interviewee reporting that someone else behaved recklessly ($r = .36$), the interviewee pausing for a long time between the end of a question and the beginning of the answer ($r = .35$), the interviewee giving clear answers ($r = .30$), the individual blaming someone else ($r = .30$), the interviewee reporting that the symptoms should only last a short time ($r = .28$) and the interviewee pausing for a long time when asked about symptoms compared to when asked about other things ($r = .27$) (see Table 6). It can be seen that at least the top 8 criteria in Table 6 show strong effects and the next 5 show medium effects.

As can be seen from Table 6, the significant criteria discriminated between judgements of truthful and exaggerated statements in the directions we would have predicted. Nine of the significant checklist items were the same as those highlighted in the discriminant analysis above. For the additional significant checklist items for this second analysis, plausible answers

and long answers were rated as more present in the accounts that were judged as truthful compared to those that were judged as exaggerated (regardless of actual credibility). Long pauses in general and long pauses before answering questions (response latency) were all rated as more present in accounts rated as exaggerated compared to those rated as truthful.

Discussion

Interviewers were provided with a checklist drawn from the literature into lying and malingering, and were asked to indicate whether or not they believed their interviewees undergoing a cold pressor procedure were being truthful or exaggerating their physical and/or psychological symptoms. Interviewers using the checklist achieved 75% correct classification of truthful interviewees and 66% correct classification of those who were exaggerating their symptoms. These data are similar to those reported when investigative interviewers are trained to use the CBCA and RM techniques for laboratory studies. Success rates in these studies have ranged from 65% to 80% (Vrij, 2008). Interviewers who were not given the checklist did not classify their interviewees at a level significantly better than chance (50%). These findings are also similar to those found in previous studies in the context of investigative interviews. A meta-analysis of 206 studies showed an average hit rate of 54% for interviewers with no training in lie detection (Bond & DePaulo, 2006).

Discriminant analysis yielded an overall classification rate of 81% (78% for truthful interviewees and 84% for exaggerating interviewees). Nine checklist items significantly discriminated between truth tellers and malingerers. Of these, three elicited strong effect sizes (Cohen's $d > 0.8$) reporting improbable levels of severity of symptoms, confirming the existence of many symptoms and reporting symptoms that don't readily fit those expected. Furthermore the total checklist score was significantly higher for exaggerated accounts than for truthful accounts. These findings support the hypotheses and the proposal that it may be possible to improve the detection of lying or malingering in the medico-legal interview setting with provision of a checklist.

The nine items that significantly discriminated between truth tellers and exaggerators were drawn mainly from the work of Rogers (1997) and Ferrari et al. (1999) with two from CBCA. The two successful CBCA criteria in the current study were 'accounts of own mental state' and 'inclusion of unusual or unexpected detail'. It is somewhat surprising that 'quantity of detail' did not appear as a good predictor in this study as it has consistently been shown in

previous research to be a reliable indicator with truth tellers giving significantly more detail than liars. That said, Vrij (2008) suggests that it could be the quality of statements (e.g. inclusion of unusual detail, unexpected complications) that might better discriminate truth from fabrication. Furthermore, the wording of the interview questions in our mock medico-legal interview may not have prompted as much detail from interviewees as those questions used in investigative interviews thus limiting the possibility of a significant difference in amount of detail between truth tellers and malingerers. The effect of type of interview question on elicitation of detail requires further investigation in the medico-legal setting. It is not unusual that the motivation-related criteria of CBCA (e.g. 'spontaneous corrections', 'admitting lack of memory' and 'raising doubts about answers') did not receive support. This is often the case in laboratory based studies (Akehurst, et al., 2004; Gödert et al, 2005) as the stakes for interviewees are not as high as they would be in a real world setting.

It is of interest to note that eight of the items and the checklist total score significantly discriminated between truths and lies *and* significantly discriminated between judgements of truth or lie. This would indicate that assessors considered these items to be helpful which appeared to assist in the objective differentiation of truth tellers and those exaggerating. However, there were four items which were considered to be of relevance by the assessors which were absent from the list that successfully discriminated between truths and lies. The items: Plausible answers, duration of answers, length of pauses, and response latency were judged by those undertaking the task to be valuable, but may actually have been irrelevant. The valuing of these questions by assessors in their deliberations may actually have been unhelpful. Further examination of this aspect would appear warranted in an endeavour to ensure both the efficient brevity of the checklist, and the avoidance of drawing assessor's attention to aspects that are associated with possible misinterpretation.

Overall, therefore, these findings indicate that the literature on lying and deception developed in the forensic setting may be less relevant to the medico-legal assessor (with the exception of reports of unusual detail and the interviewee's mental state which were taken from the CBCA technique). The items taken from the work of Rogers and Ferrari would appear to be more readily considered by the interviewer when reflecting upon an interaction and it may be that some of the items taken from CBCA and RM reflect the expected use of written transcripts and video-recorded interviews which would be rare in the medico-legal setting. This issue warrants further investigation.

The nature of medico-legal assessments precludes the reliable evaluation of such a checklist in the field (i.e., with real claimants). Claimants cannot be expected to admit to deception before or after an interview while claims progress, and follow-up enquiries after settlement of claims would be unreliable and difficult for a number of reasons (e.g., when attempting to establish ground truth). The present study therefore sought to create a situation which offered an opportunity to explore various key aspects of the assessment of interviewees in a format which offers reasonable ecological validity. There are, however, in this kind of research, unavoidable limitations as a result of the need to avoid causing participants unacceptable or lasting discomfort and distress. The motivation to maintain deception would be distinct in the medico-legal setting, although the motivation to be believed would be perhaps largely similar.

High rates of malingering have been estimated in various medico-legal settings. Ardolf, Denney and Houston (2007) suggested that the combined rate of probable and definite malingered neurocognitive dysfunction in a sample of criminal defendants was 54.3%. Greve, Bianchini and Curtis (2009) supported the proposal that the presence of external incentives can lead to high rates of exaggeration. However, in the broader medico-legal literature, estimates of malingered symptoms after personal injury have ranged from 1% to 50%, depending on whether rates are reported by psychiatric studies, insurance companies or plaintiffs' legal representatives (Appelbaum et al., 1993). The impact of base rates on the psychometric reliability of any measure in the applied setting warrants careful investigation. In this laboratory study, the ratio of truth tellers to liars or exaggerators was controlled, although assessors had no basis on which to make predictions of relative frequencies. This aspect could be manipulated in future studies.

This study explored the utility of a checklist drawn up by the authors based on their respective expertise and experience. There was therefore a degree of subjectivity in the choice of items, and some items required changes in wording and the addition of illustrations so as to be appropriate to the experimental setting. Further research will explore the views of other experts on the item selection, and review the process of selection where appropriate. In addition, further exploration of possible influence of illustrative examples needs to be undertaken.

The specific focus of the present study should be noted. The current study did not test each of Resnick's (1995) types of malingering, rather focussing on partial malingering. In

addition, the limitations imposed by the experimental procedure here used did not allow a full exploration of the potential impact of the factors identified by Resnick in an aide-memoire of the type envisaged. The indication that CBCA and RM criteria may not be as helpful in the detection of malingering warrants further investigation, and it may be the case that they would be more predictive in the detection of pure malingering or false imputation, than of partial malingering. Furthermore, both CBCA and RM are usually used to assess deception using transcript/video of an interview, and this may therefore explain their apparent limitations when used in the manner adopted here. It could also be that some items were easier to understand than others and that, with clearer definitions of the CBCA and RM criteria, interviewers may use them to better advantage. Thus it could be the case that training in CBCA would increase the utility of the CBCA related items for medico-legal experts, but such training in the absence of evidence of significant benefit may well be seen as too costly both in terms of time and financial outlay. The aide memoire envisaged here would most probably need to be freely accessible, readily understood and require no training if it were to be widely adopted among busy professionals in practice.

Participants were motivated to mangle in the study with a suggestion that if they convinced the interviewer of their credibility they might imagine being awarded compensation. However, this does not mirror the motivation of interviewees in a medico-legal setting in the real world. Further exploration of the degree to which participants were motivated to exaggerate convincingly during their interviews is warranted, and the influence of higher degree of incentive to lie successfully in particular would be worthy of attention. In addition, experimental manipulation of high versus low stakes deception scenarios would be of interest, noting that such issues are of particular relevance in the context of the medico-legal setting.

In the current study, both liars and truth tellers were instructed to "include lots of detail about your physical and psychological symptoms" and this may have reduced the effectiveness of certain checklist items relating to quantity and clarity of detail. In the real world it is unlikely that such an explicit instruction would be given to interviewees therefore instructions given pre-interview warrant further exploration.

The Diagnostic and Statistical Manual (DSM-5, 2013) proposes that malingering is strongly suspected if there is any combination of: a medicolegal context; discrepancy between the claim and the findings; lack of cooperation and antisocial personality disorder. The DSM-5 definition of malingering refers to the false or grossly exaggerated physical or psychological

symptoms in the context of external incentives, and there are various other definitions that require the presence of an external incentive (Larrabee, 2007; Sweet et al., 2000). In this study, the external incentive could be seen as an aspect that could be manipulated more explicitly, and so future research might wish to explore the effect of increasing the incentives on performance of both those seeking to mislead and those endeavouring to detect exaggeration.

Conclusion

The current study indicates that further exploration of the use of a checklist by interviewers in the medico-legal setting is warranted. Refinement of the checklist in the laboratory setting would assist in development of a better understanding of the key features and processes of the use of such a device. Inter-rater reliability and test-re-test reliability can be assessed, and it would be preferable to explore the use of the device by expert witnesses in the laboratory setting. The use of such a checklist in the medico-legal assessment setting, were it to be shown to offer acceptable levels of reliability and validity in the laboratory setting, might be considered as relevant at various levels. In the first instance, it might be that just using the checklist could offer the Courts some reassurance that the expert assessor has considered more relevant (rather than potentially irrelevant) aspects of an interviewee's presentation. It is possible that the numerical values produced from the scoring on the checklist could be used to offer some statistically derived indication of likelihood of truthfulness or deception, as well as some measure of confidence in the indications. However, we would be wary of any use of such a tool which encourages undue confidence in its validity or reliability in an interview. We do believe that these results suggest that the further examination of the process would be appropriate. Nevertheless, we note criticism of measures such as the Structured Inventory of Reported Symptoms (Rogers et al., 1992; Rogers, Sewell & Gillard, 2010) with regard to over-reliance in the court setting (DeClue, 2011), and again emphasise that the aim of the tool presented here in preliminary form is to draw the attention of assessors to potentially relevant issues rather than serve as some form of psychometric test. It may be the case that an aide memoire of the kind presented here in outline would be more widely adopted in settings where the more formal existing assessments of deception, are either considered to be impractical, given time constraints, or where experts do not feel that evidence for validity and reliability of devices currently available warrants their use in the medico-legal setting.

References

- Akehurst, L.G., Köhnken, G., Vrij, A., & Bull, R. (1996). Lay persons' and police officers' beliefs regarding deceptive behaviour. *Applied Cognitive Psychology*, 10, 461-471. doi: 10.1002/(SICI)1099-0720(199612)10:6<461::AID-ACP413>3.0.CO;2-2.
- Akehurst, L., Manton, S., & Quandt, S. (2011). Careful calculation or a leap of faith? A field study of the translation of CBCA ratings to final credibility judgements. *Applied Cognitive Psychology*, 25(2), 236-243. doi: 10.1002/acp.1669.
- Akehurst, L., Bull, R., Vrij, A. & Köhnken, G. (2004). The effects of training professional groups and lay persons to use Criteria-Based Content Analysis to detect deception. *Applied Cognitive Psychology*, 18, 877 – 891.
- Appelbaum, P. S., Jick, R. Z., Grisso, T., Givelber, D., Silver, E., & Steadman, H. J. (1993). Use of posttraumatic stress disorder to support an insanity defense. *The American Journal of Psychiatry*, 150(2), 229.
<http://europepmc.org/abstract/MED/8422072/reload=0;jsessionid=l2h4Z6WTifluElx44IVu.16>
- Ardolf, B. R., Denney, R. L., & Houston, C. M. (2007). Base rates of negative response bias and malingered neurocognitive dysfunction among criminal defendants referred for neuropsychological evaluation. *The Clinical Neuropsychologist*, 21, 899-916. doi: 10.1080/13825580600966391.
- Blandón-Gitlin, I., Pezdek, K., Lindsay, D.S., & Hagen, L. (2009). Criteria-based Content Analysis of True and Suggested Accounts of Events. *Applied Cognitive Psychology*, 23, 901–917. doi: 10.1002/acp.1504.
- Bond, C.F Jr., & DePaulo, B.M. (2006). Accuracy of deceptive judgements. *Personality and Social Psychology Review*, 10, 214 – 234. doi: 10.1207/s115327957pspr1003_2.
- DePaulo, B.M., Malone, B., Lindsay, J., Muhlenbruck, L., Charlton, K. & Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, 129 (1), 74 – 118. doi: 10.1037/0033-2909.129.1.74.
- Diagnostic and Statistical Manual 5 (2013). American Psychiatric Association.

- DeClue, G. (2011). Harry Potter and the structured interview of reported symptoms? *Journal of Forensic Psychology*, 3, 1-17.
http://www.forensicpsychologyunbound.ws/OAJFP/Forensic_Instruments_files/DeClue%202011.pdf
- Drob, S.S, Meehan, K.B & Waxman, S (2009). Clinical and conceptual problems in the attribution of malingering in forensic evaluations. *J Am Psychiatry & Law*, 37, 98-106.
doi: <http://www.jaapl.org/content/37/1/98.full.pdf+html>.
- Edens, J., Otto, R., Buffington, J., Tomicic, T., & Poythress, N. (2001) Factors differentiating successful versus unsuccessful malingerers. *Journal of Personality Assessment*, 77, 333-338. doi: 10.1207/S15327752JPA7702_13.
- Ekman, P., O'Sullivan, M., & Frank, M. G. (1999). A few can catch a liar. *Psychological Science*, 10, 263-266. doi: 10.1111/1467-9280.00147.
- Faust, D. (2012). Coping with psychological and psychiatric testimony. 6th Ed.. OUP. New York.
- Frederick, R.I. (2012). Malingering/Cooperation/Effort. Chapter 10 in Faust, D., 2012, Coping with psychological and psychiatric testimony. 6th Ed.. OUP. New York.
- Ferrari, R., Kwan, O., Russell, A., Pearce, J., Schrader, H. (1999). The best approach to the problem of whiplash? One ticket to Lithuania please. *Clinical and Experimental Rheumatology*, 17, 321-326. <http://www.ncbi.nlm.nih.gov/pubmed/10410265>
- Gödert, H.W., Gamer, M., Rill, H.G. & Vossel, G. (2005). Statement Validity Assessment: Inter-rater reliability of Criteria-Based Content Analysis in the mock-crime paradigm. *Legal and Criminological Psychology*, 10, 225 – 245.
- Greve, K. W., Ord, J. S., Bianchini, K. J., & Curtis, K. L. (2009). The prevalence of malingering in chronic pain patients referred for psychological evaluation in a medico-legal context. *Archives of Physical Medicine & Rehabilitation*, 90, 1117-1126. doi: 10.1016/j.apmr.2009.01.018.
- Hall, R., & Hall, R. (2006). Malingering of PTSD: forensic and diagnostic considerations, characteristics of malingerers and clinical presentations. *General Hospital Psychiatry*, 28 (6), 525-535. doi: 10.1016/j.genhosppsy.2006.08.011.

- Horowitz, S. W., Lamb, M. E., Esplin, P. W., Boychuk, T. D., Krispin, O., & Reiter-Lavery, L. (1997). Reliability of criteria-based content analysis of child witness statements. *Legal and Criminological Psychology*, 2, 11-21. doi: 10.1111/j.2044-8333.1997.tb00329.x.
- Johnson, J., & Lesniak-Karpiak, K. (1997). The effect of warning on malingering on memory and motor tasks in college samples. *Archives of Clinical Neuropsychology*, 12, 231-238. doi: 10.1093/arclin/12.3.231.
- Johnson, M.K., and Raye, C.L. (1981). Reality Monitoring. *Psychological Review*, 88, 67-85. doi: <http://doi.apa.org/index.cfm?fa=search.displayRecord&uid=1981-06694-001>.
- Kramer, G.M., & Gagliardi, G.J. (2009). Forensic Evaluation of Insanity: Assessing Valid Symptom Report in Defendants with Major Mental Disorder. *Journal of Forensic Psychology Practice* 9(1), 92 – 102. doi: 10.1080/15228930802427171.
- Lanyon, R. (1997) Detecting Deception: Current models and directions. *Clinical Psychology: Science and Practice*, 4, 377-387. doi: 10.1111/j.1468-2850.1997.tb00128.x.
- Larrabee, G. J. (2007). *Assessment of malingered neuropsychological deficits*. Oxford University Press.
- Levine, T. R., Feeley, T. H., McCornack, S. A., Hughes, M., & Harms, C. M. (2005). Testing the effects of nonverbal behavior training on accuracy in deception detection with the inclusion of a bogus training control group. *Western Journal of Communication*, 69(3), 203-217. doi:10.1080/10570310500202355
- Maguire, B., Harvey, A., & Shores, E. (2001). Simulated malingering in pain patients: a study with the pain patient profile. *British Journal of Clinical Psychology*, 40, 71-79. doi: 10.1348/014466501163490.
- Masip, J., Sporer, S. L., Garrido, E., & Herrero, C. (2005). The detection of deception with the reality monitoring approach: A review of the empirical evidence. *Psychology, Crime and Law*, 11, 99–122. doi:10.1080/10683160410001726356.
- Melzack, R. (1987). The short-form McGill Pain Questionnaire. *Pain*, 30, 191-197. doi: 10.1016/0304-3959(87)91074-8.
- Miller, H.A. (2001). *Miller-Forensic Assessment of Symptoms Test (M-FAST): Professional manual*. Odessa, FL: Psychological Assessment Resources.

- Miller, H.A. (2004). Examining the use of the M-FAST with criminal defendants incompetent to stand trial. *International Journal of Offender Therapy and Comparative Criminology*, 48 (3), 268-280.
- Porter, S., ten Brinke, L., & Wallace, B. (2012). Secrets and lies: Involuntary leakage in deceptive facial expressions as a function of emotional intensity. *Journal of Nonverbal Behavior*, 36(1), 23-37.
- Resnick, P. J (1995). Guidelines for the evaluation of malingering in posttraumatic stress disorder. In R. I. Simon (Ed.). *Posttraumatic stress disorder in litigation: Guidelines for forensic assessment* (pp. 117-134). Washington: American Psychiatric Press.
- Resnick, P. J. (1997). Malingered psychosis. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (2nd ed.) (pp. 47–67). New York: Guilford Press.
- Rogers, R. (1997) *Clinical assessment of Malingering Deception*. New York: Guilford.
- Rogers, R. (2008). Detection strategies for malingering and defensiveness. In R. Rogers (Ed.), *Clinical Assessment of Malingering and Deception*, (3rd ed., pp. 14-38). New York, NY: Guilford.
- Rogers, R., Bagby, R. M., & Dickens, S. E. (1992). *Structured Interview of Reported Symptoms* (SIRS), Professional manual. Lutz, FL: Psychological Assessment Resources, Inc.
- Rogers, R., Sewell, K. W., & Gillard, N. D. (2010). *Structured Interview of Reported Symptoms* (SIRS), 2nd Edition, professional manual. Lutz, FL: Psychological Assessment Resources, Inc.
- Sporer, S. (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.
- Steller, M., & Köhnken, G. (1989). Criteria-Based Content Analysis. In D. C. Raskin (Ed.), *Psychological methods in criminal investigation and evidence* (217-245). New York, NJ: Springer-Verlag.
- 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.

- Sweet, J. J., Wolfe, P., Sattlberger, E., Numan, B., Rosenfeld, J. P., Clingerman, S., & Nies, K. J. (2000). Further investigation of traumatic brain injury versus insufficient effort with the California Verbal Learning Test. *Archives of Clinical Neuropsychology*, *15*(2), 105-113. doi: 10.1016/S0887-6177(98)00153-X.
- Vrij, A. (2005). Criteria-Based Content Analysis: A qualitative review of the first 37 studies. *Psychology, Public Policy, and Law*, *11*, 3-41. doi: 10.1037/1076-8971.11.1.3.
- Vrij, A. (2008). *Detecting lies and deceit: pitfalls and opportunities*. Chichester: Wiley.
- Vrij, A., Akehurst, L., Brown, L., & Mann, S. (2006). Detecting lies in young children, adolescents and adults. *Applied Cognitive Psychology*, *20*, 1225 – 1237. doi: 10.1002/acp.1278.
- Vrij, A., Akehurst, L., & Knight, S. (2006). Police officers', social workers', teachers' and general public's beliefs about deception in children, adolescents and adults. *Legal and Criminological Psychology*, *11*, 297 – 312. doi: 10.1348/135532505X60816.
- Vrij, A., Akehurst, L., Soukara, S., & Bull, R. (2002). Will the truth come out? The effect of deception, age, status, coaching, and social skills on CBCA scores. *Law and Human Behavior*, *26*, 261-283. doi: 10.1023/A:1015313120905.
- Vrij, A., Akehurst, L., Soukara, S., & Bull, R. (2004a). Detecting deceit via analyses of verbal and nonverbal behavior in children and adults. *Human Communication Research*, *30* (1), 8 – 41. doi: 10.1111/j.1468-2958.2004.tb00723.x.
- Vrij, A., Akehurst, L., Soukara, S., & Bull, R. (2004b). Let me inform you about 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., & Easton, S. (2002). Fact or fiction? Verbal and behavioural clues to detect deception. *Medico Legal Journal*, *70*, 29-37. doi: 10.1258/rsmmlj.70.1.29.
- Vrij, A., Edward, K., & Bull, R. (2001). Stereotypical verbal and nonverbal responses while deceiving others. *Personality and Social Psychology Bulletin*, *27*, 899-909. doi: 10.1177/0146167201277012.

- Vrij, A., Granhag, P. A., Mann, S., & Leal, S. (2011). Outsmarting the liars: Towards a cognitive lie detection approach. *Current Directions in Psychological Science*, 20, 28-32. doi: 10.1177/1529100610390861.
- Vrij, A., Granhag, P. A., & Porter, S. B. (2010). Pitfalls and opportunities in nonverbal and verbal lie detection. *Psychological Science in the Public Interest*, 11, 89-121. doi: 10.1177/1529100610390861.
- Vrij, A., Leal, S., Granhag, P. A., Mann, S., Fisher, R. P., Hillman, J., & Sperry, K. (2009). Outsmarting the liars: The benefit of asking unanticipated questions. *Law and Human Behavior*, 33, 159-166. doi: 10.1007/s10979-008-9143-y.
- Youngjohn, J. R., Lees-Haley, P. R., Binder, L. M. (1999). Comment: Warning malingerers produces more sophisticated malingering. *Archives of Clinical Neuropsychology*, 14 (6), 511-515. doi: 10.1016/S0887-6177(98)00049-3.

Appendix 1: The checklist items and their sources

Item	Item source
1. If the individual pauses when describing symptoms, do the pauses last longer than when answering other questions?	DePaulo et al. (2003) Vrij (2008)
2. Does the individual show long latency periods i.e. the time between the end of a question and the beginning of an answer?	DePaulo et al. (2003) Vrij (2008)
3. Does the individual, when talking about the cold plunge test and subsequent experience talk in a disorganised way rather than speak in a structured and chronological order?	CBCA criterion 2
4. Does the individual, when talking about the cold plunge test and subsequent experience give substantial quantity of detail?	CBCA criterion 3
5. Does the individual, when talking about the cold plunge test and subsequent experience, describe unusual detail and/or unexpected complications?	CBCA criteria 7 and 8
6. Does the individual, when talking about the cold plunge test and subsequent experience, provide superfluous details (e.g. description of details which are not really relevant to answering the questions)?	CBCA criterion 9
7. Does the individual, when talking about the cold plunge test and subsequent experience, use terms and language or show knowledge related to their symptoms that they would not be expected to have?	CBCA criterion 10
8. Does the individual, when talking about the cold plunge test and subsequent experience, provide accounts of their own mental state?	CBCA criterion 12
9. Does the individual, when talking about the cold plunge test and subsequent experience, provide	CBCA criterion 14

<i>spontaneous corrections?</i>	
10. Does the individual, when talking about the cold plunge test and subsequent experience, tend to <i>admit to a lack of memory</i> ?	CBCA criterion 15
11. Does the individual, when talking about the cold plunge test and subsequent experience, <i>raise doubts about their own account</i> ?	CBCA criterion 16
12. Does the individual, when talking about the cold plunge test and subsequent experience, show <i>self-deprecation</i> (i.e. <i>put him or herself down</i>)?	CBCA criterion 17
13. Does the individual show <i>inappropriate behaviour</i> ?	Validity Checklist item 2
14. Does the individual give clear answers to questions about the symptoms?	Reality Monitoring criterion 1
15. Does the individual provide detail about <i>physical sensations</i> (e.g. "It felt like pins and needles")?	Reality Monitoring criterion 2
16. Does the individual provide <i>temporal details</i> ? (e.g. details about time order of the events; "the pins and needles came before the numb feeling")	Reality Monitoring criterion 4
17. Does the individual provide details about <i>duration of events</i> ? ("the numbness lasted for about 10 minutes.")	Reality Monitoring criterion 4
18. Does the individual give plausible answers to the questions about the symptoms?	Reality Monitoring criterion 7
19. Does the individual show cognitive operations: <i>inferences at the time of the event or at interview</i> ? (e.g. it appeared to me that she was enjoying me doing the ice plunge")	Reality Monitoring criterion 8
20. Is the individual falsely ascribing real symptoms to the cold plunge test (e.g. they had dry skin anyway and now blaming the cold plunge)?	Resnick (1997)

21. Does the individual show lack of co-operation at interview?	Resnick (1997)
22. Does the individual blame someone else for their symptoms?	Ferrari et al. (1999)
23. Does the individual believe that the other person behaved recklessly, so causing their discomfort?	Ferrari et al. (1999)
24. Does the individual expect their discomfort to last a short time and/or show significant improvement?	Ferrari et al. (1999)
25. Does the individual tend to confirm existence of many symptoms?	Rogers (1997)
26. Are there symptoms which do not readily fit those expected with a cold plunge test?	Rogers (1997) Validity Checklist item 1
27. Are there unlikely or contradictory patterns of symptoms?	Rogers (1997) Validity Checklist item 1
28. Are some symptoms reported as at extreme, improbable levels of severity?	Rogers (1997)

CBCA; Criteria Based Content Analysis criteria (from Steller & Köhnken, 1989). SVA; Validity Checklist Items (from Steller & Köhnken, 1989). RM; Reality Monitoring (from Vrij, 2008). Others as identified in reference section.

Table 1: Criteria Based Content Analysis criteria (from Steller & Köhnken, 1989)

CBCA Category	CBCA Criteria	Inclusion in Checklist for Current Study	Reason for Exclusion
<i>General Characteristics</i>	1. Logical structure	X	Relates to the recall of an event, not symptoms.
	2. Unstructured production	√	
	3. Quantity of details	√	
<i>Specific Contents</i>	4. Contextual embedding	X	Relates to the recall of an event, not symptoms.
	5. Descriptions of interactions	X	Relates to the recall of an event, not symptoms.
	6. Reproduction of conversation	X	Relates to the recall of an event, not symptoms.
	7. Unexpected complications during the incident	√	
	8. Unusual details	√	
	9. Superfluous details	√	
	10. Accurately reported details misunderstood	√	
	11. Related external associations	X	
			Relates to the recall of an event, not symptoms.

	12. Accounts of subjective mental state	√	
	13. Attribution of perpetrator's mental state	X	Not relevant in medico-legal setting.
<i>Motivation-Related Contents</i>	14. Spontaneous corrections	√	
	15. Admitting lack of memory	√	
	16. Raising doubts about one's own testimony	√	
	17. Self-deprecation	√	
	18. Pardoning the perpetrator	√	
<i>Details characteristic of the offence</i>	19. Details specific to a type of crime	X	Not relevant in medico-legal setting.

Table 2: Validity Checklist Items (from Steller & Köhnken, 1989)

Validity Checklist Category	Validity Checklist Criteria	Inclusion in Checklist for Current Study	Reason for Exclusion
<i>Psychological Characteristics</i>	1. Inappropriateness of language and knowledge	√	Not relevant in medico-legal setting.
	2. Inappropriateness of behaviour	√	
	3. Susceptibility to suggestion	X	
<i>Interview Characteristics</i>	4. Suggestive, leading, or coercive questioning	X	Not relevant in medico-legal setting.
	5. Overall inadequacy of the interview	X	Not relevant in medico-legal setting.
<i>Motivation</i>	6. Questionable motives to report	X	Not relevant in lab setting.
	7. Questionable context of the original disclosure or report	X	Not relevant in lab setting.
	8. Pressures to report falsely	X	Not relevant in lab setting.
<i>Investigative Questions</i>	9. Inconsistency with the laws of nature	X	Not relevant in lab setting.
	10. Inconsistency with other interviewee's description of	X	Not relevant in lab setting.

	difficulties	X	
	11. Inconsistency with other evidence		

Table 3: Reality Monitoring items (from Vrij, 2008).

Reality Monitoring Criteria	Inclusion in Checklist for Current Study	Reason for Exclusion
1. Clarity	√	
2. Perceptual information	√	
3. Spatial information	X	Relates to the recall of an event, not symptoms.
4. Temporal information	√	
5. Affect	X	Already included under (In)appropriateness of Behaviour – Validity Checklist item 2.
6. Reconstructability of story	X	Relates to the recall of an event, not symptoms.
7. Realism	√	
8. Cognitive operations	√	

Table 4. Frequency of hits, misses, false alarms and correct rejections for interviewers with and without the checklist

	Hit	Miss	False alarm	Correct rejection	Total
With checklist	21	11	8	24	64
Without checklist	18	14	13	19	64
Total	40	25	21	42	128

Table 5. Means, univariate F-values, pooled within-group correlations between checklist items and canonical discriminant functions and Cohen's d scores for difference between truthful and exaggerated interviewees for the significant checklist items and total checklist score.

Checklist item	M truthful (N = 64)	M exaggerated (N = 64)	F (1,28)	Canonical r	d score
Are some symptoms reported as at improbable levels of severity?	2.32	3.45	10.542	.582	0.83
Does the individual tend to confirm the existence of many symptoms?	3.19	4.13	10.391	.564	0.82
Are there symptoms that do not readily fit those expected?	1.84	2.81	10.007	.525	0.81
^Does the individual expect his discomfort to last a short time or show a significant improvement?	2.26	3.00	6.404	-.448	0.71
Does the individual believe that someone else behaved recklessly?	1.07	1.48	6.326	.445	0.64
Checklist total	70.16	75.56	5.782	.407	0.61
^Does the individual provide accounts of his own mental state?	3.13	3.97	5.342	-.391	0.59
^Does the individual include unusual or unexpected details?	2.84	3.58	5.032	-.369	0.58
Does the individual blame someone else?	1.16	1.52	4.539	.359	0.58
Are there unlikely or contradictory patterns of symptoms?	1.74	2.39	4.515	.340	0.54

* $p < 0.05$; ** $p < 0.01$

^ expected more often in truthful compared to exaggerated accounts so 1 = very much present and 5 = not at all present

For all other items 1 = not at all present and 5 = very much present. In sum, a higher score always depicts exaggeration rather than truth-telling

Table 6. Means, univariate F-values, pooled within-group correlations between checklist items and canonical discriminant functions and Cohen's d scores for difference between judgements of 'truthful' or 'exaggerated' for the significant checklist items and total checklist score.

Checklist item	M truthful (N = 64)	M exaggerated (N = 64)	F (1,28)	Canonical r	d score
Are there symptoms that do not readily fit those expected?	1.66	3.03	24.92	.594	1.25
^Does the individual include unusual or unexpected details?	2.59	3.83	16.50	.483	1.03
^Does the individual give plausible answers to the questions about symptoms?	1.77	2.69	16.46	.483	1.03
Are some symptoms reported as at improbable levels of severity?	2.31	3.59	14.65	.455	0.96
Are there unlikely or contradictory patterns of symptoms?	1.57	2.62	14.36	.451	0.95
Checklist total	69.46	76.90	12.83	.426	0.91
Does the individual tend to confirm the existence of many symptoms?	3.14	4.17	12.65	.423	0.91
Does the individual believe that someone else behaved recklessly?	1.06	1.52	9.00	.357	0.83
Does the individual wait a long time between the end of a question and the beginning of his answer?	1.63	2.34	8.59	.349	0.73
^Does the individual give clear	2.38	3.23	6.51	.303	0.65

answers to questions about symptoms?					
Does the individual blame someone else?	1.14	1.55	6.46	.302	0.65
^Does the individual expect his discomfort to last a short time or show a significant improvement?	2.31	3.00	5.67	.283	0.60
Does the individual pause for longer during answers about symptoms compared to other answers?	1.74	2.28	5.13	.269	0.57

* $p < 0.05$; ** $p < 0.01$

^ expected more often in truthful compared to exaggerated accounts so 1 = very much present and 5 = not at all present

For all other items 1 = not at all present and 5 = very much present