APPLIED COGNITIVE PSYCHOLOGY
Appl. Cognit. Psychol. **24**: 849–867 (2010)
Published online 22 June 2009 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/acp.1592

Turning a Blind Eye to Double Blind Line-Ups

DANIEL B. WRIGHT*, MARIANNA E. CARLUCCI, JACQUELINE R. EVANS† and NADJA SCHREIBER COMPO

Florida International University, Miami, USA

SUMMARY

Although psychologists have urged police officers to use double blind line-up procedures during their investigations, police officers state that these would be difficult to administer and most have been reluctant to implement this change. Four studies examine whether lay people's judgements about the guilt of a suspect vary according to whether a brief written summary of a case described the identification procedure as double blind or non-double blind. The effects were all small (and almost all non-significant). Most people do not treat double blind line-ups differently from non-double blind line-ups when assessing the guilt of a defendant. Either police investigators should stop using this biased method or police investigators and others in the judicial system (e.g. jurors, judges) should be informed of this bias when evaluating results from any line-up. Copyright © 2009 John Wiley & Sons, Ltd.

A century ago, Clever Hans amazed audiences with his mathematical feats; the horse was able to answer a range of arithmetic tasks (see papers in Sebeok & Rosenthal, 1981). In front of crowds of onlookers, Clever Hans would be asked a question and then beat his hoof the number of times corresponding to the correct answer. When Oskar Pfungst showed that Hans' feats depended on him picking up minute cues from the audience rather than equine intellect, the amazement shifted towards the ease with which information can be transferred, intentionally and unintentionally, during social interaction. This has become a central research area within social psychology (e.g. Rosenthal, 2003; Rosenthal & Jacobson, 1968; Rosental & Rubin, 1978). Further, humans are unable to control some of the cues about their knowledge, which is why poker tournaments are not just games of chance. We are not always able to verbalize correctly how we use certain types of information (Nisbett & Wilson, 1977) and cannot always interpret cues accurately, as in lie detection (Sporer & Schwandt, 2007; Vrij, 2008), but we are responsive to many of these cues.

Within the medical sciences, for example, when researchers conduct drug trials comparing a placebo and active drugs participants are unaware of their condition. This is called a single blind design. Medical researchers further ensure that whoever is in contact with the participants is also unaware in which condition the participants are. This is called a double blind design, and prevents researchers from treating participants in the two conditions differently. Most methodology textbooks state double blind designs are

*Correspondence to: Daniel B. Wright, Psychology Department, Florida International University, Miami, FL 33199, USA. E-mail: dwright@fiu.edu
† Present address: Department of Psychology, University of Texas at El Paso, Texas 79968, USA.

important to minimize unintentional cues, but given that some research is conducted by people who are motivated to observe certain types of results (e.g. for drug approval, for tenure), intentional cues could also be present. Sometimes it is impractical to use double blind procedures and sometimes even if double blind procedures are used the administrator and participant may figure out the condition (e.g. Douglass, Smith, & Fraser-Thill, 2005; Margraf et al., 1991). In general however, researchers are urged to use this procedure.

Within forensic circles intentional and unintentional cues can have pervasive impacts at many stages of judicial procedure, from the police stopping and searching someone 'acting suspiciously' to the parole board assessing the sincerity of a convict's pleas. In this paper, we focus on one step within this process, the line-up. In a single blind line-up administration, an eyewitness is not aware which person is the suspect. Show-ups, where a single individual is presented to the eyewitness, are not single blind as the eyewitness is aware that this person is under suspicion. The courts therefore frown on using show-ups and cases have been over-turned because a show-up has been needlessly used (State of Wisconsin v. Dubose, 2005). However, for standard multiple person line-ups all agree that the eyewitness should not know who the suspect is. Thus, these are single blind line-ups. Rather than use the phrase 'single blind', in this paper we use the phrase 'non-double blind'. This is because some discussions just use the word 'blind' to refer to 'double blind' line-ups.

Wells (1988) recommended that line-ups should be double blind. More recently, the American Psychology-Law Society white paper on eyewitness identification (Wells, Small, Penrod, Malpass, Fulero, & Brimacombe, 1998) and the British Psychological Society's response to the UK government proposals have made similar recommendations (British Psychological Society, 2007). Until recently very few line-ups were conducted double blind but in the past few years some jurisdictions have begun doing so (e.g. in Wisconsin, see http://www.doj.state.wi.us/dles/tns/ILRptResponse.pdf, accessed 17 May 2009).

The reason for the slow progress is likely because police officers say that it will be difficult to conduct double blind line-ups (Mecklenburg, 2006, p. 48). The officers in Mecklenburg's survey do state that there are some advantages to double blind line-ups, like protection from lawsuits where others might claim they influenced an eyewitness. However, they also believe there are advantages to having non-double blind administration because the officers involved in a case might be able 'to recognize relevant information during the identification procedure' (p. 57). The additional financial costs of having a double blind administrator were also mentioned. The reluctance of police to use double blind procedures means that non-double blind procedures are likely to continue to be used.

Existing data do not allow us to estimate reliably the impact of using non-double blind procedures for actual line-ups. Some laboratory data show that inadvertent cues can affect eyewitnesses' behaviour (e.g. Greathouse & Kovera, 2009; Haw & Fisher, 2004; Phillips, McAuliff, Kovera, & Cutler, 1999), but for the purpose of experimental control these studies use only a limited number of identification situations and thus cannot provide an estimate of the effect size.

One recent field study suggests the effect may be large, but it has other methodological limitations. In Mecklenburg (2006, Table 3a) non-double blind and double blind conditions were compared. The suspect was identified in 45% of double blind line-ups, but this rose to 60% in non-double blind line-ups. However, this study confounded the 'blind' variable with the use of simultaneous and sequential line-ups (Lindsay & Wells, 1985). There is much debate about the differential effectiveness of these two procedures (e.g. Carlson,

DOI: 10.1002/acp

Gronlund, & Clark, 2008; Gronlund, 2005; McQuiston-Surrett, Malpass, & Tredoux, 2006). The only undisputed conclusion is that the relative effectiveness of these two procedures is dependent on other contextual variables. For argument's sake, suppose that overall the simultaneous *versus* sequential manipulation made little difference for Mecklenburg's data. This suggests that for approximately 25% of the identifications in non-double blind line-ups the administrator either consciously or unconsciously directed the eyewitness towards the suspect. Another way an administrator could influence an eyewitness is to guide them away from filler identifications. In Mecklenburg's study, there were 9.2% filler identifications in double blind line-ups but only 2.8% in non-double blind line-ups. This suggests that whatever messages administrators convey to eyewitnesses may result in choosing fillers less than one-third of the time that they are chosen in double blind identifications. There is much discussion about this, and several other limitations of the Mecklenburg study (Spinney, 2008), including that outcomes were not properly recorded and random assignment to conditions was not done (Steblay, 2009). Even if these other flaws had not occurred, confounding double blind *versus* non-double blind with simultaneous *versus* sequential makes it difficult to reach any firm conclusions from these data. If similar findings occur with properly designed and conducted field studies, it would suggest that the effect of double blind line-up procedures may be large.

The purpose of the current studies is to examine whether people are aware of the importance of double blind line-up administration. If they are, then they could weigh the value of the identification less in their decision making if the administration was non-double blind. Of course, it would be better if more reliable identification methods were used, but at least biased methods would not unduly influence judicial procedures. If people are not aware of the importance of double blind line-up administration, the biased procedure needs to stop being used (which would be ideal, but police reluctance may prevent this) or instructions (by judge or expert) should warn jurors of the problems inherent in non-double blind administration. A similar warning should be given to police investigators who may use identification outcomes to help direct their investigations. Given that people who encountered Clever Hans were initially more willing to accept equine mathematical ability than non-verbal cueing, it is likely that people may not differentiate between double blind and non-double blind line-ups. This guides the designs of our studies, and we therefore designed sensitive tests to detect this possibility.

STUDY 1—BETWEEN SUBJECTS STUDY

In order to measure if, and to what extent, individuals are influenced by a line-up procedure being described as double blind or non-double blind, we provided participants a very brief summary of a crime adapted from Loftus and Ketcham (1991). We purposely used a brief summary so that the description of the line-up procedure would be noticeable and more likely to have an effect than if it were embedded in a lengthy trial transcript. The summary was created so that participants should vary in how guilty they believe the defendant is depending on whether or not the suspect was identified. The summary is of a rape case where the eyewitness is also the victim (see Appendix). This summary is used in Studies 1–3. There are limitations using just a single case (Wells & Windschitl, 1999). There might be something particular about our scenario or about rape cases. We choose this summary because it has been shown to yield large memory effects in previous research (Wright & Hall, 2007). The line-ups were all six person simultaneous line-ups; Wright (2007)

DOI: 10.1002/acp

compared sequential and simultaneous line-up administration for this case and found no systematic differences. In the first study, we use a between subjects design to be most similar to the situation presented to police, jurors and others when trying to decide the guilt of an individual (i.e. deciding on one case at a time). We asked participants both to say whether they thought the defendant was guilty or not, and also to provide a belief in guilt measure on a 0–100 scale. Having two measures provides two different yet related ways for measuring the effect.

If participants treat identifications according to psychological research, suspect identifications in double blind line-ups should lead to more guilty verdicts and higher beliefs in guilt than identifications from non-double blind line-ups, as the eyewitness chose the suspect without any additional cues. Further, filler identifications and non-identifications from non-double blind line-ups should be more indicative of innocence than those made from double blind line-ups: The eyewitness failed to identify the suspect even with whatever extra (conscious or unconscious) guidance may be present in non-double blind line-ups directing the eyewitness away from choosing fillers. If this 3 (choosing suspect *vs*. choosing filler *vs*. non-identification) $\times$ 2 (double blind *vs*. non-double blind) design is analysed in a single test for belief in guilt (or verdict) the combination of these effects predicts an interaction. However, given that the effects of double blind *versus* non-double blind have different predictions for suspect identifications, filler identifications and no identifications, we also examine each of these separately.

Methods

A power analysis was conducted to determine an appropriate sample size for detecting the effect for comparing the double blind *versus* non-double blind conditions for each level of who the eyewitness chose. Using Cohen's (1988) terminology, we designed the study so that for each eyewitness outcome (choosing a suspect, choosing a filler and making no identification) there was at least 80% power for detecting an effect halfway between his small and medium effect sizes for the verdict ($w = .2$). For each of these three comparisons, G*Power 3.0.5 (Faul, Erdfelder, Lang, & Buchner, 2007) suggested 197 people, so the aim was to recruit approximately 600 participants for the entire study. In total, 574 usable cases were gathered.

Participants for all four studies were recruited *via* the Florida International University Psychology Participant Pool. Participants were given course credit. They were directed to a URL to take part in the study. Only participants 18 and older were recruited. In this first study, the mean age was 21 years old with a standard deviation of 4 years. Most (69%) identified themselves as Hispanic, with the next most common ethnic category being White (14%). Sixty-seven per cent said they were female.

Participants were randomly allocated to one of six conditions of a 2 (double blind *vs*. non-double blind line-up administration) $\times$ 3 (eyewitness' choice: Suspect *vs*. foil *vs*. no identification) between subjects design. The factors were manipulated in a crime summary provided to the participants (see Appendix). The critical aspect was how the line-up administration was described.

Double blind

'A police officer took pictures of five other men, all with beards, between 25 and 30, and about the same build as Hoyle. On the 13th, a police officer not involved in the case showed these pictures with the one of Tom Hoyle to Nancy Von Roper. The officer, who did not

DOI: 10.1002/acp

know which photograph was of Hoyle, the suspect, showed all six photographs at the same time'.

Non-double blind
'A police officer took pictures of five other men, all with beards, between 25 and 30, and about the same build as Hoyle. On the 13th one of the police officers involved in the case showed these pictures with the one of Tom Hoyle to Nancy Von Roper. The officer, who knew which photograph was of Hoyle, the suspect, showed all six photographs at the same time'.

Participants were then asked to render a verdict, either guilty or not guilty. Next, they were asked how strongly they believed that the suspect was the culprit on a 0–100 scale (see Appendix for the exact wording).

Web-based studies have become quite popular in psychology, and initial research suggests they produce comparable results to traditional pen-and-paper methods (Gosling, Vazire, Srivastava, & John, 2004; McGraw, Tew, & Williams, 2000). Compared with computer tasks occurring within a laboratory setting, online studies have two difficulties that need addressing. The first concerns people taking part multiple times. In order to receive course credit in our studies, participants had to use their student number, meaning anyone repeating the study could be identified and removed from the dataset (none did). The second difficulty is not knowing the condition under which participants are taking part. Although participants are randomly allocated into conditions in this study, and therefore participants taking part under unreliable conditions should become part of the random error, we did remove some participants. Specifically, in this and subsequent studies, we removed anyone who took less than 1 minute to read through the crime summary. This produced a final sample size of 574.

All four studies received approval from Florida International University's IRB and participants received appropriate pre-study explanations and post-study debriefing.

Results

All four studies used a similar analytic method. The interest is in how the experimental manipulations affected both people's verdicts and their beliefs in guilt. These are different ways of measuring guilt and they are closely related. Therefore, we expected them to show similar effects and analysed them in a similar way. Because verdict is a binary variable, logistic regression was used. For main effects the effect size used was Nagelkerke's R^2 for the variable on its own (i.e. without the other variable or the interaction included). For the interaction partial Nagelkerke's R^2, which compares the fit of the model including the interaction and the main effects with the model with only main effects, is reported (see Appendix of Wright, Hanoteau, Parkinson, & Tatham (in press), for derivation and a function to calculate this using the freeware *R*). The belief in guilt variable is continuous, but is skewed, so this variable is transformed in order to meet the assumptions of the statistical procedures and then analysed. The effect size reported for the transformed belief variable is η^2 for the main effects and partial η^2 for the interaction. Specific tests comparing double blind with non-double blind conditions for each of the identification conditions were planned.

Table 1 lists the proportion of guilty verdicts and the mean belief in guilt for each cell of the 2×3 design. We begin with the analysis of the participants' verdicts for the full design.

DOI: 10.1002/acp

Table 1. The percentage of guilty verdicts and mean belief in guilt ratings

	Line-up administration		
	Double blind	Non-double blind	Effect sizes
Suspect picked			
Guilty verdicts %	74.7%	73.0%	$NR^2 = .001$, $p = .92$
Belief in guilt	75.82	81.10	$\eta^2 = .013$, $p = .12$
Filler picked			
Guilty verdicts %	30.1%	39.0%	$NR^2 = .011$, $p = .28$
Belief in guilt	54.08	52.64	$\eta^2 = .001$, $p = .72$
No identification			
Guilty verdicts %	33.0%	34.8%	$NR^2 = .001$, $p = .91$
Belief in guilt	54.94	53.55	$\eta^2 = .001$, $p = .64$

Note: NR^2 = Nagelkerke's R^2.
Effect sizes for comparisons shown in the final column. The *p*-values for verdicts are based on a Pearson's χ^2 test using Yates' correction. The *p*-values for belief are based on a *t*-test on the transformed belief variable

The eyewitness' choice had a significant effect, $\chi^2(2) = 80.70$, $p < .001$, Nagelkerke's $R^2 = .175$. The effect of whether or not the line-up administration was double blind was non-significant, $\chi^2(1) = 0.58$, $p = .45$, Nagelkerke's $R^2 = .001$, as was the interaction, $\chi^2(2) = 1.10$, $p = .58$, partial Nagelkerke's $R^2 = .003$. Most of the eyewitness choice effect was from the planned contrast between the suspect identifications and the other two conditions, $\chi^2(1) = 80.66$, $p < .001$, Nagelkerke's $R^2 = .175$. The planned contrast between the filler and no identification conditions was non-significant, $\chi^2(1) = 0.04$, $p = .84$, Nagelkerke's $R^2 = .001$.

Similar effects were found for the belief in guilt variable (see Table 1 for means). The belief variable was negatively skewed, skewness = −.50, se = .10. Squaring the variable reduced the magnitude of its skew, skewness = .12, se = .10. This transformed variable was therefore used in all statistical tests. There was an effect of eyewitness choice, $F(2,567) = 67.36$, $p < .001$, $\eta^2 = .191$, due to the difference between choosing the suspect and the other two conditions, $F(1,567) = 135.15$, $p < .001$, $\eta^2 = .191$ and not differences between filler identifications and no identifications, $F(1,567) = 0.04$, $p = .85$, $\eta^2 = .000$. The main effect of double blind *versus* non-double blind was non-significant, $F(1,567) = 0.17$, $p = .68$, $\eta^2 = .000$, as was the interaction between the independent variables, $F(2,567) = 1.32$, $p = .27$, partial $\eta^2 = .005$.

Planned comparisons were done to compare double blind and non-double blind conditions separately for suspect identifications, filler identifications and non-identifications. The results are shown for the verdict and the belief in variables in the final column of Table 1. None were statistically significant.

Discussion

Psychology research showing that inadvertent cues can affect behaviour has resulted in double blind procedures being used as standard practice in much medical research. Psychologists have argued that suspects should be afforded the same precautions (British Psychological Society, 2007; Wells et al., 1998). While some jurisdictions now use double blind procedures, most do not. On the one hand, we thought it possible that lay people could recognize that non-double blind suspect identifications are less reliable than double blind suspect identifications, and as such they are less predictive of guilt. On the other hand, the

DOI: 10.1002/acp

fact that most people were willing to believe that a horse, Clever Hans, could perform mathematics meant we realized it was also possible (nah, probable) that people would not recognize the impact of having a non-double blind line-up procedure. Here, when non-double blind procedures were used suspect identifications produced 75% guilty verdicts while the double blind procedures produced 73% guilty verdicts. This small and non-significant difference is not commensurate with the psychology literature or with the large difference between double blind and non-double blind line-ups in Mecklenburg (2006).

It is worth reiterating that there were no differences between filler identifications and no identifications. Wells and Olson (2002) discuss how filler identifications should be more indicative of the suspect being innocent because the eyewitness is willing to choose somebody, and thinks someone else in the line-up looks more like the culprit than the suspect. Clark, Howell, and Davey (2008) discuss how this difference between filler identification and non-identification depends on how the line-up is constructed. The finding that people do not differentiate between these replicates past research (Wright, 2007).

STUDY 2—MANIPULATION CHECK

The null effects in Study 1 could have arisen for several reasons. One is that the participants did not process whether the line-up was double blind or not. We used a short scenario in hopes that participants would process all important aspects of the scenario, but clearly it is possible that people may not have processed *every* aspect. From an applied perspective, it is important to know if people were not processing the line-up administration information in the first place. From a theoretical perspective if people do not process whether the line-up was conducted double blind or non-double blind this would explain the results of Study 1, albeit with a theoretically uninteresting explanation.

Method

To examine whether people were generally processing the critical information in Study 1, we designed Study 2 to test whether at least 50% of the people process this information. Our critical measure is a true/false item. If 0% of the people process the information and are guessing, we would expect 50% to be correct. If 50% process the information (and all of these people answer correctly) and 50% are guessing, we would expect 75% to be correct. As this is the null hypothesis, we test whether our sample is higher than 75% (a one-tailed test). From G*Power, to detect an effect of 10% above 75% with $\alpha = .05$ and power of .80 a sample of 103 is suggested.

One hundred and twenty-nine participants were recruited in the same manner as Study 1, but during a different year. The sample was: 82% female; 74% Hispanic, 10% White and 16% other ethnicity or not specified; the mean age was 22 years (SD = 5 years).

Participants were randomly assigned to one condition of a $2 \times 2 \times 2$ between subjects design. The factors were: Whether the line-up was double blind or non-double blind, whether the eyewitness identified the suspect or not and a methodological test to detect whether it makes a difference asking if the person administering the line-up knew which person was the suspect (positive phrasing) or if the person administering the line-up did not know which person was the suspect (negative phrasing). This methodological factor is also necessary for assuming that half of the people who guess should be correct.

DOI: 10.1002/acp

The procedure was similar to Study 1. As with Study 1, participants read through the instructions, the crime scenario and then made a verdict and a belief in guilt judgment. Next, they were asked to freely recall everything that they could remember about the case. They were then asked six true/false questions. These were:

1. The rapist drove a Toyota Camry.
2. Hoyle's alibi was that he was at his parents' house.
3. Hoyle failed a lie detector test.
4. Nancy Von Roper picked Hoyle out of the line-up.

5a. The person administrating the line-up knew which person was the suspect.
or
5b. The person administrating the line-up did not know which person was the suspect.

6. Nancy Von Roper was able to correctly report part of the rapist's license plate number.

The focus is on the responses to questions 4 and 5, because these address the information presented differently across conditions. The responses to the other questions did not vary significantly among the conditions. Overall, 93% of responses were correct for questions 1–3. We did not analyse question 6 because participants could have different criterion for what constitutes 'correctly report part of . . .'.

Results

The percentage of guilty verdicts and belief in guilt judgments for the four conditions[1] are shown in Table 2. The analyses of the data in this table will follow the same format as used in Study 1. There was a main effect of witness identification on verdict decisions, $\chi^2(1) = 34.24$, $p < .001$, Nagelkerke's $R^2 = .311$, but the main effect for double blind *versus* non-double blind was not significant, $\chi^2(1) = 0.10$, $p = .76$, Nagelkerke's $R^2 = .003$. The interaction was also not significant, $\chi^2(1) = 0.07$, $p = .79$, partial Nagelkerke's $R^2 = .001$.

The belief variable was negatively skewed, skew $= -0.58$, se $= 0.21$. To decrease distribution skewness, the variable was squared, skew $= 0.15$, se $= .21$ and this transformed variable was used for the ANOVAs. There was a significant main effect of witness identification, $F(1,125) = 31.10$, $p < .001$, $\eta^2 = .195$. There was no main effect of line-up

Table 2. Percentage of guilty verdicts and percentage for belief in guilt for Study 2

	Guilty verdicts (%)	Belief in guilt
Non-double blind—filler picked	22.6 (21.4)	57.5 (56.5)
Non-double blind—suspect picked	70.4 (73.9)	70.7 (72.6)
Double blind—filler picked	22.9 (21.9)	45.0 (44.2)
Double blind—suspect picked	75.0 (75.8)	75.5 (75.9)

The numbers in parentheses are for the participants who correctly answered whether the line-up administration was double blind or not.

[1]The analyses are only for two of the dependent variables (eyewitness identification and line-up administration) because the third dependent variable (how question 5 was worded) occurred after these responses.

DOI: 10.1002/acp

administration, $F(1,125) = 0.09$, $p = .76$, $\eta^2 = .000$, and no significant interaction, $F(1,125) = 3.32$, $p = .07$, partial $\eta^2 = .026$.

We then examined responses to the fourth and fifth true/false questions (manipulation checks). The fourth question asked if the eyewitness identified the suspect and participants were generally accurate. Overall 87% were accurate: 97% of those who read that the eyewitness identified a filler were correct, and 76% of those who read that the eyewitness identified the suspect were correct. This difference was statistically significant, $\chi^2(1) = 87.08$, $p < .001$, Nagelkerke's $R^2 = .183$. There were no statistically significant differences between the two line-up administration conditions, $\chi^2(1) = 0.11$, $p = .74$, Nagelkerke's $R^2 = .002$.

The fifth question asked whether the line-up was double blind. Half the participants were asked to respond 'true' if the line-up was double blind and half were asked to respond 'true' if the line-up was not double blind. For half of the people in each of these conditions the correct answer was 'true' and for half it was 'false'. A logistic regression was done to predict whether people were correct. The main effect of question wording, $\chi^2(1) = 0.90$, $p = .34$, Nagelkerke's $R^2 = .017$; the main effect of whether the line-up was double blind or not, $\chi^2(1) = 0.46$, $p = .50$, Nagelkerke's $R^2 = .002$ and their interaction, $\chi^2(1) = 0.14$, $p = .71$, partial Nagelkerke's $R^2 = .002$; were not statistically significant.

Overall participants were 90% accurate (95% BCa bootstrap CI from 83 to 94%). This is significantly higher than the 75% which would correspond to half of the people having a memory for it (one-tailed binomial test with H0 = 75%, $p < .001$). This is about the same level of accuracy as for the fourth question, McNemar's $\chi^2(1) = 0.35$, $p = .56$.

The analyses for Table 2 were rerun on the verdict and belief in guilt variables excluding those participants (10%) who incorrectly answered the question about whether the line-up was double blind or not. The conclusions were the same (see percentages within parentheses in Table 2). For verdict, the effect for the suspect being picked was statistically significant, $\chi^2(1) = 34.80$, $p < .001$, Nagelkerke's $R^2 = .346$. The effect for double blind *versus* non-double blind, $\chi^2(1) = 0.02$, $p = .89$, Nagelkerke's $R^2 = .002$, and the interaction, $\chi^2(1) = 0.01$, $p = .94$, partial Nagelkerke's $R^2 = .000$, were non-significant. For belief, the effect for the suspect being picked was statistically significant, $F(1,112) = 32.53$, $p < .001$, $\eta^2 = .222$. The effect for double blind *versus* non-double blind, $F(1,112) = 0.35$, $p = .56$, $\eta^2 = .000$, and the interaction, $F(1,112) = 1.86$, $p = .17$, partial $\eta^2 = .016$, were non-significant.

Discussion

A possible explanation for the null effects of Study 1 was that participants were not processing whether the line-up administration was double blind or not. If people were not processing this information this could have important applied implications. Results from Study 2 show that people *were* processing this information. In fact, the accuracy for the true/false memory items were high, 90%, which is about the level of accuracy for remembering whether the suspect was chosen, which had a large effect on verdict and belief in guilt. Although the sample size was smaller than Study 1, it is important that the null effects of having a double blind line-up remained after removing participants who incorrectly answered this question. Thus, this study shows that the lack of double blind *versus* non-double blind line-up effects in Studies 1 and 2 is not due to failure to process critical information.

DOI: 10.1002/acp

STUDY 3—VARYING THE DEGREE OF NON-DOUBLE BLINDNESS

It could be argued that participants are processing whether the line-up is double blind or not, but assume that the administrator in a non-double blind line-up has no beliefs about whether the suspect is guilty or not. Perhaps participants assume that in a non-double blind line-up the administrator stays out of the eyewitnesses' sight to decrease any influence. We examined this notion by using five different descriptions of non-double blind line-ups. These are: Non-double blind without further description; the administrator believes the suspect is not guilty; the administrator is out of the eyewitnesses' sight; the administrator believes the suspect is guilty and the administrator helps the eyewitness.

Methods

The study had a 2×6 between subjects design. The first factor was whether the eyewitness identified the suspect or a filler. The second was how the line-up administration was described. A power analysis was done to help determine the sample size. To be able to detect an effect halfway between Cohen's small and medium values ($w = .2$) with power of .95 and α of .05 for a 5 degree of freedom test, G*Power suggests 495 total participants. Participant recruitment was the same as in Studies 1 and 2, with a different set of participants. Five hundred and sixty-six participants volunteered and completed the study. The demographics are similar to the previous studies (i.e. mean age 21 years, SD = 5 years; 69% female; 71% Hispanic, 11% White). Participants read the same scenario as in Study 2, but with the following differences.

Double blind
The officer, who did not know which photograph was of Hoyle, showed ...

Non-double blind
The officer, who knew which photograph was of Hoyle, showed ...

Non-double blind believes not guilty
The officer, who did not believe that Hoyle was guilty, showed ...

Non-double blind out of sight
The officer, who stood out of sight behind Nancy Von Roper during the line-up, showed ...

Non-double blind believes guilty
The officer, who believed that Hoyle was guilty, showed ...

Non-double blind administrator helps witness
The officer, who helped Nancy Von Roper throughout the procedure, showed ...

Participants were then asked to make a verdict decision (guilty or not guilty), a belief in guilt judgment on a 0–100 scale, and answer the six true/false questions about the scenario from Study 2.

Results

Table 3 shows the percentage of guilty verdicts and the mean belief in guilt broken down by conditions. The data analysis followed the same pattern as Studies 1 and 2.

DOI: 10.1002/acp

Table 3. Percentage of guilty verdicts and percentage for belief in guilt for Study 3

	Guilty verdicts (%)	Belief in guilt (%)
Suspect picked		
Double blind	71.4 (71.4)	74.4 (72.5)
Non-double blind	75.7 (75.9)	77.3 (78.8)
Believed not guilty	69.1 (73.2)	72.3 (76.9)
Out of sight	67.9 (66.7)	68.6 (62.7)
Believed guilty	60.5 (59.4)	69.0 (69.1)
Helped out	59.2 (60.0)	67.8 (68.8)
Filler picked		
Double blind	26.5 (20.0)	46.3 (46.8)
Non-double blind	15.4 (14.3)	48.9 (50.5)
Believed not guilty	24.0 (25.0)	51.3 (50.1)
Out of sight	34.9 (38.7)	52.2 (55.5)
Believed guilty	36.3 (34.1)	54.5 (53.5)
Helped out	29.8 (28.2)	46.3 (43.1)

The numbers in parentheses are for participants who correctly answered whether the line-up administration was double blind or not.

A logistic regression was used to predict verdict from eyewitness identification and description of line-up administration. The main effect of whether the eyewitness was identified was statistically significant, $\chi^2(1) = 91.42$, $p < .001$, Nagelkerke's $R^2 = .202$. The main effect of line-up administration was not statistically significant, $\chi^2(5) = 3.35$, $p = .65$, Nagelkerke's $R^2 = .008$. The interaction was also non-significant, $\chi^2(5) = 9.60$, $p = .09$, partial Nagelkerke's $R^2 = .024$. Because we are predicting null results it is worth exploring the non-significant findings further. We re-ran the analyses comparing the 'double blind' condition with each of the others. In each case, the effect of whether the suspect was identified was statistically significant, the effect of line-up administration was not statistically significant, and the interaction was not statistically significant.

As with Study 2, we also examined responses to the True/False questions. Overall, 77% of people correctly answered whether the line-up was double blind or not. The value was brought down by those who were told that the line-up administrator knew which person the suspect was, but that the administrator was out of sight. Those participants were more likely to say the administration was double blind than the other non-blind conditions, $\chi^2(1) = 15.43$, $p < .001$, Nagelkerke's $R^2 = .050$. The analysis of verdicts and belief judgments was repeated for people who correctly answered this question (see percentages in parentheses in Table 3). The inferential statistics were nearly identical to those reported above.

Discussion

The null effects from Study 1 could have arisen because people thought that even if administrators know who the suspect is, those administrators might not believe the suspect was guilty, or would not be a position to influence the eyewitness. In this study, we included non-blind conditions which suggested these situations, but also included a condition where we explicitly state the administrator thought the suspect was guilty and a condition where the administrator is described as trying to help the eyewitness. None of these conditions produced any significant differences. This suggests that people do not think that a line-up

DOI: 10.1002/acp

administrator can influence an eyewitness' choice. Our final study was therefore designed to explore whether it is possible, at all, to make people differentiate blind and non-blind line-ups.

STUDY 4—WITHIN SUBJECT DESIGN

Within subject designs increase the chances of observing effects in two ways. First, they produce context effects. If someone is presented with two identical phrases with only a single difference, the person is likely to notice the difference and to believe it is important (Grice, 1975). Second, within subject designs are generally more powerful than between subject designs even if the effect, in terms of the variable's units, is the same, providing the scores are correlated. Here we use a within-subjects design to maximize the chances of finding an effect, at the expense of ecological validity.

After participants completed Study 1, the 574 participants took part in an unrelated filler task for 10 minute, and then took part in this study. We found no carryover effects from the experimental conditions in study 1.

Method

A $2 \times 2 \times 3$ within subjects design was used. The factors were whether the suspect was the culprit or not, whether a double blind procedure was used or not, and whether the eyewitness chose the suspect, a filler or made no identification. Participants were told what we meant by a double blind line-up. Participants were then asked:

> Suppose there are 100 cases where the suspect is in fact guilty, and a double blind line-up is used. Out of the 100, how many times do you think the suspect will be picked? How many times do you think someone other the suspect will be picked? How many times do you think nobody will be picked?

The other three scenarios (when the suspect is not guilty and when non-double blind line-ups were used) were worded accordingly.

We had three hypotheses.

1. Participants will predict more suspect identifications in non-double blind line-ups, regardless of whether the suspect is guilty or not, because of the possibility of line-up administrators directing eyewitnesses (consciously or un-consciously) towards the suspect.
2. Participants will have fewer filler identifications in non-double blind line-ups, regardless of whether the suspect is guilty or not, because of the possibility of line-up administrators directing eyewitnesses (consciously or un-consciously) away from fillers.
3. Because hypothesis 1 should increase the number of suspect identifications and hypothesis 2 should decrease the number of filler identifications, we have no directional hypothesis about whether there should be a difference between double blind and non-double blind line-ups for the amount of non-identifications. This third hypothesis is exploratory.

Results

The top half of Table 4 shows the means for each condition and compares the double blind and non-double blind conditions. Logically the sum of responses within each scenario

DOI: 10.1002/acp

Table 4. The top half of the table shows the means for each condition, and differences between double blind and non-double blind conditions with their 95% confidence intervals

	How often picked out of 100 ...		
	Suspect	Filler	No identification
Guilty			
Double blind	52.16	34.48	15.06
Non-double blind	56.90	30.36	14.22
Difference (95% CI)	4.74 (2.86 to 6.61)	4.13 (2.62 to 5.64)	0.84 (−0.54 to 2.22)
Not guilty			
Double blind	39.11	42.23	18.96
Non-double blind	40.85	40.18	20.41
Difference (95% CI)	1.73 (−0.22 to 3.69)	2.05 (0.33 to 3.78)	1.44 (0.02 to 2.87)
ANOVA statistics			
Main effect guilty	$F(1,573) = 260.01$ $p < .001$ $\eta_p^2 = .31$	$F(1,573) = 117.89$ $p < .001$ $\eta_p^2 = .17$	$F(1,573) = 78.75$ $p < .001$ $\eta_p^2 = .12$
Main effect double blind	$F(1,573) = 20.12$ $p < .001$ $\eta_p^2 = .03$	$F(1,573) = 28.86$ $p < .001$ $\eta_p^2 = .05$	$F(1,573) = 0.31$ $p = .58$ $\eta_p^2 < .01$
Interaction	$F(1,573) = 5.25$ $p = .02$ $\eta_p^2 = .01$	$F(1,573) = 3.08$ $p = .08$ $\eta_p^2 = .01$	$F(1,573) = 6.16$ $p = .01$ $\eta_p^2 = .01$

Note: Ninety-five per cent confidence intervals asymptotic.
The bottom half shows the 2×2 within subject ANOVAs for the different line-up outcomes.

should be 100, but given that some participants could feel mathematics anxiety if we forced them to do this, we did not require this. Most participants' responses did sum to 100, so those responses were not independent. Therefore we analysed them separately for suspect identifications, filler identifications and no identifications.

The distributions were examined for all variables. Only those for the no identification conditions showed marked skewness (greater in magnitude than 1.0), the situation of least importance as no directional hypotheses were made. As the main interest was in the differences between the double blind and non-double blind conditions, and these were all relatively symmetric (mean skewness = −0.08), traditional ANOVAs on the untransformed variables were used for these analyses.

People indicated that the suspect will be chosen more often when s/he is guilty (around 55%) than when s/he is not guilty (around 40%). While in the predicted direction, the effect is small in comparison with other studies of the impact of eyewitness evidence on people's decision making (e.g. Loftus, 1974; Wright, 2007). Of most interest for the current study is the effect that a non-double blind line-up administration did significantly produce higher estimates of suspect identifications (hypothesis 1). This main effect has to be viewed in light of a significant interaction; the effect was larger when the suspect was guilty. The 95% confidence interval for the guilty suspect identifications spans from about 3 to 9%, while the interval for the not guilty suspect identifications spans from about 0 to 4%. Thus, it appears our participants thought that having the line-up administrator know which person is the suspect increases the chances of this person being chosen, but only if they are guilty.

Filler identifications showed the opposite pattern of the suspect identifications, as predicted (hypothesis 2). More fillers were predicted to be chosen in double blind line-ups

DOI: 10.1002/acp

than in non-double blind line-ups. There were fewer non-identifications than the other two categories and no significant main effect of line-up administration (hypothesis 3). There was a significant interaction corresponding to more non-identifications from double blind line-ups when the suspect is guilty, but fewer non-identifications from double blind line-ups when the suspect is not guilty.

Discussion

This large sample using a within subject design allowed us to detect an effect of using double blind *versus* non-double blind line-up administration. There are three ways to interpret the findings: A strict null hypothesis testing way, a skeptical way and a pragmatic way. If following a strict null hypothesis significance testing approach, a researcher could proclaim that participants only think suspects are at an increased risk of being chosen with non-double blind line-ups if they are guilty. This would be particularly worrisome as this suggests that participants would actually prefer non-double blind line-ups because they increase the chance of accurate suspect identifications without increasing the chance of inaccurate suspect identifications. They might believe, as some of the respondents in Mecklenburg's (2006) survey apparently do, that police can 'recognize relevant information during the identification procedure' (p. 57) so that they can direct eyewitnesses towards guilty suspects. However, numerous papers, books, journal guidelines, etc., have described how the significance testing approach is often short-sighted and it is not wise to adhere strictly to it (Wilkinson et al., 1999). The skeptic's interpretation of these data would be that the task of estimating the numbers of different types of identifications for different types of line-ups is so different from real-world situations that jurors are in when they interpret the reliability of an identification that the data should be ignored. As applied researchers we sympathize with this view, but our original intention was to test whether it was possible to detect any effects of double blind *versus* non-double blind line-up administration in any situation. We felt it was worth sacrificing ecological validity in this study to make this attempt.

The pragmatic approach concentrates on the size of any effects while considering the context within which they are observed. The task was designed to increase the chances of an effect so any estimates are likely to be larger than is typical when people are asked to make a single judgment about either a double blind line-up or a non-double blind line-up. For example, if one group of people are asked 'how often they are satisfied with their lives?' and another group is asked 'how often they are very satisfied with their lives?', responses for the two groups are approximately the same (O'Muircheartaigh, Gaskell, & Wright, 1993). However, if the same people are asked both questions together (in a within subject design), response patterns differ because people focus on the *difference* between the questions: More people are 'satisfied' than are 'very satisfied'. Similarly, asking people about both double blind and non-double blind line-up administration will direct their focus to their difference and hence magnify any possible effect of this variable. Even with this powerful context, the effects were small. The largest shift was for suspect-guilty identifications with only 5 out 100 different identifications, and the smallest effects were near zero. The effect in Mecklenburg (2006) was approximately 15%. The main conclusion from our data is that while it is possible to create an effect for how people view double blind *versus* non-double blind line-up administration differently, even with a study designed to maximize the effect, the size of the effect is small and inconsistent.

DOI: 10.1002/acp

GENERAL DISCUSSION

Loftus (1974) and many others have shown that when considering different types of evidence in a criminal case people weigh eyewitness identifications highly when judging the guilt of defendant (Winter & Greene, 2007). Loftus (1974) showed that the weight given to eyewitness evidence did not substantially decrease when it was made clear that the identification was unreliable. Much research over the last three decades has examined the conditions under which eyewitness identifications are likely to be more or less reliable. There have been numerous reports and recommendations made for conducting line-ups. While some of the recommendations have been controversial among psychologists (e.g. on using sequential or simultaneous line-ups, McQuiston-Surrett et al., 2006), a consistent recommendation is that double blind line-up administration should be used (British Psychological Society, 2007; Wells et al., 1998). The science is unequivocal that people transmit information unintentionally to others and this argument has persuaded many in the medical sciences. No one should trust the effectiveness of a drug if the only studies showing positive effects were conducted when the experimenter knew which patients had the placebo and which patients had the drug. The argument from Wells and colleagues is that we should afford suspects the same level of protection as we do patients (Wells et al., 1998). However, law enforcement officers typically state that they are already so over-stretched that even the minimal amount of extra effort required for double blind line-up administration would be difficult to implement (Mecklenburg, 2006). Thus, despite the recommendations it appears that non-double blind line-ups will continue to be conducted in many jurisdictions.

Therefore, it is clearly important to see whether people recognize the inherent bias and unreliability of non-double blind line-ups. Given that so many people were duped into thinking Clever Hans (Sebeok & Rosenthal, 1981) was able to perform mathematics, we felt that people might not differentiate between the reliability of double blind and non-double blind line-ups. Our first study gave people a brief crime scenario where either a double blind or a non-double blind line-up was used. Because of our belief that any effects might be small, we used a relatively large sample, and still found non-significant effects. The second study showed that this null effect was not due to participants simply not processing this information. The third study showed that the null effects remained in a variety of line-up administration contexts, including when participants were told that the administrator helped the eyewitness. An important caveat, however, is that we used only a single case summary. We urge further research to examine if these null effects remain in other contexts.

The final study involved participants rating several different scenarios including both double blind and non-double blind line-up scenarios and as such was designed to maximize our chances of detecting an effect at a cost of ecological validity. While an effect was detected, it was small. It was also only present when the suspect was presumed guilty. When the suspect was not guilty, participants said that 39% of the time s/he would be chosen in a double blind line-up and 41% in a non-double blind line-up. While in the predicted direction, the effect is both non-significant (despite $n = 574$ and a within subject design) and small.

Our conclusions are that the effect of any differences between double blind and non-double blind line-up administrations on how people weigh eyewitness identifications is minute. Even our first study, using a short crime summary, was designed to produce a larger effect than might occur in some situations, and the effect was small and non-significant.

DOI: 10.1002/acp

The final study was designed to maximize the effect, and while it was statistically significant it was small. It is questionable whether Clever Hans could ever amaze audiences with his mathematical abilities again because it is likely skeptics would mention how society had already been duped once. However, our data suggest that people are still not aware of the communicative powers that people can inadvertently (or even purposely) employ.

The implications of this research are clear. It is critical to stress to people who rely on eyewitness identifications that non-double blind line-ups are less reliable than double blind line-ups. This is a problem for law enforcement because it could lead them down incorrect investigative routes and therefore be costly in terms of investigative time and effort, and allow the true culprit to evade justice. Perhaps if law enforcement officers became more aware of the unreliability and potential costs, the perceived hassle that they report for conducting double blind line-ups would seem less. It is also a problem for jurors as the ultimate triers of fact in many criminal cases. If simply explaining to jurors what double blind and non-double blind line-ups are and which was used produced appreciable effects then at least at this stage of judicial proceedings there would be a safeguard. However, our studies show that this is not the case. It is necessary for jurors to weigh the identifications more appropriately. There are two options for this: Expert testimony and judicial instructions. Expert testimony is not viable simply because of the number of cases involving non-double blind line-ups. There are not enough memory experts to provide testimony for all cases. Therefore, the next step needs to be to write judicial instructions and to calibrate the instructions to the size of the double blind *versus* non-double blind effect found in the field. If the Mecklenburg (2006) study is used to provide an estimate, the suspect identifications rose by about 15% in non-double blind conditions. However, the plethora of errors in the Mecklenburg study makes it unwise to rely on those data. Further studies are necessary to provide estimates of this effect so that the size of the effect can be used to help guide the value placed on identifications from non-double blind line-up.

REFERENCES

British Psychological Society. (2007). British psychological society submission to the home office consultation on 'Modernising Police Powers: Review of the Police and Criminal Evidence (PACE) Act 1984'. http://www.bps.org.uk/downloadfile.cfm?file_uuid=32109AE2-1143-DFD0-7E9E-4B5091B061CA&ext=pdf accessed 17 May 2009.

Carlson, C. A., Gronlund, S. D., & Clark, S. E. (2008). Lineup composition, suspect position, and the sequential lineup advantage. *Journal of Experimental Psychology: Applied*, *14*, 118–128.

Clark, S. E., Howell, R. T., & Davey, S. L. (2008). Regularities in eyewitness identification. *Law and Human Behavior*, *32*, 187–218.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Douglass, A. B., Smith, C., & Fraser-Thill, R. (2005). A problem with double-double blind photospread procedures: Photospread administrators use one eyewitness's confidence to influence the identification of another eyewitness. *Law and Human Behavior*, *29*, 543–562.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about Internet questionnaires. *American Psychologist*, *59*, 93–104.

DOI: 10.1002/acp

Greathouse, S. M., & Kovera, M. B. (2009). Instruction bias and lineup presentation moderate the effects of administrator knowledge on eyewitness identification. *Law and Human Behavior*, *33*, 70–82.
Grice, H. P. (1975). Logic and conversation. In P. Cole, & J. L. Morgan (Eds.), *Syntax and semantics: Speech acts* (pp. 41–58). New York: Academic Press.
Gronlund, S. D. (2005). Sequential lineup advantage: Contributions of distinctiveness and recollection. *Applied Cognitive Psychology*, *19*, 23–37.
Haw, R. M., & Fisher, R. P. (2004). Effects of administrator-witness contact on eyewitness identification accuracy. *Journal of Applied Psychology*, *89*, 1106–1112.
Lindsay, R. C. L., & Wells, G. L. (1985). Improving eyewitness identification from lineups: Simultaneous versus sequential lineup presentations. *Journal of Applied Psychology*, *70*, 556–564.
Loftus, E. F. (1974). Reconstructing memory: The incredible eyewitness. *Psychology Today*, *8*, 116–119.
Loftus, E. F., & Ketcham, K. (1991). *Witness for the defense: The accused, the eyewitness, and the expert who put memory on trial*. NY: St. Martin's Press.
Margraf, J., Ehlers, A., Roth, W. T., Clark, D. B., Sheikh, J., Agras, W. S., et al. (1991). How ''blind'' are double-blind studies? *Journal of Consulting and Clinical Psychology*, *59*, 184–187.
McGraw, K. O., Tew, M. D., & Williams, J. E. (2000). The integrity of web-delivered experiments: Can you trust the data? *Psychological Science*, *11*, 502–506.
McQuiston-Surrett, D., Malpass, R. S., & Tredoux, C. G. (2006). Sequential vs. simultaneous lineups: A review of methods, data and theory. *Psychology, Public Policy and Law*, *12*, 137–169.
Mecklenburg, S. H. (2006). Report to the legislature of the state of Illinois: The Illinois pilot program on sequential double-blind identification procedures. http://www.psychology.iastate.edu/~glwells/Illinois_Report.pdf, accessed 17 May 2009.
Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*, 691–692.
Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*, 231–259.
O'Muircheartaigh, C. A., Gaskell, G. D., & Wright, D. B. (1993). Intensifiers in behavioral frequency questions. *Public Opinion Quarterly*, *57*, 552–565.
Phillips, M. R., McAuliff, B. D., Kovera, M. B., & Cutler, B. L. (1999). Double-double blind photoarray administration as a safeguard against investigator bias. *Journal of Applied Psychology*, *84*, 940–951.
Rosenthal, R. (2003). Covert communication in laboratories, classrooms, and the truly real world. *Current Directions in Psychological Science*, *12*, 151–154.
Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom*. New York: Holt, Rinehart & Winston.
Rosental, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences*, *1*, 377–386.
Sebeok, T. A. & Rosenthal R. (Eds.). (1981). *The Clever Hans phenomenon: Communication with horses, whales, apes, and people*. New York: The New York Academy of Sciences.
Spinney, L. (2008). Line-ups on trial. *Nature*, *453*, 442–444.
Sporer, S. L., & Schwandt, B. (2007). Moderators of nonverbal Indicators of deception: A meta-analytic synthesis. *Psychology, Public Policy, and Law*, *13*, 1–34.
State of Wisconsin v Tyrone L. Dubose (2005 WI 126). http://www.wicourts.gov/sc/opinion/DisplayDocument.html?content=html&seqNo=19016, accessed 17 May 2009.
Steblay, N. (2009). It's more complicated than that! Lessons from the Evanston IL field data. *Paper presented at the American Psychology-Law Society conference*, San Antonio, TX, March 2009.
Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities*. Chichester, England: Wiley.
Wells, G. L. (1988). *Eyewitness identification: A system handbook*. Toronto: Carswell.
Wells, G. L., & Olson, E. (2002). Eyewitness identification: Information gain from incriminating and exonerating behaviors. *Journal of Experimental Psychology: Applied*, *8*, 155–167.
Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. A. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior*, *22*, 603–647.
Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling in social psychological experimentation. *Personality and Social Psychology Bulletin*, *25*, 1115–1125.

DOI: 10.1002/acp

Wilkinson, L., and the Task Force on Statistical Inference APA Board of Scientific Affairs. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.
Winter, R. J., & Greene, E. (2007). Juror decision making. In F. T. Durso, R. S. Nickerson, S. T. Dumais, S. Lewandosky, & T. J. Perfect (Eds.), *Handbook of Applied Cognition* (2nd ed., pp. 739–761). Chichester, UK: John Wiley & Sons.
Wright, D. B. (2007). The impact of eyewitness identifications from simultaneous and sequential lineups. *Memory*, *15*, 746–754.
Wright, D. B., & Hall, M. (2007). How a "Reasonable Doubt" instruction affects decisions of guilt. *Basic and Applied Social Psychology*, *29*, 85–92.
Wright, D. B., Hanoteau, F., Parkinson, C., & Tatham, A. (in press). Perceptions about memory reliability and honesty for children of 3–18 years old. *Legal and Criminological Psychology.*

APPENDIX

Crime scenario for Studies 1–3. For Studies 2 and 3, the car was described as a Honda Civic rather than a Chevette. Other differences are described in the text.

Please read carefully

Seventeen-year-old Nancy Von Roper was hitch-hiking at about 6:30 pm on the 12th of October, 1980. A light-blue compact car, driven by a bearded man, stopped and she was offered a ride. After a few minutes he exited the motorway, before he should have for the location she thought that she was being taken. When she looked confused he said 'I have to stop and see my sister'. He drove some more and turned down a dirt road. He pulled to the side of the road. She felt an object at her throat. He said 'Do as I say, or I'll hurt you'. He had her undress and then raped her. He then had her get out of the car and he drove off.

She called the police at 7:22 pm. She described the rapist as 25–30 years old, six foot, medium build, full beard and wearing a three-piece cream-coloured suit. The car was described as blue with temporary license plates displayed on the rear window. She remembered something hanging from the rearview mirror and a brown folder on the backseat. She showed police where the rape occurred. Tire tracks were photographed. At 1:20 am, police spotted a light-blue Chevette with temporary plates parked near to where the rape occurred. Tom Hoyle was in a restaurant with his fiancée, Gretchen Abraham. Hoyle was a manager of a restaurant chain, and drove the light-blue Chevette, a company car. Hoyle also fit the basic description given by Nancy Von Roper, except for being only 5′8″ (the victim never saw the person standing). The couple drove off and were soon pulled over by police.

The police asked Hoyle where he had been in the afternoon and evening. He explained he had been at his parents' house—it was his father's birthday—until 6:10 pm. He went to his best friend's to watch TV at about 6:50 pm and made some phone calls. Telephone records verified that there was a call made at 7:00 pm from the house. Given the distance between his parents and his friend's, it would have been unlikely that he could have committed rape in this time. He left his friend's house to pick up Gretchen at 9:20 pm. He gave police permission to search his car. He also gave them permission to take his photograph.

Tom Hoyle was arrested and his car impounded. He failed a lie detector test when asked if he had raped Nancy Von Roper. This can occur for a number of reasons, including the person telling the truth but being nervous. The victim stated that the number of the

DOI: 10.1002/acp

temporary license plate was either 667 or 776. The number on the license in Hoyle's car was six digits, 661–677. It is unlikely that there would be another automobile fitting the description with such a similar temporary license number. The police also reported that there was a folder like the one the victim described in the back of Hoyle's car.

A police officer took pictures of five other men, all with beards, between 25 and 30, and about the same build as Hoyle. On the 13th, one of the police officers involved in the case showed these pictures with the one of Tom Hoyle to Nancy Von Roper. The officer, who knew which photograph was of Hoyle, the suspect, showed all six photographs at the same time. Nancy Von Roper identified Tom Hoyle's photograph as the person who raped her.

In criminal trials, in order to make a verdict of guilty, you must believe in guilt beyond a reasonable doubt. Please tick appropriate for the verdict that you would give if you were a juror in this case and had this evidence presented.

[ ] Guilty [ ] Not Guilty

Using a 0–100% scale, how strongly do you believe that Tom Hoyle is the person who raped Nancy Von Roper. Zero per cent means that you are sure that he is not the rapist and 100% means that you are sure that he is the rapist.

Belief (write a percentage between 0 and 100%) _____

DOI: 10.1002/acp