

The Influence of Delay and Item Difficulty in Criminal Justice Systems on Eyewitness Confidence and Accuracy

Jacqueline M. Wheatcroft^{1,*}, Graham F. Wagstaff¹ and Brian Manarin²

¹*Institute of Psychology, Health & Society, University of Liverpool, UK*

²*Ministry of the Attorney General, Windsor, ON, Canada*

Abstract: There is international concern about the negative effects of delays in Criminal Justice Systems. Problems include the deleterious effects that delay can have on witnesses' memory accuracy and witnesses' ability to calibrate their memories accurately. Little empirical work has been conducted on these issues combined with item difficulty and the relationship between accuracy and confidence. This paper investigates these issues.

21 witnesses were interviewed about an observed crime and required to answer lawyerly questions used in cross-examination relating to target items classified as 'easy', 'moderate' and 'difficult', in terms of memorability. Participants were interviewed again, 6 months later. A 6 month delay significantly reduced memory accuracy for all levels of question difficulty. Within-subjects C-A relationships seemed to be relatively unaffected by delay; i.e. they tended to be positive for easy and moderate items, and negative for difficult items. Between-subjects C-A relationships were also positive for both easy and moderate items, but improved after 6 months; whereas C-A relationships for the difficult items remained negative and statistically insignificant following the 6 month delay. Delay can have a profound negative effect on witness accuracy that is not likely to be compensated for by improvements in C-A calibration.

Keywords: Criminal justice, delay, witness, confidence, accuracy.

INTRODUCTION

It has long been recognised, internationally, that delays in criminal justice systems are endemic and problematic (Samuels 1997; Manarin 2009). For example, in 1999, the UK Home Office produced a protocol report in an attempt to reduce delays in the Youth Justice System (Home Office 1999); this was followed by two further reports focused upon this process element (Brown 2000; Ernst and Young Independent Consultants 1999). The latter report was originally borne from the Prison Service chief Martin Narey's review of delay in 1997. In New Zealand too, particular concern has been expressed about delays that child witnesses face in giving evidence in court (Hanna, *et al.* 2010). Also, in recognition of the problems associated with delays, in 2008, in Ontario, Canada, the Ministry of the Attorney General announced its 'Justice on Target' initiative to reduce delays in the province's courts, which was deemed by many legal practitioners/scholars as long overdue (Manarin 2009). And, most recently, the judiciary in England and Wales have introduced a 'Stop Delaying Justice' initiative' (see Riddle 2012), which aims to reduce delays in the system; it is intended that every magistrate, legal advisor and prosecutor will be trained in the requirements of the scheme.

There are a number of reasons why the reduction of delays in the system might be beneficial. For example, delays may affect general confidence in the legal system, may cause protracted frustration, worry and distress to those involved, and may prevent victims of crime, and those associated with them, from 'moving on' (Hanna, *et al.* 2010; Manarin 2009). However, perhaps most significantly, it has been argued that delay may have an adverse effect on the memory of those required to give evidence. As Manarin (2009:125) has emphasized, the most valuable commodity possessed by a witness called to testify at a criminal trial is his or her memory. This is because, when credibility is not an issue, accurate memory should help guide the trier of fact towards an accurate verdict; hence he says, 'When a trial takes place without unreasonable delay, with all witnesses available and memories fresh, it is far more certain that the guilty parties who committed the crimes will be convicted and punished and those that did not, will be acquitted and vindicated'.

LITERATURE REVIEW

Empirical evidence from the psychological literature suggests that concern about effects of delay on memory may be justified. For example, a variety of evidence indicates that people typically remember less about an event when recall takes place after a delay; moreover, the deterioration in recall performance as a result of delay is often observed even when participants have been given an earlier opportunity for

*Address correspondence to this author at the Institute of Psychology, Health & Society, Department of Psychological Sciences, Witness Research Group, University of Liverpool, Eleanor Rathbone Building, Bedford Street South, Liverpool L69 7ZA, UK; Tel: 44 (0)151 795 0513; E-mail: jacmw@liv.ac.uk

recall (Flin, Boon, Knox and Bull 2011; La Rooy, Pipe and Murray 2005; Lipton 1977; Turtle and Yuille 1994).

However, in courtroom situations, in particular, the perceived accuracy of witness memory is influenced fundamentally by the confidence that the witness displays in that memory. If a witness recalls a fact, but expresses no confidence whatsoever in the reliability of his or her memory for that fact, it is unlikely to be given any credence by a jury. Indeed, in the absence of other forensic information, confidence is the main indicator used by jurors to determine the accuracy of a witness's memory (Wells 1985; Wells, Lindsay and Ferguson 1979; Wheatcroft, Wagstaff and Kebbell 2004). However, few studies have actually examined the effects of delay on the relationship between confidence and accuracy, and those that have, have produced conflicting results. For example, two studies found that repeated questioning of the same material increased confidence after 2 days (Hastie, Landsman and Loftus 1978), and after 3 weeks (Turtle and Yuille 1994), regardless of accuracy. In contrast, Ryan and Geiselman (1991) reported that confidence decreased following repeat questioning after 1 week, but more so for incorrect than correct answers; i.e. there was better correspondence between confidence and accuracy over time. However, Granhag (1997) suggests that the positive effect after a short one week delay (1 week) on confidence-accuracy (C-A) may have been more a function of making repeated confidence ratings *per se*, than the effects of the delay involved. In support of this interpretation, Granhag (1997) found that whether C-A relationships improved after a 1 week delay rested crucially on whether participants were repeatedly tested, or tested for the first time after a week. If they were tested once shortly after the event, and then, a week later, given an opportunity to review their previous responses, and tested again, calibration improved; however, if they were tested for the first time after a week, it deteriorated. However, as yet, no work has examined the effects of repeated recall over a time period more akin to that experienced by witnesses in the Criminal Justice System, such as, six months or more. Even if repeated testing after a week can sometimes increase C-A relationships, will it still do so after 6 months?

Significantly also, as yet, no research has examined the possible differential effects of delay and memory on accuracy and confidence for different kinds of target items. As background to this, it can be noted that general findings regarding the relationship between confidence and accuracy in witness memory have been

mixed. For example, a number of reviews have suggested that there is either no relationship, or only a small positive relationship between witness confidence and accuracy (Bothwell, Deffenbacher and Brigham 1987; Deffenbacher 1980; Fruzzetti, Toland, Teller and Loftus 1992; Penrod, Loftus and Winkler 1982; Sporer, Penrod, Read and Cutler 1995; Wells 1993; Wells and Murray, 1984). Nevertheless, other research suggests that C-A relationships vary according to a variety of factors, including the method used to calculate the correlations. For example, in an attempt to explain negative findings, Smith, Kassin and Ellsworth (1989) suggested that researchers have concentrated too much on 'between subjects' confidence-accuracy relationships which involve calculating an average confidence and accuracy score for each participant, and then comparing the accuracy of confident witnesses to less confident witnesses; this produces a single between subject correlation for the group. Smith, *et al.* suggest that higher correspondence between confidence and accuracy might be obtained by examining the relationships within participants' own judgments by calculating a separate confidence-accuracy correlation for each participant; the individual correlations can then be averaged to produce a single mean 'within subjects' correlation for the group. Moreover, arguably, as within-subjects relationships provide a finer discrimination for each witness, they maybe more meaningful in real life courtroom situations, especially when there are only a few key witnesses. Nevertheless, Smith, *et al.*, found, using a forced choice task, that both within- and between-subjects C-A correlations were still very low. Nevertheless, Perfect, Watson and Wagstaff (1993) repeated Smith, *et al.*'s study with a control for guessing, and found that within subjects C-A correlations were higher than between subjects C-A correlations.

In a further attempt to explain the variability in C-A findings, therefore, Kebbell, Wagstaff and Covey (1996) argued that a critical variable may be item difficulty. They suggested that, in laboratory studies, to avoid ceiling effects, researchers will tend to avoid 'easy' questions that more or less anyone will get right. However, little relationship will be found between confidence and accuracy when witnesses are asked to remember a relatively homogeneous pool of difficult or moderately difficult items (for example, items concerning peripheral detail such as details of mailboxes or pictures); but if variance is increased through the inclusion of items that are easier to

remember (for example, items reading the sex of the perpetrator, or whether it was night or day at the time), confidence-accuracy relationships may improve. A number of studies have found support for these predictions for a variety of stimulus conditions including eyewitness identification; they also indicate that C-A correlations tend to be higher for easy items generally, because easy items tend to be remembered in an 'all or none' fashion (Kebbell, *et al.* 1996; Lindsay, Read and Sharma 1998; Wheatcroft, *et al.* 2004). These findings suggest that any study looking at the effects of delay on memory in a legal context should examine, not only between and within subjects C-A correlations, but also the effects of using items of varying difficulty.

A final consideration, as yet unexamined in research on the effects of delay on witness memory, is the use of what is known as the 'lawyerese' questioning style (Wheatcroft, Kebbell and Wagstaff 2001; Wheatcroft, *et al.* 2004; Wheatcroft 2012). Cross-examination procedures have long been thought by the legal profession to be crucial for probing the accuracy of evidence obtained in the examination-in-chief, and to expose unreliable or dishonest witnesses (Stone 1988). However, despite an extensive psychological literature pointing to the dangers of leading questions in producing memory distortions (see, for example, Loftus 1979; 2003), a firm rationale has developed in legal culture whereby leading questions may be permitted during cross-examination (Keane and Fortson 2011; Wheatcroft 2012). Moreover, it is generally contended that asking questions containing false pre-supposition is a normal, useful, and effective procedure for verifying doubtful information and introducing new information (Hickey 1993); i.e. the admissibility of leading questions seems to be based upon the notion that they serve to calibrate or assess the memories of witnesses. However, leading questions of the type asked during cross-examination are not only usually suggestive to a degree, but also tend to limit responses made to a two-alternative forced choice alternative (i.e. yes/no), with a particular emphasis on the encouragement of affirmative responses ('e.g. 'The car was black, wasn't it?', 'You would agree that'). This not only encourages 'yea saying', but gives witnesses little opportunity to elaborate or expand on their answers (Harris 1984; Kebbell, Deprez and Wagstaff 2003, Taylor 2004). Consequently, serious concerns have been raised with regard to basic paradigms of justice and fairness in that 'lawyerese' questions can suggest or compel responses not made to other less directive forms of questioning (Brennan 1995; Wheatcroft and

Woods 2010; Wheatcroft and Ellison 2012). Perhaps not surprisingly, this particular kind of question style during interview has been demonstrated to have the most detrimental effect on confidence-accuracy relationships (Wheatcroft, *et al.* 2004). Despite this knowledge, leading questions put in cross-examination are still considered by lawyers as part of "the greatest legal engine ever invented for the discovery of the truth (Wigmore, 1940; p. 29); though see Wheatcroft, Caruso and Krumrey-Quinn - forthcoming

In view of these considerations, to mirror more closely the operation of the legal system, the following study investigated the effects of a six month delay on memory, examining both accuracy and confidence, when participants were interviewed using 'lawyerese' questioning.

CURRENT STUDY

Given that in general memory tends to decline over time (Jonides, *et al.* 2007) one might reasonably expect that both confidence and accuracy in memory to reduce over a long delay. However, it would not necessarily follow that the relationship between confidence and accuracy would decrease (i.e. C-A correlations would be reduced). Some studies have shown that, when participants are absolutely sure of a response, they invariably tend to be accurate (Kebbell, *et al.* 1996; Wheatcroft, *et al.* 2004). Arguably, such responses may be least likely to decline over time, not only because of the strength or salience of the original association, but also because they might be the most obvious candidates for rehearsal. Consequently, although there might be a decline in the overall number of correct responses over time, there might also be a corresponding increase in confidence-accuracy relationships as the distinction between the two types of response (accurate-confident, inaccurate-unsure) becomes more exaggerated (see also Ryan and Geiselman 1991). However, item difficulty might be a crucial variable here. For difficult items, where confidence may be low in the first place, further reductions in confidence and increasing ambiguity about memory for items, delay might intensify poor C-A correspondence. Thereby the research investigated the following hypotheses (H); H1: A decline in confidence will be found to be associated with delay; H2: A decline in accuracy will be found to be associated with delay; H3: C-A relationships will improve over time; H4: Poorer C-A correspondence will be observed for difficult items.

METHOD

Participants

The participants were 4 males and 17 females drawn from an opportunity sample. The mean age of the sample was 20.95 ($SD = 5.45$; range = 18-44). All participants were either students at the University of Liverpool or members of a research panel in the School of Psychology at the University.

Materials and Procedure

All participants were told that they were to observe a videotape of an event for a period of around 4-5 minutes, and afterwards, they would be required to answer some questions. All participants were then shown, individually, a five minute colour video depicting a criminal offence in which a woman at a bus stop is abducted and forced into a car. During the incident a gun is pointed at a witness. Following this each participant was required to complete a filler task for a period of five minutes, which involved reading unrelated material.

Participants were then interviewed individually by an interviewer who had undergone training to learn to modulate the voice consistently throughout the interviews. In the interview, participants were required to give answers to three sets of target items; 14 were designed to be 'easy', 14 'moderate' and 10 'difficult'. Item difficulty was determined using accuracy data from previously published studies (Wheatcroft, *et al.* 2001; Wheatcroft, *et al.* 2004). However, unlike in most previous studies, the questions were phrased so as to replicate the manner in which lawyers conduct cross-examinations in court; i.e. they involved 'lawyerese' questioning. The phrasings were taken directly from the examination of several Crown Court transcripts. For example, for an easy question, instead of asking the witnesses, 'did two men carry out the attack on the victim?', they were asked, 'do you also remember that two men carried out the attack on the victim?' Similarly, for a moderate question, instead of asking witnesses, 'did the victim have long hair?', they were asked, 'you would agree that the victim's hair was long?' And, for a difficult question, instead of being asked, 'would you say that this car had four doors?', witnesses were asked, 'isn't it also right this car had four doors?'. For answers to be correct, all required an appropriate affirmative or negative response (half affirmative and half negative within each category of item difficulty).

As noted previously, questions that prompt simple 'yes' or 'no' responses, but lean, in particular, towards

the elicitation of affirmative responses, are typical of 'lawyerese' questioning. However, lawyers also tend to target critical items that are associated with inconsistency both within and between witnesses' statements. Given that critical items are items about which there may be some ambiguity, this will most likely to be the case for items that are difficult to remember. From the cross examining lawyers' point of view, the strategy of targeting difficult items makes sense, as cross-examiners are considerably less likely to influence the witness' response to a very easy item, especially if the witness is perceived to be potentially harmful to the defence of an accused person. However, in doing so, lawyers will tend to target only a few critical items to prevent their strategy from becoming too transparent. Consequently, when wishing to cast doubt on the reliability of a particular witnesses' testimony overall, they will tend target fewer difficult items and set these up as exemplars. This procedure was therefore adopted in the present study; i.e. there were fewer items in the 'difficult item' category.

Following each question, participants were also asked to rate their confidence in the response they had given on a Likert scale ranging from 1 to 9, where (1) represented 'pure guess' and (9) represented 'absolutely certain'. All interviews were transcribed. At the end of the interview schedule, each participant was thanked for his/her participation and debriefed.

RESULTS

Total accuracy (total number of items correct), mean confidence (mean scores on the 1-9 Likert scale; where 1 represented 'pure guess' and 9 represented 'absolutely certain'), within subjects confidence-accuracy (C-A) correlations (mean of the point-biserial confidence-accuracy correlations for each participant), and between subjects confidence-accuracy correlations (the Pearson's correlation between total correct responses for each participant and mean confidence scores for each participant, for the group as a whole) were calculated. The overall results (i.e. easy, moderate and difficult items combined) are shown in Table 1.

Preliminary analysis showed that overall accuracy was significantly lower at 6 months ($t = 6.67$, $df=20$, $p=.001$). Also, between subjects C-A was found to be significant only at 6 months ($r=.52$, $p<.02$).

As there were unequal numbers of easy, moderate and difficult items, to explore the effects of item

Table 1: Overall Results for Mean Total Accuracy, Mean Confidence, Mean Within-Subjects (w-s) C-A correlations, and Group Between-Subjects (b-s) C-A Correlations, at 5 Minutes and 6 Months

		Accuracy	Confidence	C-A (w-s)	C-A (b-s)
5 Minute Delay (n=21)	Overall	27.95 (2.16)	6.29 (1.00)	.35 (.14)	.12
6 Month Delay (n=21)	Overall	23.48 (2.79)	4.55 (1.25)	.31 (.15)	.52*

Note: Standard deviations are shown in parenthesis; * $p < .05$.

difficulty, the data for correct answers were transformed to percentage accuracy scores and analysed using a 3X2 mixed ANOVA (easy/moderate/difficult X 5 minutes/6 months), with repeated measures on the second factor.

As expected, given the data already analysed (see Table 1), a main effect was found for delay $F(1,40)=38.41$, $p < .001$, $\eta^2_p = .49$; overall percentage accuracy was lower after 6 months ($M=58.75$, $SD=19.12$) than after 5 minutes ($M=70.95$, $SD=12.13$). A significant main effect was also observed for item difficulty, $F(1.36, 54.27)=71.23$, $p < .001$, $\eta^2_p = .64$ (n.b. Mauchly's test indicated that the assumption of sphericity had been violated; therefore, degrees of freedom were calculated using the Greenhouse-Geisser correction). Further univariate F comparisons showed that easy ($M=86.56$, $SD=1.63$), moderate ($M=69.89$, $SD=2.62$) and difficult items ($M=38.09$, $SD=3.23$) all differed significantly from each other ($p < .05$); i.e. easy items were answered most accurately, and difficult items, least accurately. These results further validated the selections of items. No significant interaction was observed, $F(1.36, 54.27) = .44$, $p > .05$, $\eta^2_p = .01$; i.e. delay did not differentially affect accuracy across the three item difficulty categories (see Table 2).

A similar 3X2 mixed ANOVA was also conducted on the mean confidence scores. A main effect of delay was found (see Table 1), $F(1,40) = 23.91$, $p < .001$, $\eta^2_p = .37$; i.e. overall, confidence was lower after a 6 month delay. In addition, a significant main effect was shown for item difficulty, $F(2,80) = 101.78$, $p < .001$, $\eta^2_p = .72$; confidence was highest for easy items ($M=6.49$, $SD=.18$). Further univariate F comparisons showed that easy items differed from moderate ($M=4.80$, $SD=.20$) and difficult items ($M=4.80$, $SD=.20$); though moderate and difficult items did not differ from each other ($p > .05$). No interaction was observed, $F(2,80)=2.26$, $p > .05$, $\eta^2_p = .05$.

A further 3X2 on the within-subjects C-A correlations showed no main effect for delay, $F(1,40) = .62$, $p > .05$, $\eta^2_p = .02$ (see Table 1). However, there was a significant main effect for item difficulty, $F(2,80)=50.10$, $p < .001$, $\eta^2_p = .56$. Further F comparisons showed that the mean correlations for easy ($M=.42$, $SD=.29$) and moderate ($M=.43$, $SD=.27$) items were significantly higher than those for difficult items; indeed, overall, the latter were negative ($M=-.16$, $SD=.35$). The correlations for easy and moderate items did not differ from each other ($p > .05$). The interaction with delay was not significant, $F(2,80) = .38$, $p > .05$, $\eta^2_p = .01$ (see Table 2).

Table 2: Mean Percentage Accuracy Scores, Mean Confidence Ratings, Mean w-s C-A Correlations, and Group b-s C-A Correlations, for Easy, Moderate, and Difficult items, at 5 Minutes and 6 Months

	5 Minute Delay (n=21)			6 Month Delay (n=21)		
	Easy	Mod	Diff	Easy	Mod	Diff
% Accuracy	92.52 (6.57)	74.15 (10.47)	46.19 (19.36)	80.61 (13.39)	65.65 (21.62)	30.00 (22.76)
Conf	7.49 (.98)	5.67 (1.26)	5.49 (1.10)	5.50 (1.32)	3.92 (1.29)	4.09 (1.49)
C-A (w-s)	.40 (.33)	.47 (.22)	-.11 (.33)	.43 (.25)	.40 (.32)	-.19 (.37)
C-A (b-s)	.28	.21	-.12	.50*	.52**	-.38

Note: Standard deviations are shown in parenthesis; * $p < .05$, ** $p < .02$.

Table 3: Within-Subjects Mean Confidence in Correct and Incorrect Answers for Easy, Moderate, and Difficult Items at 5 Minutes and 6 Months

	5 Minutes Delay (n=21)			6 Month Delay (n=21)			TOTAL		
	Easy	Mod	Diff	Easy	Mod	Diff	Easy	Mod	Diff
Conf Corr	7.70 (.95)	6.37 (1.29)	4.70 (1.83)	6.04 (1.31)	4.66 (1.32)	3.41 (2.21)	6.87 (1.41)	5.51 (1.55)	4.06 (2.11)
Conf Inco	2.77 (2.57)	3.57 (1.79)	5.76 (1.29)	3.15 (1.59)	2.38 (1.37)	4.15 (1.75)	2.96 (2.12)	2.97 (1.68)	4.95 (1.73)

Note: Standard deviations are shown in parenthesis.

As an alternative way of construing the within subjects confidence and accuracy data, for each participant, mean confidence in 'incorrect' answers score and a mean confidence in 'correct' answers, was calculated. Mean confidence in correct answers was compared to mean confidence in incorrect answers for the two levels of delay (5 Minutes/6 Months) for the easy, moderate, and difficult items using a 3X2X2 mixed ANOVA (Easy/Moderate/Difficult X 5 Minutes/6 Months X Confidence Incorrect/Confidence Correct), with repeated measures on the last factor.

In addition to the ANOVA results already reported on confidence overall, confidence in correct answers was higher ($M=5.48$, $SD=1.70$) than confidence in incorrect answers ($M=3.63$, $SD=1.84$), $F(1, 40) = 91.88$, $p<.001$, $\eta^2_p = .70$. Also, a significant interaction was observed for item difficulty and confidence, $F(2,80) = 64.12$, $p<.001$, $\eta^2_p = .62$ (see Table 3). Further F analyses ($p<.05$) showed that, whereas confidence in correct answers was significantly higher than in incorrect items for both easy and moderate questions, there was a non-significant trend in the opposite direction for difficult questions.

Also, a significant interaction was observed for item difficulty, confidence and delay, $F(2,80) = 3.72$, $p<.05$, $\eta^2_p = .08$ (see Table 3). To explore this interaction further, three 2X2 (Minutes/6 Months X Confidence Incorrect/Confidence Correct) were conducted on the data, for the easy, moderate and difficult items separately. Only easy items showed a Delay X Confidence interaction, $F(1,20) = 14.64$, $p<.01$; further analyses showed that, whereas confidence in correct answers at 5 minutes was significantly higher than confidence in correct answers at 6 months, there was no difference between confidence in incorrect answers at 5 minutes and 6 months ($p>.05$); the easy condition was the only condition not to show a decrease in confidence in incorrect answers after the delay.

Between subjects C-A relationships were also calculated for each delay period for the three classes of

item difficulty (easy, moderate, difficult; see Table 2). At 5 minutes, no significant C-A correlations were found, and none differed significantly from the others ($p>.05$). However, between subjects C-A correlations for 'easy', and 'moderate' items were both found to be significant at 6 months ($r=.50$ and $r=.52$, respectively), whereas the correlation for 'difficult' items was not. Indeed, again, the latter was negative ($r=-.38$, $p>.05$). Moreover, although the latter correlations for the easy and moderate items did not differ significantly from each other ($z = -0.081$, $p>.05$) both were found to differ significantly from the correlation for the 'difficult' items ($z = -2.85$, $p<.01$ and $z = -2.93$, $p<.02$, respectively).

Finally, out of 798 possible responses given by the 5 minutes interval group, 285 were rated in terms of confidence as "absolutely certain"; of these, 89% were correct. For the 6 months interval group, 103 were rated as "absolutely certain" and, of these, 86% were correct. Overall, 87.5% of "absolutely certain" answers were correct, but there was a 23% drop in the number of such responses after 6 months.

DISCUSSION

In sum, as expected, the overall data showed that a 6 month delay before further questioning significantly reduced both overall accuracy and confidence in responding. However, the effects on C-A relationships were more complex. Within subjects C-A relationships seemed to be relatively unaffected by the delay; i.e. regardless of the delay interval, they tended to be positive for easy and moderate items (and significantly so for the mean confidence in correct and incorrect items analysis), but insignificant with a negative trend for difficult items (participants were more confident in their incorrect responses). Between subjects C-A relationships were also positive for both easy and moderate items, but were only significant after 6 months; whereas between subjects C-A relationships for the difficult items remained negative and statistically insignificant following the 6 month delay. Taken

together these results suggest that, whilst accuracy and overall confidence tend to decline over time, C-A relationships tend to remain positive for easy and moderate items. However, for difficult items, C-A relationships remain insignificant, with a negative trend, regardless of delay. These trends, however, seem to exaggerate over time for between subjects C-A correlations; i.e. the positive relationships improve, whilst the negative relationship for difficult items remains the same, or even worsens slightly. Given the high standard of proof that is required in a criminal prosecution, that being proof beyond a reasonable doubt, the significance of such correlates cannot be underemphasized.

With regard to the between-subject C-A findings, further examination of the data showed more consistent evidence of greater sample variability in scores after the delay. For example, Tables 1 and 2 show larger accuracy and confidence score *SDs* for all levels of item difficulty after the delay. The trend for greater variability is also shown in the range scores; for example, for the overall data the range on accuracy after 5 minutes is 8, after 6 months it is 12. Any increase in heterogeneity would allow higher between subjects correlations (Kebbell, *et al.* 1996). Further examination of the raw data also showed a trend for participants who were generally accurate but not confident after 5 minutes, to become inaccurate and not confident after the 6 month delay, again raising the C-A relationship. One possible explanation for this is that people who 'guess' may be more likely to be correct in their guesses after 5 minutes because of the availability of partial memory traces. However, their guesses may become correspondingly inaccurate after 6 months as the weak traces decay (Weingartner and Parker 1984). In other words, the initially weak correlations for the easy and moderate items were not so much due to participants who were confident in incorrect answers, but those who were *not* confident in their correct answers. However, the converse would be the case for difficult items where the C-A correlation was initially negative. If anything, the effect would be worsened if those who were confident in wrong answers tended to maintain their confidence in these incorrect answers whilst those who were less confident in correct answers tended to forget their correct answers over time.

The present results offer no support for previous findings indicating an increase in within subjects C-A correlations with repeated testing after a delay (for example, Granhag 1997; Ryan and Geiselman 1991).

The most obvious difference between the present study and these others is that a longer time delay was used. Though also, unlike in Granhag's (1997) study, participants were not given an opportunity to formally review their previous responses. A formal review of previous confidence ratings was not used here because of considerations of ecological validity. Of course, there are occasionally situations in the courtroom in which witnesses do have some opportunity to review what they have previously said, such as when a police officer is allowed to consult his or her notebook; however, this is not in any sense equivalent to a detailed formal review of the kind used by Granhag. It is possible that within-subjects C-A might also have improved in the present study if participants had been given an opportunity to review and change their previous judgments; however, the applicability of Granhag's findings might rest crucially on the validity of the witnesses' original C-A assessments. If witnesses' original within subjects C-A judgments were poor or even negatively related, the effects of confirming their confident responses might result even greater calibration inaccuracy. However, the effects of reviewing previous responses on C-A relationships for items of varying difficulty have yet to be investigated.

Obviously, to be generalizable, the present findings need replication with larger samples, and over different stimulus conditions. Nevertheless, as far as they go, the present results suggest that one cannot make the general assumption that any decline in the accuracy of witnesses' testimony over time can, in some way, be compensated for an increase and improvement in their C-A calibration. It appears that the effects depend critically on what is being recalled and how it is measured. The present results support previous evidence that delay generally has a negative effect on accuracy regardless of whether participants undergo repeated testing (Flin, Boon, Knox and Bull 1992; La Rooy, Pipe and Murray 2005; Lipton 1977; Turtle and Yuille 1994), and add to this by showing that it occurs for all levels of question difficulty. Irrespective, whether prosecution offices would have the resources to conduct repeated witness interviews to gauge memory for upcoming trials is questionable. However, the present results also suggest that after a 6 month delay, even with repeated testing, there is no improvement in within-subjects C-A relationships. Moreover, although there may be improvement in between-subjects C-A relationships, this is limited to items classed as easy and moderate. If anything, the calibration is made

worse for difficult items, and there are no obvious grounds for arguing that difficult items will be any less forensically relevant in the courtroom. On the contrary, as mentioned previously, such items are more likely to be targeted by lawyers in the courtroom. It is possible that C-A calibration might be further improved if witnesses were given an opportunity to formally review and change their previous responses; however, even notwithstanding practical limitations, it is not clear whether this might actually be detrimental for items that are difficult to remember. Another point to consider concerns the 23% drop in 'absolutely certain' responses after the 6 month delay in the present study. As such responses tend to coincide quite closely with accurate recall (Kebbell, *et al.* 1996), any drop in their frequency is likely to indicate a decline in the quality of evidence provided.

CONCLUSION

Taken together, for all practical purposes, the present results would support the views of those who have concerns that delays in justice systems can have a significant and deleterious effect on the witness's ability to provide accurate testimony in court. The tendency that effects can be worsened over time if those who were confident in wrong answers remained confident in their incorrect answers later in court means that such evidence could critically influence outcomes.

REFERENCES

- Bothwell, R.K., Deffenbacher, K.A. and Brigham, J.C. 1987. "Correlation of eyewitness accuracy and confidence: Optimality hypothesis revisited." *Journal of Applied Psychology*, 7: 691-695.
<http://dx.doi.org/10.1037/0021-9010.72.4.691>
- Brennan, M. 1995. "The discourse of denial: Cross-examining child victim witnesses. Special issue: Laying down the law: Discourse analysis of legal institutions." *Journal of Pragmatics*, 23: 71-91.
[http://dx.doi.org/10.1016/0378-2166\(94\)00032-A](http://dx.doi.org/10.1016/0378-2166(94)00032-A)
- Brown, D. 2000. "Reducing delays in the Magistrates' Court." Home Office Research Findings, No. 131. London: Home Office Research, Development and Statistics Directorate.
- Deffenbacher, K.A. 1980. "Eyewitness accuracy and confidence." *Law and Human Behaviour*, 4: 243-260.
<http://dx.doi.org/10.1007/BF01040617>
- Ernst and Young Independent Consultants 1999. "Reducing delay in the Criminal Justice System: evaluation of the pilot schemes." Home Office: London.
- Flin, R., Boon, J., Knox, A. and Bull, R. 2011. "The effect of a five-month delay on children's and adults' eyewitness memory." *British Journal of Psychology*, 83(3): 323-336
<http://dx.doi.org/10.1111/j.2044-8295.1992.tb02444.x>
- Fruzzetti, A.E., Toland, K., Teller, S.A. and Loftus, E.F. 1992. "Memory and eye witness testimony." Pp. 18-50 in *Practical Aspects of Memory: Current Research and Issues*, Vol. 1, edited by M.M. Gruneberg and P.E. Morris. London: Routledge.
- Home Office 1999. "Listing arrangements for criminal cases in the Magistrates' Courts: A draft protocol." Located. www.cjsonline.org/legal/trials_issue_group.htm
- Granag, P.A. 1997. "Realism in eyewitness confidence as a function of type of event witnessed and repeated recall." *Journal of Applied Psychology*, 82(4): 599-613.
<http://dx.doi.org/10.1037/0021-9010.82.4.599>
- Harris, S. 1984. "Questions as a mode of control in magistrates' courts." *International Journal of Society and Language*, 49: 5-28.
<http://dx.doi.org/10.1515/ijsl.1984.49>
- Hanna, K., Davies, E., Henderson, E., Crothers and Rotherham, C. 2010. "Child witnesses in the Criminal Courts: A Review of Practice and Implications for Policy." Auckland: The New Zealand Law Foundation and Institute of Public Policy, Auckland University of Technology.
- Hastie, R., Landsman, R. and Loftus, E.F. 1978. "Eyewitness testimony: The dangers of guessing." *Jurimetrics Journal*, 19: 1-8.
<http://dx.doi.org/10.2307/29761641>
- Hickey, L. 1993. "Presupposition under Cross-Examination." *International Journal for the Semiotics of Law*, 1(16): 89-109.
<http://dx.doi.org/10.1007/BF01458741>
- Jonides, J., Lewis, R.L., Nee, D.E., Lustig, C.A., Berman, M.G. and Moore, K.S. 2007. "The mind and brain of short term memory." *Annual Review of Psychology*, 59: 193-224.
<http://dx.doi.org/10.1146/annurev.psych.59.103006.093615>
- Keane, A. and Fortson, R. 2011. "Leading Questions: A critical analysis." *Criminal Law Review*, 4: 280-295.
- Kebbell, M.R., Deprez, S. and Wagstaff, G.F. 2003. "The direct and cross-examination of complainants and defendants in rape trials: A Quantitative Analysis of Question Type." *Psychology, Crime & Law*, 9: 49-59.
<http://dx.doi.org/10.1080/10683160308139>
- Kebbell, M.R., Wagstaff, G.F. and Covey, A.C. 1996. "The influence of item difficulty on the relationship between eyewitness confidence and accuracy." *British Journal of Psychology*, 87: 653-662.
<http://dx.doi.org/10.1111/j.2044-8295.1996.tb02614.x>
- La Rooy, D., Pipe, M.-E. and Murray, J.E. 2005. "Reminiscence and hypermnnesia in children's eyewitness memory." *Journal of Experimental Child Psychology*, 90: 235-254.
<http://dx.doi.org/10.1016/j.jecp.2004.11.002>
- Lindsay, D.S., Read, J.D. and Sharma, K. 1998. "Accuracy and confidence in person identification: The relationship is strong when witnessing conditions vary widely." *Psychological Science*, 9: 215-218.
<http://dx.doi.org/10.1111/1467-9280.00041>
- Lipton, J.P. 1977. "On the psychology of eyewitness testimony." *Journal of Applied Psychology*, 62: 90-95.
<http://dx.doi.org/10.1037/0021-9010.62.1.90>
- Loftus, E.F. 1979. *Eyewitness Testimony*. Cambridge, Mass: Harvard University Press.
- Loftus, E.F. 2003. "Make-believe memories." *American Psychologist*, 58: 867-873.
<http://dx.doi.org/10.1037/0003-066X.58.11.867>
- Manarin, B. 2009. "Bedeviled by delay: Straight talk about memory loss, procedural manipulation, and the myth of swift justice." *Windsor Review of Legal and Social Justice*, 27: 117-144.
- Penrod, S., Loftus, E. and Winkler, J. 1982. "The Reliability of Eyewitness Testimony: A Psychological Perspective." Pp. 119-168 in *The Psychology of the Courtroom*, edited by N.L. Kerr. New York: Academic Press.
- Perfect, T.J., Watson, E.L. and Wagstaff, G.F. 1993. "Accuracy of confidence ratings associated with general knowledge and eyewitness memory." *Journal of Applied Psychology*, 78: 144-147.
<http://dx.doi.org/10.1037/0021-9010.78.1.144>

- Riddle, H. 2012. "The Stop Delaying Justice! Initiative." The Law Gazette (19th January), accessed at <http://www.lawgazette.co.uk/in-practice/practice-points/the-stop-delaying-justice-initiative>
- Ryan, R.H. and Geiselman, R.E. 1991. "Effects of biased information on the relationship between eyewitness confidence and accuracy." *Bulletin of the Psychonomic Society*, 29: 7-9. <http://dx.doi.org/10.3758/BF03334752>
- Samuels, A. 1997. "Custody time limits." *Criminal Law Review*: 260-268.
- Smith, V.L., Kassin, S. and Ellsworth, P.E. 1989. "Eyewitness accuracy and confidence: Within-versus between-subjects correlations." *Journal of Applied Psychology*, 74: 356-359. <http://dx.doi.org/10.1037/0021-9010.74.2.356>
- Sporer, S.L., Penrod, S.D., Read, D. and Culter, B.L. 1995. "Choosing, confidence, and accuracy: a meta-analysis of the confidence-accuracy relation in eyewitness identification studies." *Psychological Bulletin*, 118: 315-327. <http://dx.doi.org/10.1037/0033-2909.118.3.315>
- Stone, M. 1988. "Cross-examination in Criminal Trials." London: Butterworths.
- Taylor, S.C. 2004. *Court licensed abuse: Patriarchal lore and the legal response to intra-familial sexual abuse of children*. New York: Peter Lang Publishing.
- Turtle, J.W. and Yuille, J.C. 1994. "Lost but not forgotten details: Repeated eyewitness recall leads to reminiscence but not hyperamnesia." *Journal of Applied Psychology*, 79: 260-271. <http://dx.doi.org/10.1037/0021-9010.79.2.260>
- Weingartner, H. and Parker, E.S. 1984. *Memory consolidation: Psychobiology of cognition*. NJ: Lawrence Erlbaum.
- Wells, G.L. 1985. "Verbal descriptions of faces from memory: Are they diagnostic of identification accuracy." *Journal of Applied Psychology*, 70: 619-626. <http://dx.doi.org/10.1037/0021-9010.70.4.619>
- Wells, G.L. 1993. "What do we know not about eyewitness identification?" *American Psychologist*, 48(5): 577-580. <http://dx.doi.org/10.1037/0003-066X.48.5.577>
- Wells, G.L. and Murray, M. 1984. "Eyewitness confidence." Pp. 155-170 in *Eyewitness Testimony: Psychological Perspectives*, edited by G.L. Wells and E.F. Loftus. New York: Cambridge University Press.
- Wells, G.L., Lindsay, R.C.L. and Ferguson, T.J. 1979. "Accuracy, confidence and juror perceptions in eyewitness testimony." *Journal of Applied Psychology*, 64: 440-448. <http://dx.doi.org/10.1037//0021-9010.64.4.440>
- Wigmore, J.H. 1940. *A Treatise on the Anglo-American System of Evidence in Trials at Common Law*, Vol 5, 3rd Ed. Boston: Little Brown.
- Wheatcroft, J.M. 2012. *Judicial Processes*. Pp. 229-246 in *Forensic Psychology: Crime, Justice, Law, Intervention*, 2nd Edn, edited by G. Davies and A. Beech, (pp. 229-246). John Wiley & Sons: Chichester.
- Wheatcroft, J.M. and Ellison, L.E. 2012 "Evidence in Court: Witness preparation and cross-examination style effects on adult witness accuracy." *Behavioral Sciences & the Law*, 30: 821-840. <http://dx.doi.org/10.1002/bsl.2031>
- Wheatcroft, J.M. and Woods, S. 2010. "Effectiveness of witness preparation and cross-examination non-directive and directive question styles on witness accuracy and confidence." *International Journal of Evidence & Proof*, 14(3): 189-209. <http://dx.doi.org/10.1350/ijep.2010.14.3.353>
- Wheatcroft, J.M., Caruso, D. and Krumrey-Quinn, J. (in press) *Rethinking Leading: The directive, non-directive divide*. *Criminal Law Review*, 5: 340-346.
- Wheatcroft, J.M., Kebbell, M.R. and Wagstaff, G.F. 2001. "The influence of courtroom questioning style on eyewitness accuracy and confidence." *Forensic Update*, 65: 20-25
- Wheatcroft, J.M., Wagstaff, G.F. and Kebbell, M.R. 2004. "The influence of courtroom questioning style on actual and perceived eyewitness confidence and accuracy." *Legal & Criminological Psychology*, 9: 83-101. <http://dx.doi.org/10.1348/135532504322776870200>

Received on 08-01-2015

Accepted on 17-03-2015

Published on 03-04-2015

© 2015 Wheatcroft *et al.*; Licensee Lifescience Global.

This is an open access article licensed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted, non-commercial use, distribution and reproduction in any medium, provided the work is properly cited.