

Rapid judgements in Assessing Verbal and Nonverbal Cues:
Their Potential for Deception Researchers and Lie Detection

Aldert Vrij¹²

Hayley Evans

Lucy Akehurst

Samantha Mann

University of Portsmouth

Psychology Department

Running head: rapid judgements

¹ Correspondence concerning this article should be addressed to: Aldert Vrij, University of Portsmouth, Psychology Department, King Henry Building, King Henry 1 Street, Portsmouth PO1 2DY, United Kingdom or via email: aldert.vrij@port.ac.uk

² This project was sponsored by a grant from the Nuffield Foundation (grant URB/00689/G). Stimulus materials used in this study were derived from a project sponsored by the Economic and Social Research Council (grant R000222820).

Abstract

In the present study it was investigated to what extent observers (i) could make rapid yet reliable and valid judgements of the frequency of verbal and nonverbal behaviours of interviewees (liars and truth tellers) and (ii) detect deceit after making these rapid judgements. Five observers watched 52 videoclips of 26 liars and 26 truth tellers. The findings revealed that rapid judgements were reliable and valid. They also revealed that observers were able to detect truths and lies well above the level of chance after making these rapid judgements (74% accuracy rate was found). The implications of these findings for deception researchers and lie detection are discussed.

Rapid judgements in Assessing Verbal and Nonverbal Cues:
Their Potential for Deception Researchers and Lie Detection

Research has demonstrated that lie detection is a difficult task during which incorrect judgements are commonly made. In typical scientific lie detection studies, observers are given videotapes and asked to judge whether each of a number of people is lying or telling the truth. In the vast majority of these studies (see Vrij, 2000, 2002, for reviews) the accuracy rates (percentages of correct lie and truth detection) varied between 45% and 60%, regardless of whether the observers were laypersons (typically university students) or professional lie catchers, such as police officers (although some groups of professional lie catchers (e.g. CIA agents) are more accurate (Ekman & O'Sullivan; 1991; Ekman, O'Sullivan, & Frank, 1999)).

Research has also provided evidence that people become better lie detectors when they conduct detailed analyses of "diagnostic"¹ nonverbal and verbal cues displayed by truth tellers and liars. For example, Ekman, O'Sullivan, Friesen, & Scherer (1991) analysed liars' and truth tellers' smiles and pitch of voice, and could correctly classify 86% of liars and truth tellers on the basis of these measurements. Frank and Ekman (1997) examined signs of emotions which emerged via (micro) facial expressions, and could correctly classify around 80% of liars and truth tellers on the basis of these facial expressions. Also, on the basis of verbal detection tools such as Criteria-Based Content Analysis (CBCA) (Köhnken & Steller, 1988; Raskin & Esplin, 1991; Steller, 1989; Steller & Köhnken, 1989) or Reality Monitoring (Alonso-Quecuty, 1992, 1996; Johnson & Raye, 1981; Sporer, 1997; Vrij, Akehurst, Soukara, & Bull, in press; Vrij, Edward, Roberts, & Bull, 2000), around 70% of truths and lies can be correctly classified (see Vrij, 2000, for reviews of CBCA and Reality Monitoring research).

However, scoring nonverbal and verbal behaviours is a time-consuming exercise. In order to accurately score the frequency of occurrence of one single category of nonverbal behaviour (for example, trunk movements), observers have to watch a videotape several times, and often have to watch parts of the videotape in slow motion. This process is then repeated when observers move on to score the frequency of occurrence of a second behavioral cue (for

example, head movements). Scoring other aspects of nonverbal behaviour, such as pitch of voice, may even require sophisticated equipment (Ekman, Friesen, & Scherer, 1976).

Scoring verbal behaviours is equally time consuming. CBCA assessments require written transcripts of statements. Therefore, accounts need to be described, from audio or videotape, and it is necessary to read them several times before accurate CBCA ratings can be made. Scoring verbal criteria which are included in the Reality Monitoring list (again from written transcripts) is also time-consuming, although less time-consuming than CBCA coding (Sporer, 1997; Vrij et al., 2000, in press).

The first aim of the present study was to examine whether accurate estimates of the frequency of occurrence of a range of diagnostic verbal and nonverbal behaviours can be made on the basis of quick global assessments ("rapid judgements"). In the present study, observers were shown videotaped statements of 52 liars and truth tellers. We investigated to what extent observers agreed amongst each other, after rapid judgements, regarding the frequency of occurrence of a range of verbal and nonverbal cues displayed by the liars and truth tellers (reliability), and to what extent these rapid judgments accurately reflected the actual frequency of occurrence of the verbal and nonverbal cues displayed (validity). We predicted that these rapid judgements would be reliable and valid. We based this prediction on the findings of several training studies which revealed that asking people to pay attention to some diagnostic cues to deception (both verbal and nonverbal cues) does increase their ability to detect deceit. See Vrij (2000) for a review of training studies, and see Porter, Woodworth, and Birt (2000) for an example of a very successful training study. Obviously, a training effect can only be obtained if observers are capable of spotting the cues they are asked to look for.

The second aim of the study was to determine whether observers would be able to detect deceit after they have made their rapid judgements regarding the frequency of occurrence of several diagnostic verbal and nonverbal behaviours. We predicted that they would. Teaching observers how to score a variety of diagnostic verbal and nonverbal cues and informing these observers how these cues are related to deception is, in fact, training observers how to detect deceit, and research has shown that people do become better lie detectors when they are

instructed to look at diagnostic cues (see above).

In addition, by asking observers to primarily concentrate on scoring the frequency of verbal and nonverbal cues rather than to attempt to detect lies, we, in fact, encouraged them to detect lies implicitly, which has been shown to be more successful than explicit lie detection (DePaulo, Anderson, & Cooper, 1999; Vrij, 2001). For example, in Vrij, Edward, and Bull's (2001b) study, police officers watched videotapes of truth tellers and liars. Some participants were asked whether each of the people were lying (direct lie detection method), others were asked to indicate for each person whether that person "had to think hard" (indirect lie detection method, they were not informed that some people were actually lying). The police officers' responses distinguished between truths and lies, but only by using the indirect method. When detecting deceit directly, police officers' judgements about deceit were significantly correlated with increases in gaze aversion and movements shown by the people on the videotape. In the indirect method, however, police officers' decisions were significantly correlated with a decrease in hand and finger movements. A decrease in hand and finger movements is a more diagnostic cue to deception than, for example, gaze aversion (DePaulo, Lindsay, Malone, Muhlenbruck, Charlton, & Cooper, 2003; Vrij, 2000). This suggests that by asking lie detectors to employ the indirect method, they are subtly directed to the more valid cues of deception. Method

Participants

Five observers, two males and three females aged 19 - 21 participated in the study. They were all undergraduate students, and were not acquainted with the undergraduate students that were used to make up the stimulus material.

Stimulus Material

The stimulus material, videotaped interviews with 26 liars and 26 truth tellers, was derived from a previous experiment (Vrij et al., in press). In that study, 196 participants from different age groups participated, including 52 adults (college students). The interviews with these 52 adults were used as stimulus material in the present study. These 52 adults lied or told the truth about playing a game of Connect 4 with a confederate and rubbing a maths formula from the blackboard. In order to motivate the adults, they were promised £5 if they were able to

tell a "convincing story" and were threatened that they had to write an essay in case their story was not convincing. All 52 adults told a convincing story and received £5. The average length of the deceptive and truthful interviews were $M = 125.5$ seconds ($SD = 48.7$) and $M = 161.4$ seconds ($SD = 43.4$) respectively. The difference in length between the truthful and deceptive interviews was significant, $F(1, 50) = 7.86$, $p < .01$. See Vrij et al. (in press) for more details about this deception task.

Verbal and Nonverbal Behaviours Used in the Rapid Judgement Task

The verbal and nonverbal behaviours used in the rapid judgement task were selected on the basis of the findings of two of our previous studies (Vrij et al., 2000, in press). In Vrij et al.'s (2000) experiment, 73 nursing students either lied ($N = 39$) or told the truth ($N = 34$) about a film they had just seen which depicted the theft of a handbag in a hospital. Vrij et al. (in press) is already discussed in the Stimulus Material section above.

In Vrij et al. (2000, in press) experiments, detailed coding of a range of nonverbal and verbal behaviours took place on the basis of coding systems used by us before (Vrij, Semin, & Bull, 1996; Vrij, Edward, & Bull, 2001a, c). Differences between liars and truth tellers regarding these variables were examined and Table 1 provides an overview of the findings.

Table 1 about here

The 12 cues indicated with an asterisk (*) are included in the rapid judgement task. With the exception of latency period and speech hesitations, all the selected cues revealed significant differences between liars and truth tellers in both data sets. Latency period and speech hesitations were added to increase the number of nonverbal judgements.² Most of the selected cues also emerged as indicators of deception in several reviews and meta-analyses concerning cues to deception (DePaulo et al., 2003; Vrij, 2000).

Definitions of the Twelve Variables

(1) latency period: period of time between the question being asked and the answer being given; (2) hand and finger movements: movements of the hands or fingers without moving the arms; (3) speech hesitations: saying 'ah' or 'mm' between words; (4) quantity of details: specific descriptions of place, time, persons, objects and events; (5) contextual embeddings: descriptions

of time and location (e.g. "He was sitting on a bench during lunch time"); (6) reproduction of conversation: speech reported in its original form; (7) description of other's mental state: description of other people's feelings, thoughts or motives (e.g. "He looked really scared"); (8) visual details: description of details which the interviewee saw (e.g. "He wore a red shirt"); (9) auditory details: description of details which the interviewee heard: "He knocked loudly at the door", (10) spatial information: information about locations and about how objects were related to each other (e.g. "And then the pieces of Connect 4 fell on to the floor"); (11) temporal details: information about time and duration of events (e.g. "We kept on playing for a while"); (12) cognitive operations: thoughts and reasonings ("Because she was quite clever, she won the game").

Training

First, a research assistant (an undergraduate psychology student) read some relevant book chapters regarding the twelve verbal and nonverbal cues under investigation. She then received training concerning the twelve cues by the first author. In the training session examples of the twelve cues were given. It was also explained how the variables were related to deception³. When the research assistant felt that she understood the meaning of the cues and how to rate them, both trainer and trainee watched an example videofragment of an interviewee (examples were derived from Vrij et al., 2000), and independently from each other made rapid judgements regarding the occurrence of the three nonverbal behaviours (latency period, speech hesitations, and hand and finger movements). They then watched the same fragment again and made judgements concerning the nine verbal behaviours. All rapid judgements were given on 5-point Likert scales ranging from (1) absent to (5) very much present. After completing these twelve judgements, the raters compared their ratings. "Substantial differences" between the two raters, that is, a difference of more than 1 point on the 5-point scale, were resolved by discussion, often after watching the fragment again. The final ratings were used as "anchor scores" in the second training session with the remaining observers (see below). After watching and rating five example interviews both raters felt confident about their judgements and felt that watching further examples was not necessary.

Subsequently, the research assistant held a training session with the four remaining observers (the research assistant herself was also an observer in the study). This training session was similar to that described above. The five observers watched the same five example interviews (i.e. those that the first author had used to train the research assistant). After watching each example videofragment, the judgement ratings were compared and discussed. During these discussions, the research assistant revealed the anchor scores (the agreed ratings between herself and the first author) and the four other observers were asked to use these ratings as guidance. After approximately ninety minutes of training in which five example interviews were watched and discussed, all observers felt that they knew what was required of them and felt confident that they could accurately perform the task. Also at this stage, the research assistant felt confident that there was sufficient agreement between all five observers in their rating.

Each of the 52 clips was watched twice by each of the five observers (independent of one another). The three nonverbal judgements were made after the first viewing and the nine verbal judgements after the second viewing. After completing the twelve rapid judgements (given on 5-point Likert scales ranging from (1) absent to (5) very much present), the observers indicated whether or not they thought that the person was lying (dichotomous scale). All responses (rapid judgements and veracity judgements) were recorded on answer sheets. All responses were provided in silence and no comparisons between the observers were made at any time during the judgement task. The judgement task lasted 1.5 days (including breaks) and the observers were paid £100 for their efforts. All five observers were blind to the actual veracity of the statements, to the ratio of truthful and deceptive statements, and to the event that the persons on the video were talking about. Neither did they discuss any aspect of their judgement work during the task or during the breaks.⁴

Results

Reliability of Rapid Judgements

Table 2 about here

In order to examine the reliability of the rapid judgements, interrater agreement scores

(Cronbach's alphas) between the five raters were calculated. Results revealed satisfactorily agreement between the five observers for all variables except cognitive operations (see the first column of Table 2). Combining the scores for the five observers is therefore justified for all variables, except cognitive operations. Due to the low interrater agreement score for cognitive operations, this variable was ignored in most of the further analyses.

Validity of Rapid Judgements

Validity of the rapid judgements was tested in two ways. First, Pearson correlations were conducted between the rapid judgements of the observers and the actual frequency of occurrence of the verbal and nonverbal criteria in the statements (the actual frequencies were calculated in Vrij et al.'s, in press, study). Pearson correlations were computed for the five observers individually and the scores of the five observers combined (see Table 2).⁵ The results revealed rather high correlations between the rapid judgements of the criteria (combined scores of the five observers) and the actual frequency of occurrence of these criteria (see last column of Table 2). All these correlations were significant. Results for each individual observer (Table 2, columns 2 to 6) showed that, in general, high positive correlations were found for each individual observer. However, not all correlations were significant.

Table 3 about here

Second, ANOVAs were conducted with the verbal and nonverbal cues as dependent variables and the veracity of the statement as independent variable. On the basis of actual frequency scoring (Table 3, left half) significant differences were found between truth tellers and liars for hand and finger movements, number of details, contextual embedding, reproduction of conversation, descriptions of other's mental state, visual details, sound details, spatial details and temporal details. ANOVAs regarding the rapid judgements (five observers combined, Table 3, right half) revealed the same significant differences as were found with the actual frequency data, except for hand and finger movements and descriptions of other's mental state.⁶

These correlational and ANOVA findings combined revealed that observers are able to make reliable and valid rapid judgements of verbal and nonverbal behaviours.

Accuracy Rates of Observers

Table 4 about here

The total accuracy rates of truths and lies (correct classifications of truth tellers and liars combined) was rather high at 74% (see Table 4), with an 82% accuracy rate for truths (correct classifications of truth tellers) and 65% accuracy rates for lies (correct classifications of liars).⁷ All three accuracy rates were significantly above the level of chance (50%) (all t -values > 3.09). The lie detection and truth detection rates did not differ significantly from each other, $F(1, 50) = 3.69$, *ns*. Total accuracy rates for the five individual observers (see Table 4) ranged from a modest 56% (Observer 4) to a high 85% (Observer 5). All observers, except Observer 4, performed significantly above the level of chance.⁸

Table 5 about here

We also looked at the relationships between rapid judgements and veracity judgements. Pearson correlations (for the scores of the five observers combined) and Spearman correlations (for each observer; Spearman correlations are appropriate because for each individual observer the veracity judgement was a dichotomous variable) were carried out between rapid judgements and the decision to classify the interviewee as a liar or truth teller.⁹ Regarding the judgements for the five observers combined, Table 5 reveals several significant correlations between veracity judgements and most rapid judgements. The correlation for number of details was the highest, with the fewer details mentioned by the interviewees, the more likely that the observers classified the interviewee as a liar. A regression analysis (with veracity judgement as criterion and the verbal and nonverbal cues that reached significant correlations with veracity judgements as dependent variables) revealed two predictors explaining 67% of the variance ($F(2, 49) = 50.90$, $p < .01$). As can be seen in Table 5 (last column) quantity of details was the strongest predictor of veracity judgements. None of the nonverbal behaviours emerged as a predictor in the regression analysis.

Results for each individual observer showed numerous significant correlations. Logistic regressions (appropriate because in the analyses per individual observer the veracity judgement was a dichotomous variable) revealed that number of details emerged most frequently (three times) as a predictor. Reproduction of conversation, visual details and cognitive operations each

appeared twice as a predictor. Again, more verbal than nonverbal behaviours emerged as predictors.¹⁰

Discussion

Reliable and Valid Rapid Judgements

In this study, observer's ability to make reliable and valid rapid judgements of verbal and nonverbal cues to deception was investigated. The findings revealed that they could. Generally, (i) there was good interrater agreement between the different observers and (ii) correlations between rapid judgements and actual frequency scoring were satisfactory; (iii) differences found between truth tellers and liars on the basis of actual frequency scoring were also found on the basis of rapid judgements; and (iv) observers could detect truths and lies after making rapid judgements.

There were some exceptions to these general findings. First, we failed to find a reliable interrater agreement score for cognitive operations. Perhaps our instructions to observers about cognitive operations were not detailed enough for them to fully comprehend the concept. Indeed, cognitive operations are not always easy to score. For example, do examples such as (i) "Her shoes looked quite big", (ii) "I think she wiped off the board", and (iii) "She was quite clever" contain cognitive operations? In our definition, examples one and two don't and example three does, but we realise that this may not be immediately obvious to all observers.

Second, although significant differences were found between liars and truth tellers regarding descriptions of other's mental state and hand and finger movements on the basis of actual frequency scoring, these effects were not significant on the basis of rapid judgements. In other words, some valuable information about cues to deception was lost by making rapid judgements. Results from frequency scoring revealed that references to other's mental state were rarely made (they appeared on average $M = .13$ per statement and appeared in only 15% of the statements). Regarding hand and finger movements, our findings showed that the rapid judgements of two out of five observers did accurately reflect that truth tellers made more of these movements than liars. Such movements, however, are typically very subtle and therefore hard to spot, which may explain the absence of significant effects for three observers.

Our findings are beneficial to deception researchers. Actual frequency coding is very time-consuming and therefore expensive. Making rapid judgements is therefore an attractive alternative and our findings revealed that such judgements really do reflect actual frequency scoring.

Lie Detection

The fact that four out of five observers were able to detect both truths and lies above the level of chance after making rapid judgements, makes the findings relevant for lie detection. It suggests that when observers are asked to count the frequency of a series of "diagnostic deception cues" they will be able to detect truths and lies above the level of chance. Although we did not actually test this, we believe that the frequency coding was crucial in the success obtained. We believe that merely informing observers prior to the assessment task how the verbal and nonverbal cues were related to deception (but not actually asking them to rate the frequency of occurrence of these cues) would not have led to the same results. The counting task probably absorbed each observer's full attention, and left him/her with no time to think about lie detection. This makes our assessment task an implicit lie detection task, demonstrated to be superior to explicit lie detection tasks. Future studies could test this hypothesis.

The accuracy rates found in this study (74% total accuracy) are relatively high, and higher than found in the vast majority of previous deception studies. The present accuracy rates are comparable to those obtained with groups of specialised lie detectors, such as CIA agents (Ekman et al., 1999), and comparable to accuracy rates which were obtained after an extensive 2-day workshop about deception (Porter et al., 2000). Unfortunately, one observer, Observer 4, failed to achieve high accuracy rates. Analyses revealed that Observer 4 achieved high accuracy rates (82% total rate) while judging the first one third of the clips (clip 1-17), but performed considerably worse during the remaining part of the task (35% accuracy rate for clips 18-34 and 50% accuracy rate for clips 35-52). This suggests that Observer 4 might have been prone to a "fatigue effect": judging 52 clips is cognitively tiring, and exhaustion might have impaired performance.

The regression analyses for the five observers individually, and also the regression

analysis for the five observers combined, revealed that observers were more guided by verbal criteria than by nonverbal criteria. On the one hand, this could simply be an order effect. Veracity judgements always directly followed the verbal rapid judgements, and that may have resulted in a larger impact of verbal rapid judgements on the veracity judgements. On the other hand, it might be a real effect. Verbal information is more meaningful than nonverbal information, that is, each verbal detail has a meaningful content, whereas each nonverbal behaviour has not. This probably makes verbal information more vivid than nonverbal information and therefore likely to have a stronger impact on observers (Nisbett, Borgida, Crandall, & Reed, 1976; Nisbett & Ross, 1980).

Alternative Reasons for High Accuracy Rates

We believe that we obtained high accuracy rates because we asked our observers to assess the frequency of occurrence of diagnostic verbal and nonverbal cues before we asked them to make their veracity judgements. However, we do realise that, in principle, other explanations are possible, but don't believe that any of these explanations are strong enough to challenge the reasoning outlined previously on this issue.

First, perhaps our five observers were particularly good lie detectors. There is no reason to assume that they were. They were ordinary undergraduate students and none of them has shown particular interest in deception research before. In other words, they were lie detectors highly comparable to the lie detectors used in typical deception studies with laypersons in which generally lower accuracy rates are obtained.

Second, perhaps our 26 liars were particularly poor liars. Again, there is no reason to believe they were. The 26 liars used in this study were a random sample of college students and highly representative for the liars typically used in other lie detection studies.

Third, observers saw each clip twice before they made their veracity judgements. Perhaps they did benefit from repeated exposure to the stimulus material. Research suggests that this is an unlikely explanation. In a series of lie detection studies, Mann (2001) asked observers to make veracity judgements after watching clips of liars and truth tellers once (Studies 3 and 4) or twice (Study 2). The three studies were highly comparable as the same stimulus material was used in

all three studies. The three studies revealed similar accuracy rates, indicating that repeated exposure had no effect on the accuracy rates.

Fourth, the fact that observers saw so many clips ($N = 52$) may have resulted in a "learning effect". Perhaps, after hearing numerous statements they may have worked out the facts of the staged event that could have facilitated lie detection. We found no evidence for a learning effect (see endnote 7). On the contrary, as mentioned before, there was some evidence that Observer 4 experienced a "fatigue effect" which had a negative impact on accuracy scores. In fact, the accuracy scores for the first 17 clips (one third of the total number of clips they saw) were exceptionally high across the five observers, with a 84% total accuracy score (89% truth accuracy and 78% lie accuracy).

Fifth, while making their veracity judgements, observers may have been influenced by an obvious difference between liars and truth tellers. For example, truthful statements were significantly longer than deceptive statements and observers may have been guided by the length of the statements. There is evidence that they did not do this. In none of the regression analyses which were carried out to examine by which cues the observers were influenced while making their veracity judgements length of speech emerged as a predictor.

Compared to other lie detection studies, the present study had one major advantage. The observers were exposed to a large number of clips ($N = 52$) which is a more valid test of examining people's lie detection skills than providing observers with only a limited number of clips. A disadvantage of such a comprehensive lie detection task is that only a few observers could be used. However, using a few observers also had an advantage. It enabled us to report analyses for each individual observer which is generally impossible (and never happens) when more observers are involved.

References

- Alonso-Quecuty, M. L. (1992). Deception detection and Reality Monitoring: A new answer to an old question? In F. Lösel, D. Bender, & T. Bliesener (Eds.), *Psychology and law: International perspectives* (pp. 328-332). Berlin: Walter de Gruyter.
- Alonso-Quecuty, M. L. (1996). Detecting fact from fallacy in child and adult witness accounts. In G. Davies, S. Lloyd-Bostock, M. McMurrin, & C. Wilson (Eds.), *Psychology, law, and criminal justice: International developments in research and practice* (pp. 74-80). Berlin: Walter de Gruyter.
- DePaulo, B. M., Anderson, D. E., & Cooper, H. (1999, October). *Explicit and implicit deception detection*. Paper presented at the Society of Experimental Social Psychologists, St. Louis.
- DePaulo, B. M., Lindsay, J. L., Malone, B. E., Muhlenbruck, L., Charlton, K. & Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, 129, 74-118.
- Ekman, P., Friesen, W. V., & Scherer, K. R. (1976). Body movement and voice pitch in deceptive interaction. *Semiotica*, 16, 23-27.
- Ekman, P., & O'Sullivan, M. (1991). Who can catch a liar? *American Psychologist*, 46, 913-920.
- Ekman, P., O'Sullivan, M., & Frank, M. G. (1999). A few can catch a liar. *Psychological Science*, 10, 263-266.
- Ekman, P., O'Sullivan, M., Friesen, W. V., & Scherer, K. (1991). Face, voice, and body in detecting deceit. *Journal of Nonverbal Behavior*, 15, 125-135.
- Frank, M. G., & Ekman, P. (1997). The ability to detect deceit generalizes across different types of high-stake lies. *Journal of Personality and Social Psychology*, 72, 1429-1439.
- Johnson, M. K., & Raye, C. L. (1981). Reality Monitoring. *Psychological Review*, 88, 67-85.
- Köhnken, G., & Steller, M. (1988). The evaluation of the credibility of child witness statements in German procedural system. In G. Davies & J. Drinkwater (Eds.), *The child witness: Do the courts abuse children?* (Issues in Criminological and Legal Psychology, no. 13) (pp. 37-45). Leicester, United Kingdom: British Psychological Society.

Mann, S. (2001). *Suspects, lies and videotape: An investigation into telling and detecting lies in police / suspect interviews*. Unpublished PhD-thesis, University of Portsmouth, Psychology Department.

Nisbett, R. E., Borgida, E., Crandall, R., & Reed, H. (1976). Popular induction: Information is not always informative. In J. S. Carroll & J. W. Payne (Eds.), *Cognition and social behavior, volume 2* (pp. 227-236). Hillsdale, NJ: Erlbaum.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall.

Porter, S., Woodworth, M., & Birt, A. R. (2000). Truth, lies, and videotape: An investigation of the ability of federal parole officers to detect deception. *Law and Human Behavior, 24*, 643-658.

Raskin, D. C., & P. W. Esplin (1991). Assessment of children's statements of sexual abuse. In J. Doris (Ed.), *The suggestibility of children's recollections* (pp. 153-165). Washington DC: American Psychological Association.

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.

Steller, M. (1989). Recent developments in statement analysis. In J. C. Yuille (1989). *Credibility Assessment* (pp. 135-154). Deventer, the Netherlands: Kluwer.

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

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

Vrij, A. (2001). Implicit lie detection. *The Psychologist, 14*, 58-60.

Vrij, A. (2002, September). *Telling and detecting true lies: Investigating and detecting the lies of murderers and thieves during police interviews*. Paper presented at the Twelfth European Conference of Psychology and Law, Katholieke Universiteit Leuven, Faculty of Law,

Leuven, Belgium, September 14-18.

Vrij, A., Akehurst, L. Soukara, S., & Bull, R. (in press). Detecting deceit via analyses of verbal and nonverbal behavior in children and adults. *Human Communication Research*.

Vrij, A., Edward, K., & Bull, R. (2001a). People's insight into their own behaviour and speech content while lying. *British Journal of Psychology*, 92, 373-389.

Vrij, A., Edward, K., & Bull, R. (2001b). Police officers' ability to detect deceit: The benefit of indirect deception detection measures. *Legal and Criminological Psychology*, 6, 2, 185-197.

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

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

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

Table 1.
Schematic representation of differences in nonverbal and verbal behavior between liars and truth tellers in Vrij et al. (2000, in press).

	Vrij et al. (2000)	Vrij et al. (in press)
<u>nonverbal behaviour</u>		
gaze aversion	-	-
self manipulations	-	-
illustrators	<	-
hand/finger movements*	<	<
foot/leg movements	-	-
latency period*	>	-
speech rate	-	-
speech hesitations*	>	-
speech errors	-	-
<u>Criteria-Based Content Analysis criteria</u>		
logical structure	-	-
quantity of details*	<	<
contextual embedding*	<	<
description of interactions	-	<
reproduction of conversation*	<	<
own mental state	<	-
other's mental state*	<	>
spontaneous corrections	<	<
admitting lack of memory	-	-
raising doubts about memory	-	-
<u>Reality Monitoring criteria</u>		
visual details*	<	<
auditory details*	<	<
spatial details*	<	<
temporal details*	<	<
cognitive operations*	<	>

- < liars displayed the cue significantly less than truth tellers
- > liars displayed the cue significantly more than truth tellers
- no difference between liars and truth tellers
- * cues selected for the rapid judgement task

Table 2.

(i) Interrater agreement scores between the five raters (Cronbach's alpha), and (ii) Pearson correlations between rapid judgements and actual frequency scoring for the five observers separately and the five observers combined

	Cronbach's alpha	Correlations between actual scoring and rapid judgements						
		Observer1	Observer2	Observer3	Observer4	Observer5	combined	
<u>nonverbal behaviours</u>								
latency	.69	.32*	.61**	.32*	.39**	.44**	.60**	
hand and finger	.66	.50**	.47**	.45**	.48**	.16	.54**	
speech hesitations	.77	.22	.20	.39**	.19	.45**	.38**	
<u>Criteria-Based Content Analysis criteria</u>								
number of details	.84	.53**	.72**	.70**	.64**	.49**	.78**	
contextual embedding	.76	.60**	.58**	.44**	.34*	.56**	.71**	
reproduction of conversation	.89	.73**	.53**	.64**	.67**	.57**	.71**	
other's mental state	.92	.30*	.44**	.34*	.44**	.46**	.44**	
<u>Reality Monitoring criteria</u>								
visual details	.69	.49**	.61**	.53**	.21	.43**	.69**	
sound details	.81	.49**	.32*	.53**	.30*	.49**	.58**	
spatial details	.71	.20	.19	.29*	.66**	.19	.43**	
temporal details	.79	.33*	.41**	.42**	.63**	.24	.54**	
cognitive operations	.48	.20	-.09	.32*	.21	.16	not calculated	

** p < .01, * p < .05

Table 3.
Verbal and nonverbal cues (measured via actual frequency scoring and via rapid judgements) as a function of deception

	actual (Vrij et al., in press)		rapid judgements				number of observers				significant results		
	truth	lie	F(1,50)		truth		F(1, 50) who obtained						
			m	sd	m	sd	lie	m	sd	m	sd		
<u>nonverbal behaviours</u> ¹¹													
latency		2.02	1.9	1.40	1.6	1.61		1.95	.5	2.12	.6	1.21	0
hand and finger		21.24	19.1	10.55	8.5	6.76*		2.55	.7	2.37	1.1	.50	2
speech hesitations		9.15	4.9	9.77	5.3	.19		2.53	.7	2.49	.6	.04	0
<u>Criteria-Based Content Analysis criteria</u> ¹²													
number of details		51.42	19.6	38.87	16.4	6.28*		3.31	.6	2.75	.6	10.32**	4
contextual embedding		12.08	6.6	6.62	3.7	13.64**		2.80	.6	2.34	.5	8.66**	2
reproduction of conversation		2.27	.9	1.77	.8	4.31*		2.70	.7	1.93	.8	13.42**	4
other's mental state		.04	.2	.21	.4	4.28*		1.09	.3	1.28	.5	2.51	0
<u>Reality Monitoring criteria</u>													
visual details		44.03	20.7	29.46	10.9	10.11**		2.82	.4	2.44	.6	7.17**	2
sound details		6.92	4.5	3.08	2.5	14.66**		2.19	.5	1.77	.7	6.53*	2
spatial details		5.64	5.7	2.62	2.3	6.21*		2.37	.5	2.08	.5	3.82 ^a	2
temporal details		6.64	4.4	3.58	2.3	9.80**		3.01	.5	2.52	.5	13.25**	4

** p < .01, * p < .05, ^a p = 0.56

Table 4.
Accuracy rates for each observer and the five observers combined

	accuracy scores					
	truth		lie		total	
	m	sd	m	sd	m	sd
total	.82**	.25	.65**	.35	.74**	.31
Observer 1	.92**	.27	.77**	.43	.85**	.36
Observer 2	.81**	.40	.73**	.45	.77**	.43
Observer 3	.89**	.33	.65**	.49	.77**	.42
Observer 4	.62 ^t	.50	.50	.51	.56	.50
Observer 5	.85**	.37	.62 ^t	.50	.73**	.45

** $p < .01$, ^t $.05 < p < .10$

Table 5.
Correlations between rapid judgements and veracity judgements

	Observer1		Observer2		Observer3		Observer4		Observer5		combined		r	b
	r	R	r	R	r	R	r	R	r	R	r	R		
<u>nonverbal behaviours</u>														
latency	.43**		.10		.11		.58**	.26**	.34*		.38*			
hand and finger		-.42**		-.15		-.28*		.04		-.10		-.08		
speech hesitations		.31*		.06		.00		.39**		.26		.30*		
<u>Criteria-Based Content Analysis criteria</u>														
number of details		-.52**		-.62**		-.57**	-.39**	-.72**	-.25*	-.55**	-.17*	-.79**	-.52**	
contextual embedding	-.52**		-.54**		-.45**		-.39**		-.46**		-.72**			
reproduction of conversation	-.59**	-.21*	-.64**		-.53**		-.45**		-.54**	-.20*	-.75**	-.35**		
other's mental state		-.17		-.17		.31*		-.08		-.18		-.13		
<u>Reality Monitoring criteria</u>														
visual details		-.47**		-.70**	-.32**	-.46**		-.45**	-.20*	-.31*		-.72**		
sound details		-.44**		-.30*		-.50**		-.11		-.51**		-.65**		
spatial details		-.56**	-.18*	-.44**		-.26		-.34*		-.09		-.53**		
temporal details		-.56**		-.59**	-.18*	-.42**		-.57**		-.48**		-.75**		
cognitive operations		.51**	.24*	-.02		-.10		-.26		.25	.19*	.03		
<u>Other</u>														
length of speech		-.52**		-.61**		-.43**		-.41**		-.42**		-.61**		

** p < .01, * p < .05

1. Diagnostic cues are nonverbal and verbal behaviours which, according to deception research, are (to some extent) associated with deception. See DePaulo, Lindsay, Malone, Muhlenbruck, Charlton, & Cooper (in press), and Vrij (2000) for reviews about cues to deception.
2. By mistake, spontaneous corrections were not included in the rapid judgement task.
3. It was told that latency period, speech hesitations and cognitive operations typically increase during deception and that all the remaining variables typically decrease during deception.
4. There were two reasons for introducing this 'two-step' training programme. First, the first session was needed to obtain anchor scores that could be used in the second session. Second, the current procedure resulted in a 'responsible role' for the research assistant, which was a necessary requirement for obtaining the Nuffield Foundation grant.
5. In all analyses the 52 clips rather than the participants (observers) were the unit of analysis.
6. Additional ANOVAs were conducted on the rapid judgements for each individual judge (last column of Table 3). None of these rapid judgements revealed a significant difference between liars and truth tellers regarding descriptions of other's mental state, whereas the rapid judgements of two observers showed significant differences between liars and truth tellers regarding hand and finger movements.
7. The fact that observers saw so many clips ($N = 52$) may have resulted in a "learning effect". Perhaps, after hearing numerous statements they may have worked out the facts of the staged event which could have facilitate lie detection. That is, observers then only had to compare the individual statements with the known facts and could have judged a statement as deceptive when the information provided in the statement contradicted these known facts. There is no evidence for a learning effect. In order to examine this effect, the 52 clips were divided into three subgroups: clip 1-17, clips 18-34, and clips 35-52. A learning effect would have resulted in the highest accuracy rates in the third group (clips 35-52). This was not the case. An ANOVA with Group as factor and the total accuracy scores as dependent variable revealed a non significant effect, $F(2, 49) = 1.62$, *ns*. (Total accuracy scores per groups were: clip 1-17: $M = .84$, $SD = .3$, clip 18-34: $M = .65$, $SD = .4$, clip 35-52: $M = .72$, $SD = .3$).
8. We also tested for learning effects for the individual observers (see endnote 7). For Observers 1, 2, 3 and 5, ANOVAs with Group as factor (clips 1-17, clips 15-34, clips 36-52) and total accuracy rates as dependent variable (one ANOVA was conducted for each judge) resulted in non significant effects (all F s < 1.00). The effect for Observer 4 was significant, $F(2, 49) = 4.46$, $p < .05$. Mean scores revealed that the highest accuracy was achieved in the first group of clips (clips 1-17: $M = .82$, $SD = .4$; clips 18-34: $M = .35$, $SD = .5$, clips 35-52: $M = .50$, $SD = .5$). This suggests a "fatigue effect" rather than a learning effect. Theoretically, this fatigue effect could have been caused by a truth bias or a lie bias. That is, perhaps after a while Observer 4 had the tendency to judge statement as truthful (truth bias) or as deceptive (lie bias). There is no evidence for this. A truth/lie bias would result in a significant interaction effect in a 3 (Group) X 2 (Veracity of the clip) ANOVA with accuracy as dependent variable. In fact, the interaction was not significant, $F(2, 46) = .75$, *ns*.
9. The results for cognitive operations (both the results per individual observer and the combined results) were also included in the analyses reported in Table 5. Although the combined measure for cognitive operations is unreliable (see Table 2), we cannot disregard this measurement in these analyses as, in principle, the observers could have been guided by cognitive operations while making their veracity judgements.
10. Observer 1: A logistic regression revealed four predictors ($\chi^2(4, N = 52) = 50.51$, $p < .01$). Latency time (Wald = 3.51, $p = .06$, $R = .15$), reproduction of conversation (Wald = 5.05, $p < .05$, $R = -.21$), spatial

details (Wald = 4.27, $p < .05$, $R = -.18$) and cognitive operations (Wald = 6.14, $p < .05$, $R = .25$). On the basis of those four cues 90.38% of the cases could be correctly classified.

Observer 2: A logistic regression revealed two predictors ($\chi^2(2, N = 52) = 34.85$, $p < .01$). Visual details (Wald = 9.50, $p < .01$, $R = -.32$) and temporal details (Wald = 4.32, $p < .05$, $R = -.18$). On the basis of those two cues 92.31% of the cases could be correctly classified.

Observer 3: A logistic regression revealed two predictors ($\chi^2(2, N = 52) = 27.01$, $p < .01$). Number of details (Wald = 12.50, $p < .01$, $R = -.39$) and attributions of other's mental state (Wald = .06, *ns*, $R = .00$). On the basis of those two cues 82.69% of the cases could be correctly classified.

Observer 4: A logistic regression revealed three predictors ($\chi^2(3, N = 52) = 51.62$, $p < .01$). Latency time (Wald = 6.76, $p < .01$, $R = .26$), number of details (Wald = 6.31, $p < .05$, $R = -.25$) and visual details (Wald = 4.68, $p < .05$, $R = -.20$). On the basis of these three cues 92.16% of the cases could be correctly classified.

Observer 5: A logistic regression revealed three predictors ($\chi^2(3, N = 52) = 28.35$, $p < .01$). Number of details (Wald = 3.93, $p < .05$, $R = -.17$), reproduction of speech (Wald = 4.68, $p < .05$, $R = .26$) and cognitive operations (Wald = .19, $p < .05$, $R = .19$). On the basis of these three cues 84.62% of the cases could be correctly classified.

11. Hand and finger movements and speech hesitations were corrected for the length of interview and number of spoken words. Hand and finger movements scores represent the frequency of such movements per one minute of speech; speech hesitation scores represent the number of speech hesitations per 100 words.
12. Unlike nonverbal behaviours, the verbal criteria (CBCA and RM criteria) were not corrected for the number of spoken words and/or length of interview. Such a correction is inappropriate as longer speech is an automatic result of the presence of the verbal criteria. That is, the more details someone mentions, the longer the person will speak, and so on. Correcting for speech length will therefore negate the effects of the verbal criteria. Additionally, correction for speech length will substantially change the nature of these criteria, as it will provide information about the 'density of details' in a statement (that is, the more details mentioned in the fewer words, the higher the score). Verbal criteria, however, do not refer to density of details.