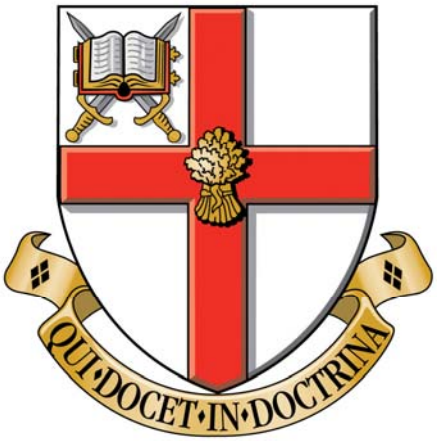

University of Chester

Abstract

To date, the majority of investigations in to accuracy in detecting deception have used low stakes lies as stimulus materials, and findings from these studies suggest that people are generally poor at detecting deception. The research presented here utilised real life, high stakes lies as stimulus materials, to investigate the accuracy of police and non-police observers in detecting deception. It was hypothesised that both police and non-police observers would achieve above chance levels of accuracy in detecting deception, that police officers would be more accurate at detecting deception than non-police observers, that confidence in veracity judgements would be positively related to accuracy, and that consensus judgements would predict veracity. 107 observers (70 police officers and 37 non-police participants) watched 36 videos of people lying or telling the truth in an extremely high stakes, real life situation. Police observers achieved mean accuracy in detecting deception of 72%, non-police observers achieved 68% mean accuracy, and confidence in veracity judgements were positively related to accuracy. Consensus judgements correctly predicted veracity in 92% of cases.

Key words: deception detection; high stakes lies

Meta-analytic findings suggest that, in general, people are poor at detecting deception. Bond and DePaulo (2006), in their meta-analysis, report an overall average accuracy in detecting deception of 54%, which is barely above the 50% accuracy that would be expected by chance. Furthermore, a review of the research has suggested that there is no relationship between accuracy and confidence in making judgements of veracity (DePaulo et al., 1997), such that when people attempt to detect deception, they report feeling as confident in their decisions when they are incorrect, as when they are correct. However, it is possible that such findings may reflect some major limitations in the dominant methodology used in deception research.

One possible explanation for the generally low accuracy rates, and lack of relationship between accuracy and confidence, is that they are an artefact of the stimulus materials used in the experiments; i.e. the large majority of studies of deception detection have used low stakes lies as stimulus materials. Miller and Stiff (1993) have suggested that in typical laboratory studies, the stakes are not high enough to elicit discernible cues to deception, and thus deceit is almost impossible to detect, as detection is dependent on the cues available. These types of low stakes lies may differ from more high stakes lies, for example, lying about committing a crime, in which the consequences of not being believed are serious for the person telling the lie. One implication is that this may lead to underestimates of the accuracy of observers in detecting deception in more realistic forensic situations. Relevant here are the meta-analytic findings of DePaulo et al. (2003), that strong motivation to succeed in the lie, lies which were identity relevant, and lies about transgressions, were all moderating factors that produced more prominent cues to deception. These are factors that are more likely to be present in high stakes situations than in low stakes situations, particularly in a forensic context. Correspondingly, one might expect high stakes situations to produce more prominent and reliable cues to deception, and consequently also improve the accuracy of observers in detecting deception.

Despite the limitations of laboratory experiments, there has been very little research investigating accuracy in detecting deception in real life, high stakes situations. The few studies that have investigated this have produced reasonably consistent results, which differ from meta-analytic findings of low stakes deception detection (Bond & DePaulo, 2006), and suggest that people may be able to detect deception with accuracy substantially above chance levels: in a series of studies using videos of police interviews with suspects as stimulus materials (Mann & Vrij, 2006; Mann, Vrij & Bull, 2004, 2006; Vrij & Mann, 2001a; Vrij, Mann, Robbins & Robinson, 2006), police officers' credibility judgements were consistently accurate at rates well above chance levels, at 68%, 65%, 69%, 64% and 72% respectively. These relatively high accuracy rates are important, as previous studies, using low stakes stimulus materials, have suggested that police officers are no more accurate than laypersons at detecting deception (for example, DePaulo & Pfeifer, 1986; Ekman & O'Sullivan, 1991; Meissner & Kassin, 2002). It would appear then, that police officers' ability to accurately detect deception increases when they view police interviews of suspects. However, due to restrictions placed on the viewing of the videos of the interviews, the authors of the aforementioned body of studies were unable to test whether non-police participants would also achieve higher accuracy than is usual in deception detection studies, leaving a number of important questions unanswered. For example, the higher accuracy rates may have been due simply to the use of high stakes lies, which may have magnified the behaviours used by the observers to differentiate between honest and deceptive behaviour, and so made them more easily discernible by all. Alternatively, it could be that police officers are particularly good at detecting deception in police interviews, due to a domain familiarity effect: research by O'Sullivan and Ekman (2004), found that law professionals were significantly more successful at identifying lies in a crime deception task than in an emotional deception task, whereas therapists showed the opposite pattern, suggesting that different groups of people may be more successful at

identifying types of lies with which they are most familiar. However, it is not known whether police officers would achieve comparable high accuracy in distinguishing truths from lies in other high stakes contexts, or whether the use of high stakes lies as stimulus materials would affect accuracy and truth bias in members of the public. Clearly, further investigation is needed into these issues.

Also, findings regarding the relationship between observer accuracy and confidence in the credibility judgements of police officers in high stakes deception research have been mixed. In two studies no relationship was found between confidence and accuracy (Mann et al., 2004; Vrij & Mann, 2001b); however, in another study participants were more confident in their correct judgements than in their incorrect judgements of veracity (Mann et al., 2006). This finding in particular is interesting, as a review of the research in low stakes situations has suggested that there is no relationship between accuracy and confidence (DePaulo et al., 1997). Again, to the authors' knowledge, there have been no published investigations of the relationship between accuracy and confidence in the general public, using high stakes lies as stimulus materials. It may be that if observers are able to base their credibility judgements on more pronounced cues to deception, they may be more accurate and also more confident in their accurate decisions.

Another issue that has not been addressed so far in any deception research, is the value of consensus judgements (i.e. a majority opinion) as predictors of veracity: if there is a strong consensus amongst observers that a 'sender' is lying, or telling the truth, how likely is it that the consensus judgement is correct? Most of the work in consensus effects psychology generally has concentrated on group influences on accuracy in decision making, in which members are required to make group decisions. The general finding is that group decisions where there is consensus are no more accurate, and sometimes less accurate, than individual decisions (Reagan-Cirincione, 1994). However, the kind of consensus effect referred to in the

present context concerns whether decisions are more accurate when there is an agreement amongst independently derived judgements. According to the general principle of aggregation, we might expect aggregated judgements to be more reliable and valid than individual scores. Perhaps the closest examples we have here are from the area of consensus forecasts (of weather, economy, production etc.), which, due to the principle of aggregation, consistently tend to be better when there is a consensus amongst independently derived forecasts (Blix, Wadefjord, Wienecke & Adahl, 2001). Although previous findings on low stakes situations have found little variance in observer accuracy in deception detection (Bond & DePaulo, 2006; Levine, 2010), this would perhaps not be surprising if there were few discernible cues to detect. However, if there are discernible cues to detect, it could be that consensus judgements derived from independent judges will most enable us to identify them accurately.

In light of the above considerations, the main aim of the research presented here was to investigate the comparative accuracy of police and non-police participants, in detecting deception in a real life, high stakes situation. It was hypothesised that participants, both police and non-police, would produce accuracy rates in detecting deception above chance, because of the high stakes nature of the stimulus materials. It was further hypothesised that, due to the likelihood of increased exposure to deceptive behaviour in the course of their jobs (domain familiarity), police officers would be more accurate at detecting deception than the general public. By the same token, it was also expected that, within the police sample, officers in CID, who routinely interview suspects, would be more accurate in judging credibility than firearms officers, who do not routinely interview suspects. Additionally, despite mixed findings in previous research, it was hypothesised that the use of real life, high stakes deception as stimulus materials, would result in a positive relationship between accurate veracity judgements and confidence in all participants. Finally, it was hypothesised that, due to the principle of aggregation, consensus judgements would accurately predict veracity.

Method

Participants

A total of 107 participants were recruited to take part in the study. The participants were 37 firearms officers, and 33 CID officers, of a large North West police service, and 37 undergraduate students participating for course credit. Students were recruited using the University online participation system, and police officers were recruited using opportunity sampling. All participation was voluntary. There were 33 male and four female firearms officers, ages ranged from 26 to 53 (M = 37.27, SD = 6.98), and years of experience as a police officer ranged from six to 29 years (M = 13.35, SD = 6.38). There were 18 male and 15 female CID officers, whose ages ranged from 26 to 46 years (M = 34.18, SD = 6.40), and years of experience as a police officer ranged from two to 23 years (M = 9.67, SD = 5.73). There were four male and 33 female undergraduate students; their ages ranged from 18 to 25 years (M = 19.19, SD = 1.58). None of the undergraduate students had been a police officer. It can be noted that although there were age and gender discrepancies between the groups, meta-analytic findings indicate that there is no effect of age or gender on accuracy in detecting deception (Aamondt & Custer, 2006)

Materials

To address the issues of ecological validity discussed above, the materials that were chosen as stimulus materials consisted of video footage of people making public appeals for help with missing or murdered relatives. It is not unusual, when a person goes missing or is killed, for a relative to appear before the press and appeal to the public to help find the missing person, or to help find out who killed the person. Sometimes the person making the appeal is honest; he or she is not involved in the death or disappearance of the relative, and is genuinely appealing to the public for help. However, sometimes the person making the appeal is deceptive; he or

she is involved in the death or disappearance of the relative, and the appeal is a deceptive communication, in which the appealer attempts to manipulate the beliefs of others by concealing knowledge of the crime, and falsifying an appeal (a deceptive appealer does not actually want the public to help find the relative, or find out who killed the relative). Thirty six appeals were used in the present study, 18 of which were honest, and 18 of which were deceptive.

There were several advantages in using appeals as stimulus materials, for example; it allowed non-police participants to view the same stimulus materials as police officers, and therefore to compare non-police and police accuracy in detecting deception in a high stakes situation; it allowed the investigation of police accuracy in detecting deception in a context other than a police interview. It also allowed the adoption of a between subjects stimulus materials design; i.e. participants observed honest and deceptive individuals. In the majority of previous research investigating deception in high stakes situations, a within subjects stimulus materials design has been used (Mann & Vrij, 2006; Mann et al., 2004, 2006; Vrij & Mann, 2001a; Vrij et al., 2006). In these studies, honest and deceptive sections of communications by individuals guilty of committing various crimes are compared, and this within subjects method is limited as it negates the possibility of examining the individuals holistically; i.e. if a person is lying in parts of his/her testimony, it does not follow that behaviours relevant to the classification of individuals as honest or deceptive, will be limited only to those parts of the testimony which are untruthful. Furthermore, although some have stressed the importance of using active interactive situations to study cues to deception (DePaulo & Bond, 2012; Vrij & Granhag, 2012), the method in which observers make judgements of deception by viewing short video clips in a passive paradigm reflects exactly how appeals are seen in real life: appeals are watched passively by observers, including police officers who are investigating the case.

Freely available video footage of people making public appeals for help with missing or murdered relatives was collected from various online news and media sites from the United States, the United Kingdom, Canada, Australia and New Zealand. Appeals were not considered for inclusion if they were recent and high profile in the UK, or if they were made many months or years after the event; all appeals included were made within one month of the relative going missing or being murdered. In all cases classified as deceptive, the appealer was convicted in a criminal court of involvement in the death (or in one case, kidnapping), of their relative. In all cases classified as honest, either another person was convicted of the death of the relative, or the relative was found with no evidence of a crime. Furthermore, appeals were considered for inclusion only if they satisfied stringent criteria for establishing ground truth; i.e. the evidence was strong enough to support the conclusion that the appeals were actually truthful or deceptive. According to other published research in this area, a number of criteria can be used to determine whether ground truth has been established; those used by ten Brinke and Porter (2012), and Vrij and Mann (2001b), were used here. Appealers were classified as honest or deceptive only if there was overwhelming evidence indicating the extent of their involvement in the death or disappearance of their relative using these criteria. For a full description of the sample used in the present study, including the criteria used for veracity classification, see Wright Whelan, Wagstaff and Wheatcroft (2013).

Table 1 around here

Procedure

Participants attended the study in a psychology laboratory at the University, in groups ranging in size from three participants to 14 participants. There were seven groups of firearms officers, three groups of CID officers, and four groups of students. This variation in group size was a

result of the number of police officers able to participate at any particular session, and did not affect the running of the study. Each participant was given an instruction and response sheet. Participants were informed that they would be viewing short video clips of people appealing for help with missing or murdered relatives, and would be asked to decide whether they thought each appealer was lying (i.e. was actually involved in the death or disappearance of their relative), or telling the truth (i.e. was not involved in the death or disappearance of their relative). For each appeal, participants were also asked to rate how confident they were in their credibility judgement, on a Likert scale of 1 (not at all confident) to 5 (very confident). Before each appeal was shown, a short summary was provided of what was publicly known of the case at the time of the appeal, for example, 'This clip features a man talking about his missing 34 year old daughter, Nancy'. Each summary included: the relationship between the appealer and the relative (e.g. parent, spouse, sibling, child); the age of the missing or murdered relative if the appealer was a parent (to allow for possible differences in observers' expectations of parental behaviour depending on whether, for example, their child was an infant or an adult); an explanation of any names, events or details mentioned in the appeal; and whether the relative was missing (i.e. no body had been found) or publicly known to be dead (i.e. the body had been found). Participants were asked to check a box on the response sheet if they were familiar with the appealer featured in the clip, or the outcome of the case, and to not complete the section of the response sheet (including the veracity judgement) for that appeal. In the group of firearms officers, the number of appeals with which any individual was familiar with ranged from zero appeals (four participants) to six appeals (one participant), $M = 2.43$, $SD = 1.39$. In the group of CID officers, the number of appeals with which any individual was familiar with ranged from zero appeals (one participant) to six appeals (two participants), $M = 3.45$, $SD = 1.62$. In the group of students, the number of appeals with which any individual was familiar with ranged from zero appeals (15 participants) to five appeals (two participants), $M = 1.05$, $SD =$

1.37. For each group, the appeals were shown in a different, randomised order. Participants were shown the first of the 36 appeals, and then completed the relevant section of the response sheet. This process was repeated for each appeal.

Results

Accuracy in Detecting Deception

To investigate the hypothesis that participants would produce accuracy rates in detecting deception above chance, overall accuracy rates were calculated for each participant. Any cases which were known to a participant were not included in the accuracy scores. Overall mean accuracy was 70.75% (*SD* = 8.68), median accuracy was 71.88 %. There were large variations in individual accuracy, which ranged from 91% to 45%. It can be noted that 50% of police officers achieved accuracy rates of 74% or higher, and correspondingly that 27% of public participants achieved accuracy rates of 74% or higher. There was no relationship between accuracy and years of experience as a police officer (r = -.09, p = .461).

As hypothesised, all groups of participants produced percentage accuracy rates above chance. Firearms officers achieved a median accuracy of 72.79%, which a sign test showed to be significantly above chance, Z = 5.92, p < .001 (all values were above 50). CID officers achieved a median accuracy of 74.17%, which a sign test showed to be significantly above chance, Z = 5.57, p < .001 (all values were above 50). The public achieved a median accuracy of 67.65%, which a sign test showed to be significantly above chance, Z = 5.50, p < .001 (35 values were above 50).

To investigate whether there were differences in accuracy between the three groups of participants, and also between veracity conditions (i.e. deceptive appeals and honest appeals), a 3 x 2 mixed ANOVA (group: firearms, CID and public x veracity: deceptive and honest),

with repeated measures on the second factor was conducted on the percentage accuracy scores; the means and *SD*s are shown in Table 2.

A significant main effect was found for group, F (2,104) = 3.37, p = .038, η^2_p = .061. Post hoc tests revealed that, as hypothesised, firearms officers (p = .031) and CID officers (p = .024) were significantly more accurate than the public. However, the difference in accuracy between firearms officers and CID officers was not significant (p = .874), offering no support for the hypothesis that CID officers would be more accurate at detecting deception than firearms officers.

There was no significant main effect for veracity, however, there was a significant interaction between group and veracity, F (2,104) = 9.97, p < .001, η^2_p = .161. Post hoc tests showed CID officers were significantly more accurate (p = .007) when judging deceptive appeals than when judging honest appeals, and the public were significantly more accurate (p = .002) when judging honest appeals than when judging deceptive appeals. Further post hoc F tests revealed that firearms officers (p < .001), and CID officers (p < .001), were more accurate than the public when judging deceptive appeals. When judging honest appeals, there were no significant differences in accuracy between any of the groups. See Table 2.

Table 2 around here

Accuracy and Confidence in Credibility Judgements

To investigate whether accuracy in credibility judgements was related to confidence in credibility judgements, mean confidence ratings were calculated for each participant for correct judgements of deceptive appeals, correct judgements of honest appeals, incorrect judgements of deceptive appeals, and incorrect judgements of honest appeals. A 3 x 2 x 2 mixed ANOVA with repeated measures on the last two factors was conducted on the mean confidence ratings,

with the factors group (firearms, CID and public), veracity (deceptive appeal and honest appeal), and judgement accuracy (correct judgement and incorrect judgement). The means and *SD*s are presented in Table 3.

As hypothesised, a significant main effect was found for judgement accuracy, F (1,102) = 126.11, $p < .001$, $\eta^2_p = .553$, so that, overall, participants were more confident in their correct judgements, than in their incorrect judgements. A significant main effect was also found for veracity, F (1,102) = 6.00, $p = .016$, $\eta^2_p = .056$, so that overall, participants were more confident in their judgements of deceptive appeals, than in their judgements of honest appeals.

A significant interaction was found between veracity and judgement accuracy, F (1,102) = 13.28, $p < .001$, $\eta^2_p = .115$. Post hoc tests showed that for both honest and deceptive appeals, participants were more confident when making a correct judgement, than when making an incorrect judgement ($p < .001$). However, the interaction was caused by a difference in confidence between correct judgements for honest and dishonest appeals. Confidence was significantly higher for correct judgements in deceptive appeals, than for correct judgements in honest appeals ($p < .001$), whereas this was not the case for incorrect judgements.

Table 3 around here

Relationship of Consensus Judgements to Veracity

To reiterate, the idea behind the consensus hypothesis is that when there is high consensus as to the veracity of a particular appeal, it is more likely to be accurate than inaccurate. So, for example, if 75% of people agree that a particular appeal is dishonest, it is likely to be dishonest. To investigate the utility of consensus judgements in determining the veracity of the appeals, therefore, for each appeal, the percentage of correct judgements of veracity ('hits'), was compared with the percentage of incorrect judgements of veracity ('misses'), using Wilcoxon

signed ranks tests. Hence this analysis looks at agreement on appeals, not whether individual observers are accurate.

As expected, the percentage of correct judgements by participants (median = 72.77), was significantly greater than the percentage of incorrect judgements (median = 27.23), Z = - 4.76, p = .001. Indeed, in 33 out of the 36 appeals, the percentage of 'hits' was greater than the percentage of 'misses'. A consensus of 75% or over was always correct, and 17 appeals had a consensus over 75%. In other words, the more likely, collectively, people were to agree on the direction of a judgement, the more correct they were.

Looking at police and public consensus separately, the percentage of 'hits' by police participants (median = 81.00) was significantly greater than the percentage of 'misses' by police participants (median = 19.00), Z = - 4.80, p < .001, and the percentage of 'hits' by public participants (median = 68.00) was significantly greater than the percentage of 'misses' by public participants (median = 32.00), Z = - 4.28, p < .001.

Table 4 around here

To investigate whether accurate consensus judgements were related to the length of the appeal, (i.e. it may be that a high level of accurate consensus is related to longer exposure to the 'sender'), a Pearson's correlation was conducted on the length of the appeals (seconds), and the number of 'hits' on the appeal. The relationship between length of appeal and percentage of correct judgements made about the appeal was not significant, r = .11, p = .513.

Discussion

The present findings support the general proposition that observers (both police and public) would be able to achieve accuracy rates in detecting deception above chance levels when

viewing ecologically valid stimulus materials, which suggests that the common proposal that people are poor at detecting deception (for example, Bond & DePaulo, 2006), may be an artefact of the methodology of utilising low stakes lies. Moreover, as the detection of deception necessarily implies differences between honest and deceptive behaviour, these findings support theoretical proposals that increasing the stakes may exacerbate factors underlying the production of cues to deception, resulting in more, and/or more prominent, cues to deception. There are several further, more specific implications: police officers in the present study were able to achieve high accuracy levels in a high stakes context other than that of police suspect interviews (although the context of appeals could still be regarded as forensic), and police officers were more accurate at detecting deception than the public. These are important findings because, as far as the authors are aware, there have been no previous investigations comparing police accuracy with public accuracy, when judging real life, high stakes lies. It may be that findings from previous research, which have reported that police participants are no more accurate than non-police participants (for example, DePaulo & Pfeifer, 1986; Ekman & O'Sullivan, 1991), may have occurred at least partly because of the low stakes contexts of the stimulus materials.

However, the proposal that police participants may achieve higher accuracy than participants from the public because they are exposed to more high stakes lies, was apparently not supported in that the number of years of experience as a police officer was not related to accuracy. Some light may be shed on this from the finding that it was the police officers' ability to correctly identify deceptive appeals that accounted for their higher overall accuracy rates, as there was no difference in accuracy between police and non-police participants in correctly identifying honest appeals. One possible explanation for this is that, perhaps because of their training, and even limited experience in the field, in terms of Signal Detection Terms, Police officers are encouraged to adopt a more lax criterion (beta shift) for the classification of

deception, and a more cautious criterion for the classification of honesty, than the public. This would result in them paying more initial attention to cues for dishonesty than honesty. This, coupled with any benefits of more experience with people who (because of other evidence) are known to be lying, may give the police an advantage over the public that plateaus with a certain amount of experience (i.e. it will not necessarily improve with years of experience). It can be noted that in the sample used here, the minimum experience was two years. If this is the case, this suggests that only limited exposure to a range of individuals who are lying may be sufficient to boost the ability to detect deception.

Considering that meta-analytic reports (mainly involving low stakes studies) indicate that there is no relationship between confidence and accuracy when people make veracity decisions (DePaulo et al., 1997), the finding that there was a positive relationship between accurate veracity judgements and confidence, is potentially an important one. Previous findings using high stakes lies as stimulus materials have used only police participants, and have produced conflicting results (Mann et al., 2004, 2006; Vrij & Mann, 2001a). However, these studies used limited ranges of stimulus materials and within subjects designs, thus restricting both the potential for variance in the data, and the robustness of the results. This fits with other more general research on witness confidence and accuracy, which suggests that, with a wide range of stimuli that include easy or 'obvious' items which increase the variance, confidence-accuracy relationships tend to be higher so that the more confident people are that they are correct, the more likely they are to actually be correct (Kebbell, Wagstaff & Covey 1996; Wheatcroft, Wagstaff, & Kebbell, 2004). Furthermore, participants were more confident when correctly identifying deceptive appealers than honest appealers, implying that deceptive appealers may be more 'obvious' to observers, than honest appealers. This could be due to a number of factors, for example, cues to deception may be more numerous and, perhaps, more prominent than cues to honesty, thereby making them more salient or 'easier' to recognise.

Also, in this context, one particular factor that might contribute to their prominence is norm violation; deceptive behaviours may be more prominent because they violate the norms of 'usual' acceptable behaviour in social contexts.

The investigation of consensus judgements in the present study was particularly important, as this is an aspect of deception detection that has not been previously explored, and the findings were important for two reasons. First, to the authors' knowledge, observer consensus as a predictor of veracity has not previously been investigated, and the results from the present study suggest that it is an area that warrants further research. Second, it is unusual in deception research for any variable to predict veracity with accuracy as high as 92%. Evidently, as the utility of consensus judgements has not been previously investigated, the findings can only be regarded as preliminary, and replication would be needed before any firm conclusions can be drawn. For example, future research investigating the value of consensus judgements in a different high stakes context, or the minimum number of judges required for this effect, would be useful. Nonetheless, the present findings suggest that, in the context of appeals at least, there are sufficient deceptive and honest behaviours produced by the 'senders', for the majority of observers to make correct veracity judgements about them. It appears that Levine's idea that above chance detection of deception is due to a few 'leaky liars' (2010) may not apply in this high stakes context, as nearly all appealers (both deceptive and honest) were sufficiently 'leaky' to be correctly classified by the consensus of observers.

Moreover, despite large differences in the length of appeals (some were less than 20 seconds, some were over two minutes long), there was no relationship between the percentage of correct credibility judgements, and the length of the appeal. This implies that, as providing observers with longer examples of behaviour did not affect their accuracy, observers may have been making rapid, accurate, global judgements, rather than relying on careful analysis of lengthy sections of behaviour. This suggests that observers may have been looking at a 'type'

of person, rather than specific behavioural cues, and this has important theoretical and methodological implications. In investigations of high stakes deception, particularly in a forensic context, honest and deceptive individuals may differ from each other not only in terms of whether or not they are lying, but also in whether or not they have committed a serious crime; if observers are able to make relatively accurate judgements about credibility using holistic judgements of the type of person who would, in this case, murder a relative, then using a between subjects design is a serious limitation, as it does not allow this difference to emerge. Moreover, theories of the production of behaviours related to deception have not generally taken in to account the possible role of individual differences between people who choose to commit serious crimes and lie about them, and those who do not (although see Wright Whelan et al., 2013).

In sum, the findings from the present study have several important implications. One of the basic propositions of the present research was that using ecologically valid stimulus materials would produce results different to investigations of accuracy in detecting deception using low stakes stimulus materials; the findings presented here support this proposition. Research reporting observer accuracy in detecting high stakes deception is scant, and although the present findings contribute to this small body of research, there is wide scope for further investigation of this area. For example, replication is needed of the findings regarding the relatively high accuracy of non-police observers, and also of the relationship between accuracy and confidence in this group. Also, replication is needed of the finding that police observers are better able to detect deception than non-police observers, and of the factors that may underlie this difference. It is hoped that these basic findings will reinforce the importance of ecological validity in deception research.

References

Aamondt, M.G. & Custer, H. (2006). Who can best catch a liar? A meta-analysis of individual differences in detecting deception. *The Forensic Examiner, 1,* 6-12

Blix, M., Wadefjord, J., Wienecke, U., & Adahl, M. (2001). How good is the forecasting performance of major institutions? *Economic Review, 3,* 451–476.

Bond, C. F. & DePaulo, B. M. (2006). Accuracy of deception judgements. *Personality and Social Psychology Review, 10* (3), 214-234. doi: 10.1207/s15327957pspr1003_2

Bond, C. F. & Uysal, A. (2007). On lie detection "wizards". *Law and Human Behavior, 31* (1), 109-115. doi: 10.1007/s10979-006-9016-1

DePaulo, B. M. & Bond Jr, C. F. (2012). Beyond accuracy: Bigger, broader ways to think about deceit. *Journal of Applied Research in Memory and Cognition, 1* (2), 120-121. doi:10.1016/j.jarmac.2012.04.010

DePaulo, B. M., Charlton, K., Cooper, H., Lindsay, J. J. & Muhlenbruck, L. (1997). The Accuracy-Confidence Correlation in the Detection of Deception. *Personality and Social Psychology Review, 1* (4), 346-357. doi: 10.1207/s15327957pspr0104_5

DePaulo, B. M., & Pfeifer, R. L. (1986). On-the-job experience and skill at detecting deception. *Journal of Applied Social Psychology, 16,* 249–267.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K. and Cooper, H. (2003). Cues to Deception. *Psychological Bulletin, 129* (1), 74-118

Ekman, P., & O'Sullivan, M. (1991). Who can catch a liar? *American Psychologist, 46,* 913–920.

Kebbell, M. R., Wagstaff, G. F. & Covey, J. (1996) Eyewitness confidence and accuracy: Is it dependent on item difficulty? *British Journal of Psychology*, 87, 653-662.

Levine, T. R. (2010). A few transparent liars. *Communication Yearbook*, 34.

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

Law, 12 (3), 307-319

Mann, S., Vrij, A. & Bull, R. (2004). Detecting true lies: Police officers' ability to detect suspects' lies. *Journal of Applied Psychology, 89,* (1), 137-149. doi:10.1037/0021-9010.89.1.137

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

Meissner, C. A., & Kassin, S. M. (2002). "He's guilty!": Investigator bias in judgments of truth and deception. *Law and Human Behavior, 26,* 469–480.

Miller, G. R. & Stiff, J. B. (1993). *Deceptive Communication.* Newbury Park, CA: Sage

O'Sullivan, M. (2007). Unicorns or Tiger Woods: Are lie detection experts myths or rarities? A response to On Lie Detection 'Wizards' by Bond and Uysal. *Law and Human Behavior, 31* (1), 117-123. doi: 10.1007/s10979-006-9058-4

O'Sullivan, M. & Ekman, P. (2004). The wizards of deception detection. In Granhag, P. A. & Stromwall, L. A. (Eds), *The Detection of Deception in Forensic Contexts* (pp.269-286). Cambridge: Cambridge University Press

Reagan-Cirincione, P**.** (1994). Improving the accuracy of group judgment: A process intervention combining group facilitation, social judgment analysis, and information technology**.** *Organizational Behavior and Human Decision Processes, 58,* 246–270**.**

Shaw, J., Porter, S., & ten Brinke, L. (2013). Catching liars: Training mental health and legal professionals to detect high-stakes lies. *Journal of Forensic Psychiatry & Psychology,* published online. doi:10.1080/14789949.2012.752025

ten Brinke, L., & Porter, S. (2012). Cry me a river: Identifying the behavioral consequences of extremely high-stakes interpersonal deception. *Law and Human Behavior* (in press)

Vrij, A. & Granhag, P. A. (2012). Eliciting cues to deception and truth: What matters are the

questions asked. *Journal of Applied Research in Memory and Cognition, 1* (2), 110-117. doi:10.1016/j.jarmac.2012.02.004

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

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

Vrij, A., Mann, S., Robbins, E. & Robinson, M. (2006). Police officers ability to detect deception in high stakes situations and in repeated lie detection tests. *Applied Cognitive Psychology, 20,* 741-755. doi:10.1002/acp.1200

Wheatcroft, J. M., Wagstaff, G. F. & Kebbell, M. R. (2004). The influence of courtroom questioning style on actual and perceived eyewitness confidence and accuracy. *Legal & Criminological Psychology*, 9, 1-20.

Wright Whelan, C., Wagstaff, G. F. & Wheatcroft, J. M. (2013). High stakes lies: Verbal and nonverbal cues to deception in public appeals for help with missing or murdered relatives. *Psychiatry, Psychology and Law.* Published online 23rd September 2013. DOI: 10.1080/13218719.2013.839931