Face composite production effects on witness memory: Case not closed.

Edward Dumbell

Faculty of Humanities

University of Cape Town

2008

Supervisor: Colin Tredoux

Word counts:

Abstract: 127

Main body: 6806

ABSTRACT

Does creating a composite sketch of a criminal impair a witness's ability to later identify that criminal? Perhaps because of variations in materials and method, evidence from laboratory investigations is highly equivocal, but if the dramatic results of Wells, Charman, and Olson (2005) can be replicated and generalized, this would have implications for forensic practice worldwide. A pilot study was conducted, partially replicating Wells et al., and in addition attempting to explore the possible moderating influences of race (same-race vs. cross-race identifications), decision strategy (automatic vs. intentional) and prospective confidence. Effects of composite construction were compared to those of verbal overshadowing and a control condition. Prospective confidence appeared to moderate the effects of composite construction on discriminability but not the effects of verbalization on shifted decision criteria.

Keywords: accuracy; composite construction; confidence; decision criteria; face recognition; prospective confidence; response bias; verbal overshadowing

BACKGROUND

A not uncommon police procedure for narrowing down suspects is to have a witness build a *composite* likeness of the perpetrator's face. This process entails selecting and combining individual features (mouths, chins, hairstyles, etc.) either from a booklet of transparencies or on a computer, until the witness is satisfied that the result resembles their memory of the criminal's face.

In 1979, Gary Dotson received a sentence of 25 to 50 years for rape and kidnapping after the teenage victim (the only witness) helped created a composite mug-shot using the process described above, and later identified him from a line-up. Ten years later, Dotson became the first person in the world to be exonerated due to DNA evidence (The Innocence Project, n.d.; Wikipedia, 2008).

In the USA alone, the number of wrongful convictions uncovered due to DNA evidence has now reached 216 (The Innocence Project, n.d.) but this is just the tip of the iceberg: DNA evidence is relevant only to the small minority of crimes that involve close personal contact with a victim and that are high-profile enough to warrant re-examination of evidence. (Death sentences are particularly over-represented). If all types of convictions are as prone to miscarriages of justice as we now know death sentences to be, this would mean a total of 29,000 innocents convicted in the USA over a period of 15 years (Gross, Jacoby, Matheson, Montgomery, & Patil, 2005).

Eyewitness testimony appears to be the major problem. It was the "primary evidence" in 36 of the first 40 DNA-exonerated cases (Wells et al., 2000, p. 587), and occurred in 88% of the rape exonerations reviewed by Gross et al. (2005). The US Department of Justice, in recognition of the problem, has started recruiting psychologists to help change its policy and guidelines. Formed in 1998, the Technical Working Group for Eyewitness Evidence included 7 eyewitness researchers in a total of 34 members, and resulted in the booklet *Eyewitness Evidence: A Guide for Law Enforcement* (Technical Working Group for Eyewitness Evidence, 1999). It appears that Psychology is finally playing a direct role in formulating forensic practises, and so more than ever before it behoves eyewitness researchers to replicate and generalise each others' findings.

It is surprising, then, that no published papers have followed up on two experiments by Wells, Charman, and Olson (2005) which demonstrated that building a composite face can drastically impair recognition ability. Participants saw a mock suspect (or *target*) and two days later had to identify this person from a photo line-up. Those who had in the intervening time constructed a composite reproduction were *five times* more likely to pick an "innocent" line-up member instead of the person they had seen before. Given the large sample size (50 participants per condition per experiment), large number of alternate stimuli used, and very large effect, the conclusion that composite production causes memory interference seems incontrovertible. However, previous research (as well as one unpublished attempt at a replication) is equivocal.

Does a Composite Production Effect Make Sense?

Although recognition memory for faces is remarkably resistant to decay over time, source monitoring can be a source of error. Brown, Deffenbacher, and Sturgill (1977) found that participants' ability to identify a person they had encountered previously was impaired if they had they had been shown mug-shots of distractor members of the line-up. Thus, although a face may be well remembered, the circumstances under which it was seen may not be (even to the extent of not knowing whether it was in a photograph or in person).

Could constructing a composite face introduce source monitoring confusion between the original face and the many features and faces-in-progress seen during the task? Hall (1979) found that working with a sketch artist to produce a likeness of a remembered face interfered with later recognition of that face, but in contrast McClure and Shaw (2002) found improved accuracy for participants who had drawn a face themselves. The key difference may be that an artist's realistic sketch is more detailed than the participant's memory, and therefore must include inaccurate detail, whereas when participants drew their own (rough or cartoonish) sketches, they could limit details to what they accurately remembered.

Verbal overshadowing (VOE)

Merely describing a face can impair the ability to later identify it in a line-up, and this effect seems to depend on the amount of detail or guessing (see Meissner & Brigham's meta-analysis, 2001), analogous to the difference between working with a sketch artist and drawing a cartoon. However, there are cases of VOE that defy any explanations based on altered or competing memories, and can only be explained in terms of a global processing shift.

Transfer-Inappropriate Processing Shift

Verbal overshadowing has been found to occur even when a completely different face is described (Dodson, Johnson, & Schooler, 1997). In their meta-analysis, Meissner and Brigham (2005) found an overall negative effects on identification due to verbalisation only for those experiments in which participants were tested within 30 minutes of verbalisation. With longer delays, a *release* occurred, and verbalisation appeared to *enhance* recognition, on average.

Schooler (2002) marshalled a variety of compelling evidence that VOE may be merely one example of a more general phenomenon having nothing specifically to do with verbalisation or memory, and rather reflects a shift to a *featural* processing style which interferes with a range of *holistic* types of sensory expertise (e.g., the recognition of faces, colours, and smells).

Macrae and Lewis (2002, cited in Schooler, 2002) induced such a processing shift with Navon letters. They had participants view "view large letters composed of many small letters, attending either to the large or the small elements" (p. 990). Those focusing at the featural level performed worse than the control group at subsequent face recognition, and those focusing at the gestalt level performed better.

In a similar vein, Dunning and Stern (1994) found that accurate line-up identifications correlated with participants' self-reported use of automatic, non-verbalisable decision strategies ("His face just popped out at me," or "I just recognized him, I cannot explain why"), and misidentifications correlated with intentional (or controlled) decision strategies ("I first eliminated the ones definitely not him").

Is this research relevant to reduced identifications due to composite construction? It seems unlikely that a processing shift played any part in the extreme effects of composite production found by Wells et al. (2005) because recognition only occurred after a 48-hr delay. But using composite production software certainly constitutes a controlled, featural task. To test whether this could affect processing strategies 48 hrs later, the current research included self-reported measures of controlled versus automatic decision strategies (adapted from Dunning and Stern, 1994).

The Composite Production Effect: An Overview of Studies

In their meta-analysis of the verbal overshadowing effect (discussed above), Meissner and Brigham (2001) also analysed the subset of studies which included composite construction tasks, and found across studies a significant facilitating effect: “Participants who generated a facial composite were 1.56 times more like to *correctly identify* the target when compared with a no-description control condition” (p. 612).

However, this meta-analysis predates the experiments of Wells et al. (2005) and excludes the only other study reporting strong interference effects (that of Comish, 1987, because she used Identi-kit face constructions as stimuli, not real photographs). See Appendix A for a condensed summary of findings of all studies accessed for the current review (and one not accessed but included in a Meissner and Brigham’s meta-analysis, 2001). Appendix A also shows potentially relevant differences in materials and designs.

Terminology and common methods

All these experiments comprise three phases. In the encoding phase, participants are presented with a picture or video of a single target person. In the interference phase, participants in the composite condition construct a composite to resemble the target face. In the recognition phase, participants must discriminate between target and *foil* photos (and sometimes indicate their level of confidence). This last phase varies in important ways between studies: Line-ups can be either sequential (also known as *show-ups*—participants must make a decision about each mug-shot before seeing the next one) or simultaneous (six mug-shots are presented at once). Simultaneous line-ups can be either target-absent (simply six foils) or target-present (five foils and the target). Participants are never told in advance whether a simultaneous line-up is target-absent or target-present, but in the latter case, if they make no identification (a miss) they may be asked who they would choose if they *had* to make an identification (a forced choice).

Face Composite Systems

The Identi-kit, used by Comish (1987), Yu and Geiselman (1993), and Mauldin and Laughery (1981) is "a set of transparent celluloid sheets, each containing a line drawing of a facial feature (e.g., many types of noses and eyes)" (Mauldin & Laughery, 1981, p. 353). The eyewitness must work with a trained technician to select the appropriate features and superimpose them to make a face. Photo-fit, used by Davies, Ellis and Shepherd (1978), is a similar system but it uses parts of photographs instead of line drawings. The other researchers used FIS, Mac-a-Mug, and FACES, which are all computer-based systems that eyewitnesses can operate on their own (after minimal instruction).

The difference in level of detail may be an important factor. Wogalter, Laughery and Thompson (n.d.) found a positive effect of composite construction when participants used FIS and a negative trend when they used Mac-a-mug pro. The authors attribute the difference in effect direction to the greater amount of detail in the Mac-a-mug system.

Measurements and testing paradigms

None of the present studies used repeated measures designs, or more than one target face per participant. This makes sense in the light of Fallshore and Schooler's (1995) finding an attenuation of the verbal overshadowing effect after the first trial.

Thus, target-present conditions yield only a two-level variable per participant, resulting in a lack of statistical power. Wogalter et al. (n.d.) and Mauldin and Laughery (1981) partly addressed this problem by giving participants a six-point scale combining three levels of confidence with the target/not-target decision, in effect weighting hits and misses according to confidence. Confounding accuracy and confidence is not necessarily desirable, especially since the two often do not correlate (see, e.g., Comish, 1987, p. 489, but c.f. Dunning & Stern, 1994).

An effect on what?

To evaluate whether participants' recognition performance is good or bad, one must measure their ability to correctly identify both old stimuli (the target) and new stimuli (foils). Thus recognition performance is two-dimensional; deciding what pattern of responses is "better" is straightforward neither for public administrators nor for cognitive psychologists. To illustrate: suppose that, on average, when presented with a line-up including a culprit, witnesses to the crime manage to identify the culprit 50% of the time; and when the culprit is not present they mistakenly identify an innocent look-a-like 20% of the time. Some proportion of those misidentified will be suspects, and a witness's error *might* land an innocent suspect in jail. Suppose that to minimise this injustice, we experiment with shining spotlights on all the members of a line-up. We find that this reduces mistaken identifications from 20% to 10%, which is good, but it also reduces correct identifications from 50% to 30%, which means more criminals on the streets. Public administrators must decide whether this is desirable or not. Cognitive psychologists have a different task: determining whether bright light, over and above making witnesses more cautious, has affected their memories or their ability discriminate. Two rough measures of *discriminability* are 1) hit rate minus false alarm rate and 2) hit rate divided by false alarm rate. By the first measure, bright light has decreased overall accuracy (50 - 30 > 20 - 10) but using the second measure suggests the opposite conclusion (50 / 30 < 20 / 10). Signal detection theory provides formulas to convert proportions of hits and false alarms into orthogonal measures of discriminability and decision criterion. Unfortunately, it is by no means clear that human recognition conforms to idealized receiver operating curves (see e.g., Kroll, Yonelinas, Dobbins & Frederick, 2002).

Thus, a major caveat to Meissner and Brigham's (2001) conclusion of a facilitative-on-average composite construction effect is that they analysed only hit-rates. As can be seen in Appendix A, the two experiments showing a clear positive effect of composite construction on target identification (Mauldin and Laughery, 1981; Wogalter et al, n.d., Exp 1) seem to suffer from floor effects in the target-absent condition (near-zero percentages of false alarms). Thus we cannot discount the possibility that increased identifications were due to participants relaxing their decision criterion.

Clare and Lewandowsky (2004) demonstrated a case of VOE that was indeed due to shifted decision criteria: The VOE was present when participants could choose "target not present", but in a second experiment employing a forced-choice decision, the VOE disappeared.

The forced-choice decision introduced by Clare and Lewandowsky simply and elegantly measures discriminability unconfounded by decision bias. Wells et al. improved on this method by first asking participants to make an unforced decision (a useful measure both because it is ecologically valid and because it allows the possibility of analysing decision criteria) and then asking those participants who (erroneously) chose the “not present” option to make a forced identification. The present experiment employs the same procedure.

Investigating the discrepancy

The purpose of this paper is to address the discrepancy between the findings of Wells et al. (2005) and the attempted replication by Maskow, Schmidt, Tredoux, and Nunez (n.d.). The similarities between these studies rule out many possible moderating variables: Both studies ruled out shifts in decision criteria by including a forced identification decision. Both included a 48-hr delay between composite construction and recognition, probably eliminating transfer-inappropriate processing shifts as an explanation. The level of detail entailed in the interference task was also matched: both sets of participants spent a similar amount of time using the same software program (FACES). Both studies used a different stimulus in line-ups than the stimulus seen at encoding.

One obvious difference between the studies is stimulus encoding time (and, therefore, difficulty of line-up task). Wells et al. used a three-minute exposure and a 30 s video. But Maskow et al., to avoid the ceiling effects they encountered in Exp 1, had to reduce their exposure time to 16 s. This suggests that their foils were easier to discriminate from the target than were those of Wells et al.

If more distinct faces can be distinguished by verbalisable features but highly similar faces can only be distinguished on the basis of holistic or visual memories, then a task which disrupts visual memories will affect performance only on more homogenous line-ups. However, Itoh (2005) found that the similarity of line-up faces moderated VOE in the opposite direction: Verbalisation facilitated recognition when the task was harder and disrupted recognition when target faces were more easily distinguishable. Therefore the uniquely long stimulus presentations times used by Wells et al. may be relevant to their results, but it is unclear how or way.

Another possible moderating variable is the race of participants and targets. Although I have referred to face recognition as a special type of automatic perceptual expertise, most of us are only experts at recognising people of our own race. Fallshore and Schooler (1995) demonstrated a VOE for own-race but not other-race recognition. Maskow et al. (n.d., Exp 2) used White males as stimuli but only 34% of participants were White—it is plausible that any effects of composite construction would have been limited to those 34% of participants. Wells et al. also used White male targets, and do not report the racial demographics of their participants (who were undergraduates at Iowa State University).

METHOD

Design

Two possible target faces and three levels of interference task (control, composite construction, verbal overshadowing) yielded six conditions to which participants could be assigned. Each block of six participants was randomly distributed among these six conditions. Unfortunately, only about 50% of those who signed up actually participated, and this resulted in unequal cell sizes.

Random effects

Exposure time was changed from 3 to 2 s after the first fourteen participants, resulting in a random effect confounded with day of participation.

Participants reporting their racial identity as “White” were scored as same-race and all others were scored as other-race.

Prospective confidence

In order to be able to partial out variance in regression models, any predictor of accuracy independent of the interference tasks was valuable. In addition to exposure and same-race, participants’ confidence that they could identify the target, reported after initial exposure, was expected to covary with accuracy, especially with a small delay introduced between encoding and confidence rating (see Busey, Tunnicliff, Loftus & Loftus, 2000). Pre-confidence was scored on an anchored scale from 1 to 5.

Other dependant variables (in the order in which they were measured)

- Unforced decision: regarding the unforced line-up decision, scored as miss, hit or false alarm. (Note that the target was always present, therefore a correct rejection was not an option.)
- Unforced accuracy: scored as 0 for those who selected "target not present" or identified a foil (miss or false alarm), 1 for a correct identification.
- Post-confidence: confidence in the unforced decision, scored from 1 to 5.
- Forced accuracy: scored as 1 for those who identified the target (either in the unforced condition or after initially choosing "target not present"), and 0 for any identification of a foil.
- Decision strategies: Participants endorsed one or more of seven statements about how they chose a face. Each strategy was intended to reflect either controlled or automatic processing, and each was coded as a binary variable.

Participants

Discounting those who failed to arrive for their first session, 91 undergraduates at the University of Cape Town participated, either as a course requirement or for R50 compensation. Fourteen participants failed to arrive for their second session and one was excluded due to experimenter error, leaving 76 usable cases. Self-reported demographics were: 54 female, 22 male, 31 White, 29 Black, 9 Coloured, 5 Indian, 2 mixed/other. Ages ranged from 18 to 25. Advertisements specified that participants must have normal or corrected-to-normal vision, and gave an indication of the tasks involved without mentioning recognition or memory interference. (Nevertheless, some participants announced that they had arrived for "the face recognition experiment.")

Others participated informally in the preparation of stimuli: Five friends and family members of the experimenter provided suspect descriptions, and 68 undergraduates played "guess the suspect" during a lecture on Psychology and Law. No permission was requested to use these results, and so they will not be discussed further.

Materials

Each participant sat at a desktop computer with a 17-inch monitor. Instructions, questions, and stimuli were presented using PowerPoint slideshows paced by the participant. Printed answer sheets included only numbers and letters to circle or space to write, so as to minimize any clues about upcoming questions—specifically, the unforced line-up decision must be unaffected by knowledge of an upcoming forced line-up decision. The on-screen questions and answer sheets are included in Appendix C.

Face stimuli

All stimuli used were high-quality mug-shot type photos of Caucasian women. In the encoding phase, one of two possible targets was shown, both in colour, from the shoulders up, in a three-quarter pose (both eyes visible). In the recognition phase, a line-up spread consisted of six black-and-white, neck-up, full-face views was shown. Target 1 always appeared top-right and target 2 always appeared bottom-middle.

Target photos were shown approximately 20 by 10 cm at encoding time. Each line-up photo was approximately 9 by 5 cm.

All photos were drawn from two sets available for research purposes. To prevent participants from making use of non-face cues, line-up photos were approximately matched for brightness and contrast, and the backgrounds of the encoding-phase photos were artificially whitened.

Composite software

Participants in the Composite condition learned to use the software program FACES: The ultimate composite picture, version 4.0. This program allows a novice user to select features (head shapes, hair, noses, smile-lines, chin dimples, etc.) to form a face. Some features can be resized or repositioned. For a very detailed description, see Wells et al. (2005).

Procedure

First session, encoding phase

The number of participants together in the same room ranged from one to seven (and was not coded as a variable). Each participant was shown to a computer with a PowerPoint presentation already open, and given a consent form and answer sheet. It was not always possible to ensure that no participant could see the screen of any other participant.

From this point, participants followed instructions on their screens. The PowerPoint presentation was self-paced except for the exposure of the target face, which was preceded by an injunction that "you will only see it once—for just a few seconds—so please watch carefully".

All participants wrote a brief description of the target's mood and personality (similar to a task used be Wells et al., 2005, to induce holistic encoding), and then rated their confidence that they could later identify the target "from a line-up of similar-looking women."

Participants in the Control and Verbal conditions were then reminded (by on-screen instructions) of their next appointment, and left

Participants in the Composite group read a description of the FACES software and an instruction to wait a few minutes for the experimenter to begin the tutorial.

First session, composite task

As soon as all other participants had left (leaving from one to three composite group participants still seated) the experimenter switched each computer to the FACES window and began reading out the scripted instructions.

Participants were encouraged to ask for help if confused or needing more time. The tutorial introduced the basic features and controls of the program (including eye-lines, smile-lines and mouth-lines) in the context of creating a likeness of George W. Bush. This took between 8 and 10 min. Then participants were told they had "about ten minutes" in which to do their best to create a likeness of the target. After 12 min, participants were thanked, reminded of their next appointment, and sent home.

Second session

After a delay of between 47 and 50 hrs, all participants gathered in the same room and were each seated at an assigned computer and given an answer sheet (see appendix C). As in the previous session, all participants followed directions from a (mostly) self-paced PowerPoint slideshow beginning with a "welcome back" screen.

Verbal overshadowing task

Participants in the control and composite groups proceeded directly to the recognition phase. Participants in the verbal condition saw slideshows which began with two additional tasks: First they were asked to spend two minutes counting backwards by sevens, starting from 1000. After 2 min the next slide appeared, asking participants to write down 20 details about the face they saw in the previous session. Instructions stressed the importance of writing a detail on each of the 20 lines, even if the participant had to resort to guessing. After that slide, the rest of the slideshow was identical to that seen by the other two groups.

Recognition phase

(For complete wording of questions for analysed dependent variables, see Appendix C). Participants saw a line-up of six faces including the target they had seen earlier. Instructions were: "If you think one of the faces is the same person you saw 2 days ago, then circle the number of that photo. If you think none of the 6 faces is that person, circle 'N'".

Next, they were asked to circle a number from 1 to 5 ("not at all confident" to "absolutely confident") indicating how confident they were in that decision.

If they had already selected a line-up member, they were instructed to skip the next question. If they had circled "N" then they viewed the same line-up a second time, but for this question they were only given options from 1 to 6, and told to choose their "best guess or the closest match".

Next, participants saw seven statements about how they might have chosen a face, and they circled one or more letters to indicate which strategy (or strategies) they actually used.

The remaining questions asked whether participants were male or female, what their racial identity was (a multiple choice question), and their age.

The final slide thanked participants and, depending on how they were recruited, either assured them that the department would be notified of their participation, or reminded to ask the experimenter for their R50 on their way out.

RESULTS

Descriptive Statistics and Contingency Tests

Discriminability

Figure 1 shows mean accuracy under forced-choice conditions for all independent variables. Unfortunately, as can be seen in Table 1, higher proportions of female and same-race participants were in the long exposure condition, and therefore differences in mean accuracy may be misleading. Because interference task was not confounded with Exposure, and because the total sample size (76) was barely adequate, the 15 long-exposure cases were included in all analyses.

A contingency test of the effect of interference task on forced accuracy (3x2) showed no evidence that forced accuracy differed across groups (Pearson $\chi^2(2) = 1.69$, $p = .43$).

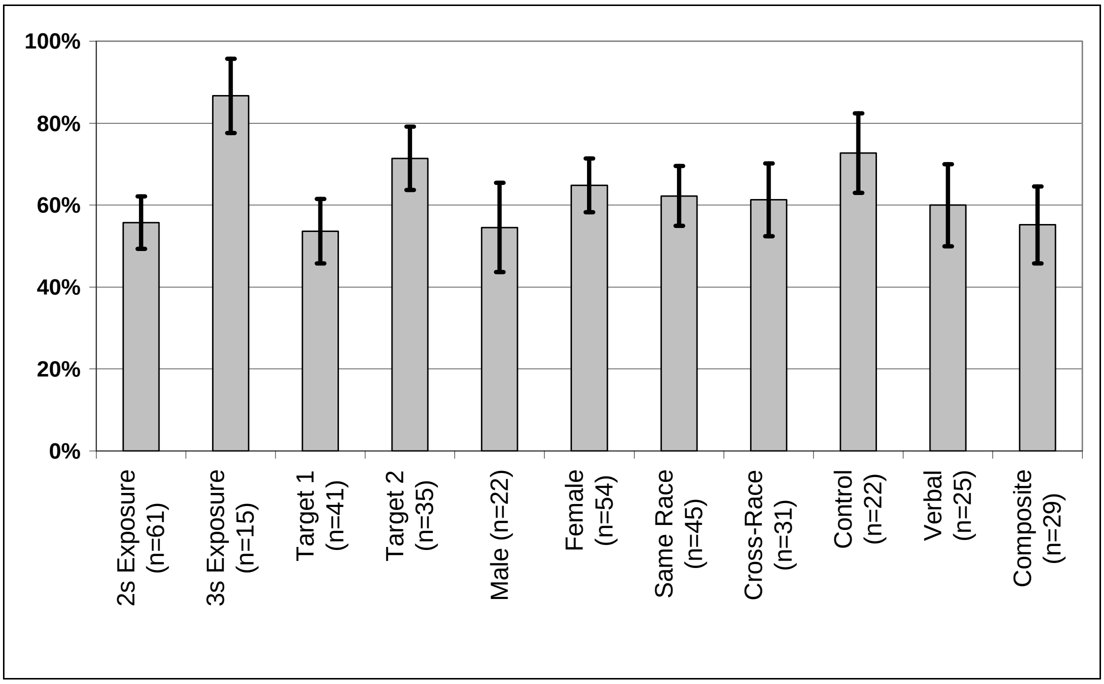

Figure 1. Forced-choice accuracy for all independent variables, +- 1 SE.

Table 1

Intercorrelations for random and unbalanced effects

	Sex	Same-race	Exposure	Target
Sex	--	.06	.24*	.12
Same-race		--	.19	.04
Exposure			--	-.13
Pre-confidence	-.02	-.04	-.01	-.01
Verbal	.01	-.07	.00	-.03
Composite	.08	.01	.02	.04
Unforced Accuracy	.04	-.11	.11	.21
Miss	.01	.10	.00	-.09
Post-confidence	.06	-.19	.00	.02
Forced Accuracy	.10	-.01	.25*	.18

Note: Reported figures are Pearson r_{pb} or φ.
*Significant at $p < .05$
Rows reflect the time-ordering of effects and measurements.

Unforced decisions

Interference task showed a marginal effect on accuracy of unforced decisions (Pearson $\chi^2(4) = 8.98$, $p = .062$). Table 2 shows the percentages of decision outcomes for each condition. The Verbal group made the fewest false alarms and the most misses, indicating stricter decision criteria. For comparison with other studies, see Appendix A.

Table 2

Results of unforced line-up decision by interference task

Interference Task	Hit	Miss	False alarm	Total
Control	64% (14)	14% (3)	23% (5)	100% (22)
Verbal	40% (10)	52% (13)	8% (2)	100% (25)
Composite	45% (13)	31% (9)	24% (7)	100% (29)
Total	49% (37)	32% (25)	18% (14)	100% (76)

Note: Reported figures are: percentage of row (*n*)

Hypothesis Testing: Alpha Levels Not Adjusted

The main hypothesis, that constructing a composite impairs forced-choice accuracy, was not supported. The justification of all further analyses is not to test hypotheses, but to identify possible effects for more rigorous testing in a future experiment (with greater power). Alpha levels are left at .05 and not controlled for multiple comparisons. Non-significant trends are discussed.

Power Maximised by Including Covariates in Logistic Regression Models

Peng and So (2002) advocate building logistic models by first constructing a univariate model for each predictor, and then, based on these results, constructing a preliminary model including all predictors that are "of importance or of interest to the researcher" (p. 54). Interactions are then tested using alternative models.

In order to partial out as much variance as possible (an ANCOVA-like design), I followed a similar procedure (although the purpose was to discover possible effects, not to build a model). I used zero-order correlations to determine "important" predictors: A covariate was included if it correlated at least marginally with the DV ($p < .1$ or $r_{pb} > .2$). (Possible cases when a mediating covariate could mask the effect of an IV are discussed.)

For each dependent variable I constructed one model to test main effects and one model with interactions of interest (if any) added.

Interpreting coefficients

All coefficients given reflect the change in the natural log of the odds ratio. A coefficient of 0.5 reflects an increase in predicted probability from 50% to ~62%, or from 80% to ~87%. A coefficient of -1 reflects a decrease from 50% to ~27%, or from 80% to ~60% (if all other predictors in the model are held constant). Note, however, that coefficients in regression models should always be regarded as a rough estimate, even if they show an excellent fit for the data used to generate the model.

Moderating effect of confidence

Although interaction terms in a logistic regression model were marginal (effect of Composite X Pre-confidence on forced accuracy: L.R. $\chi^2 = 2.6$, $p = .11$), Table 3 shows clear significant correlations under some conditions and no correlation under others.

Because pre-confidence (which occurs first) seems to moderate the effects of the interference tasks, tests of effects that follow are computed not only for the whole sample but also for the high-confidence group only (pre-confidence > 3; $n = 52$ = 68% of cases). Examining this group separately revealed significant effects which would otherwise have been missed.

Table 3
Confidence-accuracy correlations by interference task

Correlation	Control group n=22	Verbal group n=25	Composite group n=29
Pre-confidence with Forced accuracy	.48 p=.023	.23	.03
Post-confidence with Forced accuracy	.18	-.48 p=.015	.04
Pre-confidence with Misses	-.31 p=.157	-.06	-.32 p=.093
Post-confidence with Misses	.39 p=.070	.6 p=.001	.09

Note: Reported figures are Pearson r_{pb}. P values for negligible correlations are omitted. Misses and forced accuracy are not independent ($r_{pb} = -.31$, $p = .006$). Counter-intuitively, control and verbal participants' advance predictions of their accuracy were significantly more accurate than their retrospective ratings.

Effects on forced accuracy (discriminability)

In real-life forensic situations witnesses always make an unforced decision, but in the lab forced-choiced decisions are useful because it yields a measure of accuracy unaffected by decision criteria. A significant effect on this variable would constitute the most compelling evidence for changes in forensic procedure.

Table 4 shows the results of logistic regressions on forced accuracy. When limited to cases with pre-confidence > 3, the effect of composite construction was significant and the effect of verbal overshadowing showed a stronger trend. (However, as mentioned above, a contingency test and the effect of Composite X Pre-confidence suggest a less certain conclusion.)

Table 4
Effects on forced accuracy

	All cases n = 76				High-confidence cases (Pre-confidence = 4 or 5) n = 52			
Effect	B	SE B	χ^2	p	B	SE B	χ^2	p
				Simple models				
Verbal				> .3	-0.77	(0.49)	2.50	.114
Composite	-0.43	(0.32)	1.73	.188	-1.01	(0.47)	4.55	.033
Pre-confidence	0.63	(0.33)	3.55	.060	n.a.	n.a.	n.a.	n.a.
Post-confidence	-0.46	(0.33)	2.02	.156				> .8
Same-race				> .6	-0.64	(0.36)	3.19	.074
				Model with interaction terms added				
Verbal X Same-race				> .3				> .7
Composite X Same-race				> .6				> .4

Note: Negligable coefficients are omitted. Covariants included in models were: exposure, verbal, composite, Same-race, and pre-confidence.

Unforced accuracy not analysed

This is the variable most directly applicable to a “real life” line-up decision, and for purposes of comparison with other studies has been reported in Appendix A. However, I suggest that this variable is of little theoretical interest because it reflects both discriminability and decision bias, in unknown proportions. Discriminability is directly measured by forced accuracy. Shifts in decision criteria can be estimated by analysing the types of errors that are made in the unforced condition (misses or false alarms).

Misses as a measure of decision bias

The number of misses as a proportion of all errors in the unforced decision may be an excellent reflection of average decision bias. Unfortunately, it cannot be used as a DV in a regression model because it cannot be calculated for individual cases.

Instead, effects on decision bias were tested by using percentage of misses as the dependent variable in a logistic regression model and including forced accuracy as a covariate, with the intent of partialling out the effects of discriminability. Coefficients should estimate the degree to which predictor variables biased towards choosing “target not present” (a miss) over picking a line-up member. Table 5 shows that the verbalisation task caused a sizeable significant bias towards strict decision criteria, and composite production showed a smaller trend in the same direction. These effects do not appear to be moderated by pre-confidence.

Table 5
Effects on percentage misses, with forced accuracy partialled out

	All cases *n* = 76				High-confidence cases (PreConf = 4 or 5) *n* = 52			
Effect	B	SE B	L.R. χ^2	*p*	B	SE B	L.R. χ^2	*p*
				Simple models				
Verbal	0.97	0.38	6.55	0.010	1.53	0.63	5.91	.015
Composite	0.51	0.38	1.84	0.175	0.50	0.64	0.61	.435
Pre-confidence	-0.45	0.32	1.95	0.163	n.a.	n.a.	n.a.	n.a.
Same-race	0.31	0.27	1.31	.252	0.79	0.40	3.86	.050
				Model with interaction terms added				
Verbal X Same-race				> .3	--			
Composite X Same-race				> .6	--			

Note: Negligible coefficients are omitted. Covariants included in models were: exposure, verbal, composite, same-race, pre-confidence. “--“ indicates that no regression could be completed, because of empty cells.

Controlled and Automatic Decision Strategies

Participants were given four statements intended to reflect automatic processing and three intended to reflect controlled processing, and asked which applied to their decision (for wording, see appendix C). The four auto statements were endorsed by only 5, 5, 13, and 11 participants, respectively. In contrast, three controlled statements were endorsed by

44, 24, and 52 participants. Taken at face value this suggests that most participants were using controlled processing.

To test whether participants' endorsements of decision strategies *might* reflect two dissociable processes, a principle component analysis was conducted. Figure 1 shows disappointing results: If "controlled" and "automatic" items measure anything like the intended constructs, then all variables in each group should load not just on the same factor but with the same sign. Controlled2 clearly did not measure the same construct as conrolled1 and controlled3. No further analysis of this data was conducted.

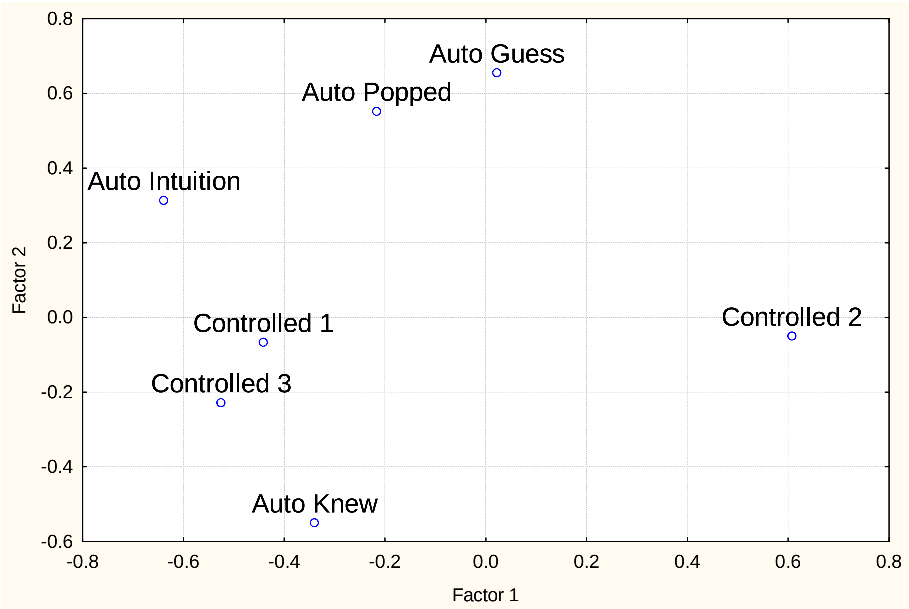

Figure 1. Factor loadings for decision strategy questions (unrotated).

Summary

Confidence-accuracy correlation

Pre-confidence (but not post-confidence) was a significant predictor of later accuracy, for the control group only. This difference in confidence-accuracy relationships across the levels of interference task strongly suggested that pre-confidence moderated the effects of the tasks. Consistent with this possibility, logistic regression models including only high-confidence cases revealed significant effects and stronger trends (see below).

Discriminability

A straight contingency test revealed no overall effect of interference task on discriminability (as measured by forced accuracy). But a logistic regression revealed some effects apparently moderated by Pre-confidence. In the high-confidence group only, constructing a composite impaired discriminability and the verbal overshadowing task showed a similar but non-significant trend. Also limited to the high-confidence group was a trend in contrast to previous research: cross-race participants were more accurate than same-race participants. No interaction between same-race and interference task was found.

Decision bias

A straight contingency test showed a near-significant relationship between interference task and unforced decision, probably reflecting more misses and fewer false alarms in the Verbal group (see Table 2). Trusting logistic regression to partial out the effects of discriminability confirmed that this trend was due to a significantly stricter decision criterion. Participants in the Composite condition showed a weaker trend in the same direction.

DISCUSSION

The Relationship Between Confidence, Accuracy, and Interference Task

Maskow et al. (n.d.) found a positive correlation between accuracy and retrospective confidence for composite constructors only, and not for the control group. Their interpretation was that the composite task improved participants' meta-memory. The current research found a different relationship: a clear negative confidence-accuracy correlation after the verbal overshadowing task. It is not clear how to disentangle the relationship between these variables.

Measurement of prospective confidence was included in the design with the intent of reducing error variance. As hoped, prospective confidence proved to be a predictor of forced-choice accuracy (i.e. discriminability). However this relationship was only significant for the control group, and disappeared for the composite group. Because Pre-confidence and interference task are independent of each other, a causal relationship can be inferred: the interaction of these variables affected discriminability. This is more clearly demonstrated by the regression analysis of the high-Pre-confidence group, among whom composite construction significantly interfered with identification (verbal overshadowing showed a similar but marginal trend).

If such a three-way interaction between prospective confidence, interference task and discriminability is borne out in future research, it may provide an explanation for discrepancies between studies. Wells et al.'s (2005) long exposure times presumably allowed participants to feel highly confident that they could recognise a face, leading to stronger negative effects of composite construction than those found in the present study.

Decision Criteria

In contrast to the effects on discriminability, the verbalisation task had a stronger effect on decision criteria than did composite construction, and these effects did not appear to be moderated by prospective confidence.

Limitations of the current research, and future directions

In my attempt to replicate Wells et al.'s (2005) experiments, I had hoped to make four improvements on Maskow et al.'s (n.d., Exp 2) attempt at a replication: 1) longer exposure times (combined with more difficult line-ups); 2) an analysis of same-race vs. cross-race participants; 3) the partialling out of error variance by means of predictive covariates (such as pre-confidence); and 4) an investigation into controlled versus automatic decision strategies, to determine if these mediated effects of accuracy. None of these improvements turned out as hoped, but all provide opportunities for further investigation.

Homogeneity of line-ups

Wells et al.'s (2005) exposure times are extremely long compared to any of the other face composite studies except Maskow et al.'s (2005) Experiment 1, which found nothing but ceiling effects (only one participant made an error). Clearly, Wells et al. (2005) used line-ups with extremely similar-looking faces. This unique aspect of their experiments may account for their unique results.

Unfortunately I was unable to find foils sufficiently similar to my targets. Despite the fact that participants in my study had to recognise full-face photos after having seen only three-quarter views at encoding time, they showed extremely good recognition ability (regardless of race). After pilot testing, exposure time had to be further reduced from 3 s to 2 s. While it is of some use to demonstrate that the composite production effect does not generalise to shorter exposure times and easier line-ups, future research should attempt a more exact replication of Wells et al.'s (2005) method.

A note on line-ups

While most aspects of method and materials are reported in great detail, face stimuli used tend to be an exception. While overall accuracy and exposure time can be carefully controlled, faces can be "different" in many ways, and the nature of differences between foils and targets may account for discrepancies in findings. Tredoux (2002) mapped facials features from photographs and computed measures of difference. However this one-dimensional measure of difference may miss important aspects of face difference relevant to interference and decision strategies that. For example: two foils, both equally discriminable from the target face, might be similar to and different from the target in different ways: one might be identical except for a having a much larger nose (and therefore discriminable by a verbalisable memory), and another might differ in a myriad subtle ways (and thus be discriminable only by "holistic" processing).

Until such issues are better understood, I suggest including a sample of stimuli in appendices (see appendix B for one of my target and line-up conditions).

Race

Presumably due to random sampling error, cross-race identifications were more accurate than same-race ones. More-over, there was no interaction between SameRace and interference task, as found by Fallshore and Schooler (1995). One drawback of using only White faces as stimuli was that English as a second language (although not measured) was probably confounded with same-race. Verbalisation in English may have different effects if English is not one's home language. Ideally, future research should balance and investigate separately the effects of race, language, and same-race identification.

Covariates

Logistic regression is considered to require a minimum of 50 cases. In this pilot study the sample size was not large enough to split into two for cross-validation, and it is not clear whether the regression models adequately accounted for small and unbalanced cell sizes (multi-collinearity). While use of covariates did appear to confirm some trends suggested by contingency tests and confidence-accuracy correlations, all findings should be treated merely as promising possibilities, until confirmed by a larger study.

Decision strategies

Dunning and Stern (1994) found that eyewitness accuracy correlated with the types of statements participants endorsed regarding their decision strategy. Specifically, statements were assumed to reflect either automatic or intentional processing. However, such a correlation can only be interpreted if the two groups of statements can be shown to manifest two latent factors. The current study found no such pattern.

Conclusion

Does the use of face-composite construction software impair later recognition? The results of this experiment fall somewhere in between the dramatic findings of Wells et al. (2005) and the null findings of Maskow et al. (n.d.). Moreover, a possibly important moderating variable has been discovered: prospective confidence.

A further experiment is indicated with the following modifications: balanced cell sizes, a longer exposure time and more homogenous line-ups, and well validated measures of decision strategy.

REFERENCES

Brown, E., Deffenbacher, K., & Sturgill, W. (1977). Memory for faces and the circumstances of encounter. *Journal of Applied Psychology, 62,* 311-323.

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic bulletin & review, 7,* 26-48.

Clare, J., & Lewandowsky, S. (2004). Verbalizing facial memory: Criterion effects in verbal overshadowing. *Journal of Experimental Psychology, 30*, 739-755.

Comish, S. (1987). Recognition of facial stimuli following an intervening task involving the Identi-Kit. *Journal of Applied Psychology, 72*, 488-491.

Davies, G., Ellis, H., & Shepherd, J. (1978). Face identification: the influence of delay upon accuracy of Photofit construction. *Journal of Police Science and Administration, 6,* 35-42.

Dodson, C. S., Johnson, M. K., & Schooler, J. W. (1997). The verbal overshadowing effect: Why descriptions impair face recognition. *Memory and Cognition, 25*, 129-139.

Dunning, D., & Stern, L. B. (1994). Distinguishing accurate from inaccurate eyewitness identifications via inquiries about decision precesses. *Journal of Personality and Social Psychology, 67,* 118-135.

Fallshore, M., & Schooler, J. W. (1995). Verbal vulnerability of perceptual expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 1608-1623.

Gross, R., Jacoby, K., Matheson, D. J., Montgomery, N., & Patil, S. (2005). Exonerations in the United States, 1989 through 2003. *Journal of Criminal Law and Criminology, 95,* 523-560.

Hall, D.F. (n.d.). *Obtaining eyewitness identifications in criminal investigations: applications of social and experimental psychology (Abstract).*

Innocence Project (n.d.). Facts on Post-Conviction DNA Exonerations. Retrieved 17/05/2008 from http://www.innocentproject.org/Content/351.php.

Itoh, Y. (2005). The Facilitating Effect of Verbalization on the Recognition Memory of Incidentally Learned Faces. *Applied Cognitive Psychology, 19*, 421-433.

Maskow, N., Schmidt, H., Tredoux, C., & Nunez, D. (n.d.). *Face composite production does not affect eyewitness identification accuracy.* Unpublished manuscript, University of Cape Town.

Mauldin, M. A., & Laughery, K. R. (1981). Composite production effects on subsequent facial recognition. *Journal of Applied Psychology, 66*, 351-357.

Meissner, C. A., & Brigham, J. C. (2001). A meta-analysis of the verbal overshadowing effect in face identification. *Applied Cognitive Psychology, 15*, 603-616.

McClure,. A. & Shaw, J. S. III (2002). Participants' free-hand drawings of a target face can influence recognition accuracy and the confidence-accuracy correlation. *Applied Cognitive Psychology, 16*, 387-405.

Peng, C. J., & So, T. H. (2002). Logistic regression analysis and reporting: a primer. *Understanding Statistics, 1,* 31-70.

Schooler, J. W. (2002). Verbalization produces a transfer-inappropriate processing shift. *Applied Cognitive Psychology, 16*, 989-997.

Technical Working Group for Eyewitness Evidence (1999). *Eyewitness evidence: A guide for law enforcement* [Booklet]. Washington, DC: United States Department of Justice: Office of Justice Programs.

Tredoux, C. G. (2002). A direct measure of facial similarity and its relation to human similarity perceptions. *Journal of Experimental Psychology: Applied, 8*, 180-193.

Wells, G.L., Charman, S.D., & Olson, E.A. (2005). Building face composites can harm lineup identification performance. *Journal of Experimental Psychology: Applied, 11*, 147-156.

Wells, G. L., Malpass, R. S., Lindsay, R. C. L., Fisher, R. P., Turtle, J.W. & Fulero, S. M. (2000). From the lab to the police station: A successful application of eyewitness research. *American Psychologist, 55*, 581-598.

Gary Dotson. (2008, June 21). In *Wikipedia, The Free Encyclopedia*. Retrieved on October 30, 2008, from http://en.wikipedia.org/w/index.php?title=Gary_Dotson&oldid=220849044

Wogalter, M. S., Laughery, K. R., & Thompson, B. G. (1986). *Eyewitness identification: Composite construction on subsequent recognition performance.* Unpublished manuscript, University of Richmond.

Yu, C. J., & Geiselman, R. E. (1993). Effects of Constructing Identi-Kit Composites on Photspread Identification Performance. *Criminal Justice and Behaviour, 20,* 280-292.

APPENDIX A
Condensed summary of studies including a composite construction condition

Study & Condition (some collapsed)	System / Task	Stimulus type	Identical stimulus at test time?	Post-construction delay	Line-up type
Thompson (1978, unpublished)	no details available				
Davies et al. (1978) Exp 2	Verbal + Photo-fit	photo	yes	immediate	seq
Davies et al. (1978) Exp 2		photo	yes	3 weeks	seq
Mauldin and Laughery (1981)	Identi-kit	photo	yes	immed. & 48 hrs	seq
Wogalter et al (n.d.) Exp 1	FIS	live human	no	immediate	seq
Wogalter et al (n.d.) Exp 2	Mac-a-mug	photo	yes	immediate	seq
Comish (1987)	Identi-kit	Identi-kit	yes	~13 min	simul
Yu and Geiselman (1993)	Identi-kit	video	no	2 days	simul
Wells et al (2005) Exp 1	FACES	photo	yes?	48 hrs	simul
Wells et al (2005) Exp 2, TP	FACES	video	no	48 hrs	simul
Wells et al (2005) Exp 2, TA	FACES	video	no	48 hrs	simul
Maskow et al (n.d.) Exp 2, TP	FACES	photo	no	48 hrs	simul
Maskow et al (n.d.) Exp 2, TA	FACES	photo	no	48 hrs	simul
Current study	FACES	photo	no	47-50 hrs	simul

Study & Condition (some collapsed)	*n*	Hit Rate		False Alarm Rate (lower=better)		Miss Rate (lower=better)		Fisher's Z_r
		Interference task						
		control	comp.	control	comp.	control	comp.	
Thompson (1978, unpublished)	50	--	--	--	--	--	--	0.26
Davies et al. (1978) Exp 2	20	90%	80%	--	--	n.a.	n.a.	-0.14
Davies et al. (1978) Exp 2	20	60%	40%	--	--	n.a.	n.a.	-0.2
Mauldin and Laughery (1981)	80	~ 60%	~ 90%	~ 4%	~ 4%	n.a.	n.a.	0.49
Wogalter et al (n.d.) Exp 1	78	60%	89%	6%	3%	n.a.	n.a.	--
Wogalter et al (n.d.) Exp 2	90	97%	84%	13%	13%	n.a.	n.a.	--
Comish (1987)	72	44%	22%	47%	50%	8%	28%	--
Yu and Geiselman (1993), TP	47	37%	40%	30%	10%	33%	50%	0.03
Yu and Geiselman (1993), TA				31%	50%			--
Wells et al (2005) Exp 1	~100	84%	10%	6%	30%	10%	58%	--
forced choice:		94%	30%	6%	70%	n.a.	n.a.	--
Wells et al (2005) Exp 2, TP	100	60%	18%	4%	20%	36%	62%	--
forced choice:		88%	32%	12%	68%	n.a.	n.a.	--
Wells et al (2005) Exp 2, TA	100	n.a.	n.a.	20%	26%	n.a.	n.a.	--
Maskow et al (n.d.) Exp 2, TP	90	87%	80%	3%	3%	10%	17%	--
Maskow et al (n.d.) Exp 2, TA	90	n.a.	n.a.	43%	32%	n.a.	n.a.	--
Current study	51	64%	45%	22%	24%	14%	31%	--
forced choice:		73%	55%	27%	45%	n.a.	n.a.	--

Note: Studies analysed by Meissner and Brigham (2001) are indicated by their Fisher's Zr.
"--" = data not available. "n.a." = data not applicable. "seq" = sequential. "simul" = simultaneous.

APPENDIX B
Photos used at encoding time (in colour) and at recognition time (black and white) for target 2

APPENDIX C
Questions for collection of dependent variables (from a PowerPoint slideshow)

Phase 1

Pre-confidence:

On a scale from 1 to 5, How confident are you that you would be able to pick out the person you just saw from a line-up of similar-looking women?

Where 1 indicates "not at all confident" and 5 indicates "absolutely confident", please circle **one** number to indicate your confidence.

Phase 2

Unforced decision:

You're about to see some photos, numbered 1 to 6. The person you saw 2 days ago might be one of the six, or all six might be look-a-likes.

Next to **"Question 2"**, please circle **one** of the options

- If you think one of the faces is the same person you saw 2 days ago, then circle the number of that photo

- If you think none of the 6 faces is that person, circle **"N"**

Post-confidence:

O.K. you should have circled "N" if you thought the original person was not one of the 6 you just saw. Or you should have circled the number of the person you recognized.

How confident are you in that decision? Next to "Question 3", please circle a number from 1 to 5, with 1 being "not at all confident", 3 being "somewhat confident" and 5 being "absolutely confident".

Forced decision:

…if you circled "N" for question 2, then look at the 6 faces again, in the next slide, and this time you **must** circle a number - either your best guess or the closest match.

Decision strategy:

What strategy (or strategies) did you use to choose a face? **You may circle more than one letter.**

(A)	I can’t explain my decision - I just knew	[Auto Knew]
(B)	I can’t explain my decision - I guessed	[Auto Guess]
(C)	I first eliminated the photos that definitely weren’t her	[Controlled 1]
(D)	I compared photos to find the closest match	[Controlled 2]
(E)	I relied on my intuition	[Auto Intuition]
(F)	I compared my memory to specific details in the photos	[Controlled 3]
(G)	One of the faces just “popped out” at me	[Auto Popped[

SameRace:

Please circle **one** letter to indicate your racial identity:

(B)	Black	[0]
(C)	Coloured	[0]
(W)	White	[1]
(I)	Indian	[0]
(O)	Any other race, or mixed race	[0]

Answer sheet, Phase 2

Question 1: [Verbal condition only]

a ______________________________
b ______________________________
c ______________________________
d ______________________________
e ______________________________
f ______________________________
g ______________________________
h ______________________________
i ______________________________
j ______________________________
k ______________________________
l ______________________________
m ______________________________
n ______________________________
o ______________________________
p ______________________________
q ______________________________
r ______________________________

Question 2: **(N) (1) (2) (3) (4) (5) (6)**

Question 3: Not at all---------Somewhat--------Absolutely
(1) (2) (3) (4) (5)

Question 4: **(1) (2) (3) (4) (5) (6)**

Question 5: **(A) (B) (C) (D) (E) (F) (G)**

Question 6: **(M) (F)**

Question 7: **(B) (C) (W) (I) (O)**

Question 8: ______

PLAGIARISM DECLARATION

1. I know that **plagiarism** is wrong. **Plagiarism** is using another's work and to pretend that it is ones own.

2. I have used the American Psychological Association (APA) as the convention for citation and referencing. Each significant contribution to, and quotation in, this essay/report/project/… from the work, or works of other people has been attributed and has cited and referenced.

3. This essay/report/project… is my own work.

4. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.

5. I acknowledge that copying someone else's assignment or essay, or part of it, is wrong, and declare that this is my own work

SIGNATURE: ___________________________

DATE: _________________

(this document copied from http://64.233.183.104/search?q=cache:x7c-GbPEyawJ:web.uct.ac.za/depts/psychology/resources/plag.doc+uct+plagiarism+declaration&hl=en&ct=clnk&cd=1&gl=za&client=firefox-a)