Evaluation of facial composite evidence depends on the presence of other case factors

Dawn McQuiston-Surrett¹*, Amy Bradfield Douglass² and Samantha G. Burkhardt¹
¹Arizona State University, Glendale, Arizona, USA
²Bates College, Lewiston, Maine, USA

Purpose. Previous research on the reliability of eyewitnesses’ facial composites suggests that some scepticism is warranted when evaluating their accuracy. The current research examines the degree to which mock jurors evaluate facial composite images as a source of incriminating or exonerating evidence against a defendant, particularly when they are inconsistent with other evidence presented.

Method. Mock jurors read a fictitious criminal trial transcript of a person charged with armed robbery. Experiment 1 manipulated the facial composite evidence such that participants viewed a facial composite which was either a good or a poor match to the defendant. Experiment 2 also manipulated the quality of the facial composite evidence, as well as the other case evidence against the defendant. After reading the transcript, participants made several testimony-relevant judgments about the case.

Results. Mock jurors reacted more favourably towards the defendant when the facial composite resembled him than when it did not, but reacted similarly when the composite did not resemble the defendant vs. no composite presented. In addition, assessments of the strength of the composite evidence varied depending on the other case evidence presented.

Conclusion. Mock jurors selectively incorporate evidence into their judgments about the credibility of the eyewitness and the culpability of the defendant. They responded to strong vs. weak facial composite evidence as one would expect: A good match composite was viewed as more favourable to the prosecution than a poor match composite. Weak facial composite evidence was largely ignored, however, participants’ assessment of the strength of this evidence was adjusted to be consistent with the other evidence presented.

An extensive research literature in eyewitness identification has shown that jurors tend to overbelieve the testimony of eyewitnesses and this testimony can be highly incriminating. Eyewitnesses are often believed in the absence of other evidence.

* Correspondence should be addressed to Dawn McQuiston-Surrett, Arizona State University, 4701 W. Thunderbird Rd., Glendale, AZ 85306, USA (e-mail: dawn.mcquiston@asu.edu).

DOI:10.1348/135532507X214192
(e.g. Loftus, 1974) and their testimony often carries greater weight than other forms of evidence like physical evidence, alibis and character evidence (see Boyce, Beaudry, & Lindsay, 2006, for a review). Jurors place particularly great weight on the level of confidence displayed by witnesses in evaluating their accuracy (e.g. Cutler, Penrod, & Dexter, 1990; Cutler, Penrod, & Stuve, 1988; Lindsay, Wells, & Rumpel, 1981; Wells, Lindsay, & Ferguson, 1979), which persists even when there is evidence that confidence has been artificially inflated (Bradfield & McQuiston, 2004). Unfortunately, we know from the work of the Innocence Project that eyewitnesses can be wrong and unwarranted belief in eyewitnesses' accuracy has contributed to the erroneous conviction of many innocent people (Scheck, Neufeld, & Dwyer, 2001).

One important factor that may have an impact on jurors’ perceptions of eyewitness accuracy is the consistency with which an eyewitness’s statements are given. Although inconsistent statements are not correlated with an eyewitness’s accuracy (Fisher & Cutler, 1995), some studies have found that jurors exposed to inconsistent witness testimony thought the witness was less credible, the defendant was less culpable, and were less likely to vote to convict than when the witness was entirely consistent (Berman & Cutler, 1996; Berman, Narby, & Cutler, 1995; Bradfield & McQuiston, 2004; Wells & Leippe, 1981). Brewer and Burke's (2002) research suggests that witness consistency has an effect only when the witness displays a high level of confidence; that is, witness inconsistency may be viewed negatively only when it leads jurors to perceive the witness as having reduced confidence (see also Brewer, Potter, Fisher, Bond, & Luszcz, 1999; Lindsay, Lim, Marando, & Cully, 1986). Together, these studies show that witness inconsistency is a potentially important variable that can interact with other evidence and impact jurors’ decisions.

The research on perceptions of witness consistency has examined this issue by exposing mock jurors to contradictory statements given by the witness either while on the witness stand or between the witness’s testimony presented on the stand and earlier statements made. What has not been examined is how jurors react to an eyewitness’s testimony when the strength of that testimony is conflicting. In other words, an eyewitness could provide some information that leads jurors to believe strongly in the accuracy of their testimony, but then provide other information which makes the reliability of their account questionable. For instance, during Timothy McVeigh’s trial concerning the Oklahoma City Federal Building bombing, a key eyewitness testified to descriptions of two men she saw shortly before the bombing, but then admitted to having problems with her memory and overall mental state due to the trauma she experienced (Clark, 1997). What will jurors do in such a situation? Will they selectively use some pieces of evidence while disregarding others? There is evidence that jurors interpret new information during the course of a trial or integrate the evidence presented in such a way that supports their pre-verdict opinion about the defendant’s guilt (Carlson & Russo, 2001; Ostrom, Werner, & Saks, 1978). Specifically, people’s selective use of eyewitness information to support their conclusion about the evidence is suggested by a handful of studies (e.g., Leippe, 1985; McAllister & Bregman, 1986, 1989). An unexplored situation in which selective use of eyewitness evidence by jurors might occur is in the evaluation of a facial composite created by the eyewitness. The current research focuses on this aspect of witness’ testimony, as composite evidence can be ruled admissible in US courts if a testifying witnesses’ account is attacked as having been fabricated. The composite evidence can serve as an attempt to rehabilitate an impeached identification (e.g. People v. Maldonado, 2002). This evidence can also be admissible in UK courtrooms (C.D. Frowd, personal communication, 9 April 2007).
Eyewitnesses are often asked by police to create a facial composite of a perpetrator. For this process, witnesses are generally presented with a collection of different facial features from which to choose in order to recreate the face, with many police forces having adopted computerized composite technologies (McQuiston-Surrett, Topp, & Malpass, 2006). Unfortunately, research has demonstrated various problems with these composite systems in their ability to accurately depict the intended individual (Cutler, Stocklein, & Penrod, 1988; Davies, van der Willik, & Morrison, 2000; Frowd et al., 2005; Koehn & Fisher, 1997; Kovera, Penrod, Pappas, & Thill, 1997; McQuiston-Surrett & Topp, 2008; Shepherd & Ellis, 1996). Various factors influence composite quality, including operator expertise (Davies & Little, 1990; Davies, Milne, & Shepherd, 1983; Ellis, Davies, & Shepherd, 1978), equipment effects (Christie & Ellis, 1981; Davies et al., 2000; Ellis et al., 1978; Laugherly & Fowler, 1980), practice constructing composites (Christie & Ellis, 1981; Wogalter & Marwitz, 1991) and witness/situational effects (Davies, 1981, 1983; Davies & Christie, 1982; Ellis & Shepherd, 1992). Other research has demonstrated that the process of building a composite can harm witnesses’ subsequent ability to accurately identify the culprit (e.g. Wells, Charman, & Olson, 2005; also see Mauldin & Laugherly, 1981).

With this evidence in mind, it is clear that some scepticism is warranted when evaluating eyewitnesses’ facial composites. However, research has not examined the degree to which laypeople evaluate a facial composite image as a source of incriminating or exonerating evidence against a defendant. A related question is, how would jurors assess the credibility of the eyewitness and the culpability of the defendant when the facial composite of the perpetrator produced by the eyewitness is inconsistent with the physical appearance of the defendant? Indeed, frequent media coverage of criminal cases has demonstrated that people who are arrested and charged with a crime sometimes do not resemble the facial composite originally constructed by an eyewitness (for an example of such a case see Montaldo, 2005; Villa, 2005), yet the witness’s testimony and identification of the defendant are still relied upon in court. This is particularly relevant considering that the match between the defendant’s appearance and the prior description of the perpetrator given by witnesses is among the criteria put forth by the US Supreme Court for evaluating eyewitness reliability (Neil v. Biggers, 1972).

How will jurors react to facial composite evidence, especially when it conflicts with otherwise strong eyewitness testimony? In other words, will jurors perceive an inconsistency between an eyewitness’s testimony indicating that (a) she had a good view of the culprit and made a positive identification from a line-up in which she was confident vs. (b) she produced a facial composite that does not resemble the defendant, which should in-turn impact their judgments? Consistent with some of the literature on the impact of witness consistency (e.g. Berman & Cutler, 1996; Berman et al., 1995), one possibility is that this inconsistency will result in positive evaluations of the defendant: Evidence of a composite that is not a good match to the defendant should impugn the witness because it implies that the eyewitness might not have gotten a good look at the culprit and thus exculpate the defendant. Alternatively, Brewer and Burke’s (2002) and others’ findings suggest that mock jurors will ignore this inconsistency and believe that weak composite evidence is irrelevant as long as the witness expresses confidence in her account. Given these competing hypotheses, Experiment 1 was designed to explore how facial composite evidence of varying degrees of resemblance to the defendant would affect assessments of the eyewitness and the defendant.
EXPERIMENT 1

Method

Participants
One hundred twenty undergraduate students ($M = 22$ years old, 71% female) attending a large Southwestern university in the United States volunteered to participate. Students were recruited in Psychology classes and received partial course credit in exchange for their participation.

Materials and design

Transcript
Participants read one of three fictitious trial transcripts, each of which was approximately eight pages long. The case involved a man charged with armed robbery and assault. He allegedly assaulted the woman outside of a shopping mall after knocking her down and stealing her purse and other belongings. The victim of the robbery, who got only a brief look at the culprit, was the sole eyewitness to the crime. The defendant was arrested the following day at the shopping mall because he matched the general description of the culprit given by the eyewitness. He was found in possession of $750 which he claimed was from his paycheck; the victim/witness said that the cash in her purse and contents of other items stolen equalled about $950 in value. He was subsequently identified by the witness from a police line-up. The victim/witness’s testimony was countered by alibi testimony from the defendant’s brother who claimed that the defendant was with him the entire day in question. Each transcript contained the trial in its entirety including the judge’s opening instructions, opening statements by both attorneys, direct- and cross-examination of all witnesses (the eyewitness, the police investigator who handled the case and an alibi witness), closing statements from both attorneys and the judge’s closing instructions to the jury.

Manipulation of facial composite evidence
Experiment 1 had one independent variable with three levels. In the first (control) condition, the eyewitness indicated that, following the crime, she was asked to view a line-up and identified the defendant. At trial, she reported being ‘positive’ in the accuracy of her identification. There was no facial composite information in this version of the transcript. In the second and third conditions, the eyewitness testified that, prior to her line-up identification, she was asked to work with a police investigator in creating a physical likeness of the culprit using a computerized composite system. When she testified at trial, she reported being ‘positive’ in the accuracy of her identification. In each of these conditions, embedded in the transcript were a photograph of the defendant and the composite of the culprit created by the eyewitness. In the second condition, the facial composite bore a strong resemblance to the defendant (good match condition), while in the third condition, the facial composite bore only a mild resemblance to the defendant (poor match condition). Experimenters created these different composite resemblance conditions by either representing all of the facial features of the defendant well in the composite (good match condition) or representing only a couple of the defendant’s facial features while widely varying the others (poor match condition). In both of the composite conditions, two defendant photographs
along with their corresponding composites were used to increase the generalizability of the results. Pilot testing \( (N = 41) \) using a 10-point scale (1, no match; 10, perfect match) determined that the similarity ratings of the good match and poor match composites to their corresponding faces were significantly different from each other for both defendant one \( (M = 6.30 \text{ and } 2.80, \text{ respectively}), t(19) = 3.26, p < .01 \) and defendant two \( (M = 7.18 \text{ and } 3.80, \text{ respectively}), t(18) = 3.63, p < .01 \). Ratings between defendant one vs. two did not differ from each other on either their good match composites, \( t(19) = -0.83, p > .05 \), or their poor match composites, \( t(18) = 1.07, p > .05 \). See Figure 1 for all images used.

**Dependent measures questionnaire**

The questionnaire included eight questions (all using 10-point scales), three of which were answered only by participants in the second and third experimental conditions (i.e. those containing facial composite information). The five critical questions answered by all participants included two concerning the eyewitness (‘Do you think Mrs Thomas, the eyewitness, made an accurate identification from the line-up?’ and ‘Do you think Mrs Thomas got a good look at the culprit from the information you received about the robbery?’) and three concerning the defence’s case (‘How strong do you think the defence’s case is?’; ‘Do you think Mr Robinson, the defendant, is guilty?’ and ‘How believable is the defendant’s alibi, Mr Robinson’s brother, in this case?’). The remaining three questions were answered only by participants who received facial composite information in their transcript: ‘How strong do you think the composite evidence in this case is against the defendant?’; ‘In general, do you think a composite that closely resembles a defendant can convince you of his/her guilt?’; and ‘How well do you think the facial composite constructed by Mrs Thomas resembles the defendant?’ The last question was designed to serve as a manipulation check.

![Defendant photographs and corresponding facial composites used in Experiments 1 and 2. The good match and poor match composites were those implemented in Experiment 1, and the poor match and mismatch composites were those implemented in Experiment 2.](image-url)
**Procedure**
Participants completed the experiment in a classroom setting. Upon arriving at the experiment location, participants completed a consent form. After reading the transcript and reviewing the materials, participants completed the Dependent Measures Questionnaire. They were debriefed, thanked and dismissed.

**Results**
Results are presented in the following order: Manipulation check, five dependent measures answered by all participants in the study and two dependent measures answered only by participants in the two composite conditions. See Table 1 for means and standard deviations for all dependent measures.

**Manipulation check**
Since participants in the two experimental conditions were presented with a transcript containing a composite that was either a good or a poor match to the defendant, there should have been a significant difference on the measure assessing participants’ ratings of how well the facial composite constructed by the eyewitness resembles the defendant. Participants in the good match condition \( M = 7.20; SD = 2.15 \) rated the facial composite as bearing a significantly greater resemblance to the defendant when compared with participants in the poor match condition \( M = 4.02; SD = 1.86 \), \( t(78) = 7.06, p < .001 \). In addition, a series of t tests was done to examine whether participants responded differently to the two defendants on any of the dependent measures. The analyses revealed no significant effects, all \( p > .21 \), therefore the following results are collapsed across the two defendants.

**Dependent measures answered by all participants**
For the five dependent measures answered by all participants, a multivariate analysis of variance (MANOVA) was conducted with transcript type as the independent variable and the following dependent variables: Evaluations of the eyewitness’s accuracy, strength of the defence’s case, quality of the eyewitness’s view, defendant’s guilt and believability of the alibi witness. This analysis revealed a significant omnibus effect of transcript type, \( F(10, 226) = 2.42, p < .01 \). To determine the effect of the facial composite variable, a series of univariate one-way analyses of variance (ANOVAs) was conducted with follow-up Tukey’s Honestly Significant Difference post hoc tests.

Results revealed that participants’ ratings of the eyewitness’s accuracy, strength of the defence’s case, quality of the eyewitness’s view and defendant’s guilt differed significantly, \( F(2, 117) > 3.61, p < .05 \). Ratings on all four variables differed between the good and poor match conditions: Participants who viewed a composite that was a good match to the defendant thought the eyewitness’s identification was more accurate, the eyewitness had a better view of the culprit and the defendant was guiltier than those who viewed a poor match composite, \( p < .05, d > .61 \). Ratings on two variables differed between the control and the good match conditions: Compared to the no composite-control condition, participants who viewed a good match composite also thought the eyewitness’s identification was more accurate and that the defence’s case was weaker, \( p < .05, d > .72 \). On these four significant dependent measures, viewing a poor match composite did not differ significantly from viewing no composite, \( p > .05 \).
| Evaluation of facial composite evidence | 285 |

<table>
<thead>
<tr>
<th>Control – no composite (N = 40)</th>
<th>Composite evidence – good match (N = 40)</th>
<th>Composite evidence – poor match (N = 40)</th>
<th>Overall (N = 120)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Belief in eyewitness accuracy</td>
<td>4.40 (2.02)</td>
<td>4.43 (1.68)</td>
<td>4.98 (2.22)</td>
</tr>
<tr>
<td>Strength of composite evidence</td>
<td>6.95 (1.81)</td>
<td>6.68 (1.76)</td>
<td>6.47 (1.92)</td>
</tr>
<tr>
<td>Defendant guilt</td>
<td>3.50 (1.83)</td>
<td>3.05 (1.30)</td>
<td>3.58 (1.80)</td>
</tr>
<tr>
<td>Eyewitness view of culprit</td>
<td>4.28 (2.61)</td>
<td>3.80 (1.74)</td>
<td>4.22 (1.92)</td>
</tr>
<tr>
<td>General composite reliability</td>
<td>5.18 (1.74)</td>
<td>5.60 (1.78)</td>
<td>5.25 (1.81)</td>
</tr>
<tr>
<td>Strength of composite evidence</td>
<td>4.97 (1.89)</td>
<td>4.02 (1.86)</td>
<td>4.62 (2.37)</td>
</tr>
<tr>
<td>Composite resemblance</td>
<td>7.20 (2.15)</td>
<td>3.88 (1.92)</td>
<td>5.62 (2.57)</td>
</tr>
<tr>
<td>Belief in alibi evidence</td>
<td>5.93 (2.41)</td>
<td>4.92 (2.06)</td>
<td>4.90 (2.42)</td>
</tr>
<tr>
<td>Composite resemblance</td>
<td>4.85 (2.29)</td>
<td>4.16 (2.17)</td>
<td>4.89 (2.17)</td>
</tr>
</tbody>
</table>

**Table 1.** Means and standard deviations for dependent measures for Experiment 1.
Last, the omnibus test for the effect of the composite manipulation on participants’ belief in the credibility of the alibi witness indicated no significant difference between-groups, $F(2, 117) = 1.25, p > .05$.

**Questions assessing facial composite evidence**

Participants in the facial composite conditions ($n = 80$) also responded to questions about the strength of the facial composite evidence and how convincing a matching composite is of guilt. Results showed that participants rated the facial composite as stronger when they received a good match composite ($M = 5.93; SD = 2.41$) than a poor match composite ($M = 3.88; SD = 1.92$), $t(78) = 4.20, p < .001; d = .47$. Ratings concerning participants’ general beliefs concerning the effects of facial composite evidence on defendant culpability were not statistically different, $t(77) = -.15$, $p > .05$.

**Discussion**

Participants read one of three fictitious trial transcripts in which facial composite evidence was manipulated. As one would expect, when presented with a facial composite that strongly resembled the defendant, participants responded favourably towards the eyewitness and punitively towards the defendant: They thought the eyewitness was more accurate, the eyewitness got a better look at the culprit, the defendant was more guilty and the composite evidence was stronger, when compared with participants who viewed a composite that was a poor match to the defendant. Consistent with the second possible inconsistency hypothesis described above, evidence that the eyewitness’s facial composite of the culprit bore a poor resemblance to the defendant did not exculpate him; participants’ ratings of eyewitness accuracy, quality of the eyewitness’s view, strength of the defence’s case and defendant guilt were comparable between the poor match condition and the no composite-control condition. This result may support previous findings that a witness’s expression of confidence can override the negative impact of testimonial inconsistencies (e.g. Brewer & Burke, 2002). At the same time, this finding seems counter-intuitive, as one might expect this evidence to impugn the eyewitness’s credibility and result in more favourable evaluations of the defendant when compared with no real exculpatory evidence. Why might a poorly-matching facial composite fail to provide stronger evidence in favour of the defendant’s innocence? We know from the significant differences in ratings of the eyewitness and defendant that participants did not completely disregard the facial composite evidence. However, perhaps the poor match composite was not poor enough to have an impact on participants’ judgments, particularly in the face of fairly strong eyewitness testimony. In other words, to better test the effects of inconsistent testimony, it could be that stronger manipulations were needed.

It was, therefore, reasoned that varying the strength of the case evidence – particularly the witness’s confidence – would influence the degree to which facial composites are considered. The literature on persuasion and jurors’ use of central vs. peripheral cues in the courtroom (Petty & Cacioppo, 1986; Williams & Jones, 2005) supports the assertion that the level of confidence displayed by eyewitnesses may be differentially processed, leading to different interpretations of the other evidence presented. For example, if the majority of the evidence is incriminating including an
extremely confident witness that confidence level might act as a peripheral cue such that jurors are so persuaded by the witness’s account that other evidence is not critically evaluated. In this scenario, we would expect the witness’s confidence to override a weak facial composite such that the composite is ignored. An unconfident eyewitness might lead jurors to instead process information through the central route whereby they feel they must carefully examine the arguments being made and be critical of all of the evidence presented. Here, exculpating evidence that includes an unconfident witness would be expected to better highlight the exculpatory nature of a poor facial composite. To test these possibilities and provide a stronger evidence inconsistency manipulation, a second experiment was conducted to systematically examine the conditions under which participants might incorporate very weak facial composite evidence into their assessments of the case.

EXPERIMENT 2

A circumstance under which weak facial composite evidence may contribute to less favourable ratings of the prosecution is when the other evidence in the case is consistent with innocence. The case that participants read in Experiment 1 focused not only on the eyewitness’s identification of the defendant, but also included an alibi witness who testified in his defence. This evidence was designed to produce a case in which the defendant’s guilt was uncertain, which was accomplished as participants in the control condition rated his guilt near the mid-point of the scale (M = 4.28 on 10-point scale). However, in order to strengthen our test of the effect of inconsistent testimony, new materials were developed in which both the strength of the case against the defendant and the facial composite-defendant resemblance were manipulated. Thus, Experiment 2 examined whether manipulations in the reliability of the evidence would highlight a facial composite which was inconsistent with the physical appearance of the defendant and lead to more favourable ratings of the defendant, compared with a no composite condition.

Three different evidence strength conditions were created. The strong evidence condition focused mostly on the eyewitness’s identification of the defendant and strong confidence statement, and there was little evidence in support of his innocence (i.e. a weak alibi). The weak evidence condition contained a very unconfident eyewitness and included a strong alibi who testified as to the defendant’s whereabouts during the time the crime was committed. The moderate evidence condition was identical to the transcript presented in Experiment 1. Three different facial composite evidence conditions were also created. Since the purpose of Experiment 2 was to explore the conditions under which weak facial composite evidence would be given greater weight by mock jurors, a good match composite condition was not included. In addition, to further strengthen the manipulation of inconsistent testimony, a condition was included in which the facial composite did not resemble the defendant at all. Thus, the facial composite variable consisted of three conditions: A no composite-control condition (identical to Experiment 1), a poor match condition (identical to Experiment 1) and a mismatch condition in which the composite did not resemble the defendant. The manipulation of these variables produced a 3 (evidence strength: Weak vs. moderate vs. strong) × 3 (facial composite: No composite control vs. poor match composite vs. mismatch composite) completely randomized between-groups factorial design.

Several predictions were made based on our assumption that weak composite evidence would be viewed as exculpatory under some conditions. First, an interaction
effect was predicted such that composite evidence would be viewed differently depending on the strength of the other case evidence (e.g. Williams & Jones, 2005). When evidence exculpates the defendant, mock jurors should be sensitive to poor composites because they corroborate other evidence indicating that the defendant is innocent. When the evidence against him is strong, poor composites should be ignored because they contradict the confident eyewitness and weak alibi. No difference between the composite conditions was expected when the strength of the other evidence was moderate, based on our findings from Experiment 1. Our second prediction, more exploratory in nature, was a main effect of the facial composite variable. We expected the mismatch composite to produce more favourable evaluations towards the defence when compared with the poor match composite.

Method

Participants
Two hundred and sixty seven undergraduate students ($M = 23$ years, 71% female) attending a large Southwestern university in the United States volunteered to participate. Students were recruited in Psychology classes and received partial course credit in exchange for their participation.

Materials and design
The transcript used in Experiment 1 was modified to create nine different versions reflecting the nine experimental conditions. The evidence strength was manipulated to include three types: Weak, moderate or strong. The transcript used in the moderate evidence condition was identical to that used in Experiment 1. To create a strong evidence condition, the eyewitness testified as being very confident in her identification of the defendant and her statements were straightforward and did not contain any pauses or qualifiers, the alibi witness’s testimony contained several inconsistencies and the prosecution presented clothing found at the defendant’s home which matched the eyewitness’s description of what the perpetrator was wearing. For the weak evidence condition, the eyewitness’s identification of the defendant was not strong as she wavered between several line-up members and her statements contained many verbal qualifiers (i.e. ‘umm . . . ’) and pauses to indicate uncertainty. The defendant had a strong alibi (a friend who gave reliable statements when questioned about having been with the defendant during the time the crime was committed) as well as evidence of a pizza delivery receipt that was time stamped around the alleged time of the crime containing what was alleged to be the defendant’s signature (Olson & Wells, 2004; Skolnick & Shaw, 2001). A convenience store surveillance videotape was also introduced that showed the defendant’s alibi and a man resembling the defendant visiting a convenience store right around the time the crime was said to have been committed.

The facial composite evidence was also varied to include either no facial composite, a composite that was a poor match to the defendant, or a composite that bore no resemblance and was a mismatch to the defendant. The poor match composites were those used in Experiment 1. The mismatch composites were created by experimenters to bear no resemblance on any facial features to the corresponding defendant photographs. For the two facial composite conditions, arguments made by the attorneys in the transcripts were modified slightly to reflect this evidence, as well as the case evidence manipulation detailed above.
The Dependent Measures Questionnaire implemented in Experiment 2 was identical to that from Experiment 1.

Pilot testing
Participants (N = 20) read one of the three trial transcripts (either weak, moderate, or strong evidence) in which the composite manipulation was removed to isolate the case evidence variable, then rated the defendant’s guilt on a 10-point scale (1, not at all guilty; 10, extremely guilty). Guilt ratings of the defendant differed in the expected directions between all three evidence groups (the differences were marginally significant, t(11) > 1.94, p < .08, for the moderate (M = 4.83, SD = 2.86) vs. weak (M = 2.43, SD = 1.13) and moderate vs. strong (M = 7.86, SD = 2.73) groups; t(12) = 4.85, p < .001 for the weak vs. strong group). A second round of pilot testing (N = 40) was conducted to ensure differences between the poor match composites (from Experiment 1) vs. the mismatch composites. Similarity ratings on a 10-point scale (1, no match; 10, perfect match) between the mismatch composites and their corresponding photos (M = 2.15, SD = .99) were significantly less than that of the poor match composite ratings from Experiment 1 (M = 3.30, SD = 2.10), t(38) = 2.21, p < .05.

Procedure
Participants completed the experiment in a classroom setting. Following informed consent, participants were randomly assigned to one of nine conditions. After reading the transcript, participants completed the Dependent Measures Questionnaire, with the same instructions used in Experiment 1. After completing the questionnaire, participants were debriefed, thanked and dismissed.

Results
Results are presented in the following order: Manipulation check, five dependent measures answered by all participants in the study, and two dependent measures answered only by participants in the two composite conditions. See Table 2 for means and standard deviations for all dependent measures.

Results of manipulation check
Participants in composite conditions viewed a facial composite that was either a poor match to the defendant or mismatch to the defendant, so there should have been a significant difference on the measure assessing participants' ratings of the how well the facial composite resembles the defendant. Participants in the poor match condition rated the facial composites as bearing a significantly greater resemblance to the defendant (M = 4.69; SD = 2.35) compared to participants in the mismatch condition (M = 3.47; SD = 2.32), t(176) = 3.51, p = .001. In addition, participants did not respond differently to the two defendants on any of the dependent measures, p > .23, so results are collapsed across both defendants.

Dependent measures answered by all participants
A two-way MANOVA was conducted to assess the effects of the evidence strength and facial composite variables on responses to the five dependent measures answered by all participants. The main effect of evidence strength was statistically significant,
<table>
<thead>
<tr>
<th>Control – no composite (N = 89)</th>
<th>Weak evidence</th>
<th>Moderate evidence</th>
<th>Strong evidence</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Belief in eyewitness accuracy</td>
<td>3.14 (1.48)</td>
<td>4.07 (1.87)</td>
<td>7.47 (1.79)</td>
<td>4.91 (2.53)</td>
</tr>
<tr>
<td>Strength of defence case</td>
<td>8.03 (1.55)</td>
<td>5.90 (2.38)</td>
<td>2.93 (1.80)</td>
<td>5.60 (2.85)</td>
</tr>
<tr>
<td>Eyewitness view of culprit</td>
<td>2.66 (1.61)</td>
<td>2.83 (1.80)</td>
<td>6.73 (2.36)</td>
<td>4.09 (2.71)</td>
</tr>
<tr>
<td>Defendant guilt</td>
<td>3.38 (2.32)</td>
<td>3.67 (1.94)</td>
<td>8.10 (2.17)</td>
<td>5.07 (3.04)</td>
</tr>
<tr>
<td>Belief in alibi evidence</td>
<td>7.59 (2.01)</td>
<td>5.67 (1.88)</td>
<td>1.57 (0.63)</td>
<td>4.91 (2.99)</td>
</tr>
<tr>
<td>Strength of composite evidence</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Composite resemblance</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>General composite reliability</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Composite evidence – poor match (N = 88)</th>
<th>Weak evidence</th>
<th>Moderate evidence</th>
<th>Strong evidence</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Belief in eyewitness accuracy</td>
<td>3.00 (1.51)</td>
<td>4.40 (2.14)</td>
<td>6.89 (1.78)</td>
<td>4.72 (2.42)</td>
</tr>
<tr>
<td>Strength of defence case</td>
<td>7.43 (1.87)</td>
<td>6.43 (1.79)</td>
<td>3.89 (2.25)</td>
<td>5.97 (2.45)</td>
</tr>
<tr>
<td>Eyewitness view of culprit</td>
<td>2.57 (1.19)</td>
<td>3.73 (2.19)</td>
<td>6.43 (2.13)</td>
<td>4.19 (2.47)</td>
</tr>
<tr>
<td>Defendant guilt</td>
<td>2.97 (1.83)</td>
<td>4.03 (2.43)</td>
<td>7.29 (2.05)</td>
<td>4.70 (2.78)</td>
</tr>
<tr>
<td>Belief in alibi evidence</td>
<td>7.20 (2.23)</td>
<td>5.60 (1.77)</td>
<td>1.57 (0.96)</td>
<td>4.86 (2.93)</td>
</tr>
<tr>
<td>Strength of composite evidence</td>
<td>4.47 (2.56)</td>
<td>4.33 (2.17)</td>
<td>5.32 (2.24)</td>
<td>4.69 (2.35)</td>
</tr>
<tr>
<td>Composite resemblance</td>
<td>3.67 (1.86)</td>
<td>4.40 (1.96)</td>
<td>5.71 (2.31)</td>
<td>4.57 (2.91)</td>
</tr>
<tr>
<td>General composite reliability</td>
<td>5.40 (2.37)</td>
<td>4.63 (2.19)</td>
<td>5.64 (1.95)</td>
<td>5.22 (2.20)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Composite evidence – mismatch (N = 90)</th>
<th>Weak evidence</th>
<th>Moderate evidence</th>
<th>Strong evidence</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Belief in eyewitness accuracy</td>
<td>3.00 (1.64)</td>
<td>4.40 (2.14)</td>
<td>3.67 (1.89)</td>
<td>4.72 (2.42)</td>
</tr>
<tr>
<td>Strength of defence case</td>
<td>7.80 (1.65)</td>
<td>6.43 (1.79)</td>
<td>6.07 (2.43)</td>
<td>5.97 (2.45)</td>
</tr>
<tr>
<td>Eyewitness view of culprit</td>
<td>2.50 (1.25)</td>
<td>3.73 (2.19)</td>
<td>3.00 (1.66)</td>
<td>4.19 (2.47)</td>
</tr>
<tr>
<td>Defendant guilt</td>
<td>2.23 (0.93)</td>
<td>4.03 (2.43)</td>
<td>4.10 (2.22)</td>
<td>4.70 (2.78)</td>
</tr>
<tr>
<td>Belief in alibi evidence</td>
<td>7.67 (1.99)</td>
<td>5.60 (1.77)</td>
<td>5.30 (2.26)</td>
<td>4.86 (2.93)</td>
</tr>
<tr>
<td>Strength of composite evidence</td>
<td>2.63 (1.69)</td>
<td>4.33 (2.17)</td>
<td>3.03 (2.12)</td>
<td>4.69 (2.35)</td>
</tr>
<tr>
<td>Composite resemblance</td>
<td>3.30 (2.00)</td>
<td>4.40 (1.96)</td>
<td>3.20 (1.90)</td>
<td>4.57 (2.91)</td>
</tr>
<tr>
<td>General composite reliability</td>
<td>5.10 (2.17)</td>
<td>4.63 (2.19)</td>
<td>5.13 (2.31)</td>
<td>5.22 (2.20)</td>
</tr>
</tbody>
</table>
Evaluation of facial composite evidence

F(10, 508) = 48.03, p < .001. The main effect of facial composite was not statistically significant, F(10, 508) = .94, p > .05, nor was the interaction, F(20, 843) = 1.22, p > .05.

According to the univariate ANOVAs for the evidence strength variable, participants’ ratings on all five dependent measures (eyewitness accuracy, defence case strength, eyewitness view, defendant guilt and alibi believability) differed significantly, F(2, 258) > 86.18, p < .001. Post hoc analyses indicated that when the evidence was strong, participants believed that the eyewitness was more accurate, the defence’s case was weaker, the eyewitness had a better view, the defendant was guiltier and the alibi was less believable, compared to both the moderate and weak evidence conditions, p < .01, d > 1.08. Differences between the moderate and weak conditions were also significant in the same directions, p < .01, d > .52, except on ratings of the eyewitness’s view of the culprit, p = .11.

Questions assessing facial composite evidence
Participants who viewed composite evidence (n = 178) answered three additional questions (including the manipulation check). A two-way MANOVA was conducted to assess the effects of the evidence strength and facial composite variables on composite resemblance, composite strength, and how convincing a matching composite is of guilt. As expected, the main effect of facial composite was significant, F(3, 170) = 4.49, p < .01. Surprisingly, the main effect of evidence strength was statistically significant also, F(6, 340) = 14.75, p < .001. The interaction was not significant, F(6, 340) = .97, p > .05.

The main effect of the composite variable on the composite resemblance measure was statistically significant (as reflected in the manipulation check above), showing that the poor match composites were found to be a better match to the defendant than the mismatch composites, F(1, 172) = 13.80, p < .001, d = .53. This variable had no significant effect on ratings of the strength of the facial composite evidence, p > .05, indicating that participants did not find a difference in the strength of the poor match composites vs. the mismatch composites despite the difference in resemblance ratings. There was also no significant effect for ratings of how convincing a matching composite is of guilt, p > .05.

There was a significant main effect of evidence strength on both the composite resemblance and the composite strength dependent measures, F(2, 172) > 7.73, p < .01. Post hoc analyses revealed that when the other case evidence against the defendant was strong, participants thought the facial composite bore a better resemblance to the defendant and provided stronger evidence than when the other case evidence was moderate or weak, p < .01, d > .57. Ratings between the moderate and weak evidence conditions were not significantly different for either dependent measure, p > .05. There was no significant main effect for the dependent measure assessing how convincing a composite is of guilt, p > .05.

Discussion
Experiment 2 was designed to provide a stronger test of the inconsistency hypothesis by varying the strength of the evidence against the defendant based on the assumption that this would result in greater weight being given to exculpating facial composite evidence. Our first hypothesis that the strength of the other case evidence against the
defendant would influence the interpretation of weak composite evidence (vs. no such evidence) was not supported by the data: There was no difference in ratings of the eyewitness’s accuracy, eyewitness’s view, defendant’s guilt, strength of the defence’s case, or believability of the alibi under the facial composite conditions vs. the no-composite condition regardless of the strength of the other case evidence. This finding replicates and extends our findings from Experiment 1, suggesting that the inclusion of weak – and even very weak – facial composite evidence for jurors to consider is not different from including no facial composite. It is possible, as described above, that a witness’s identification is so powerful that a non-matching facial composite on record is simply ignored (Boyce et al., 2006). The overexposure of composite images in the media without providing information about how often they actually assist police might also dull their impact. Laypeople probably know very little about the construction of composites and their use by law enforcement - in contrast with the more sensationalized forms of evidence (e.g. DNA and fingerprint analyses) via current television programmes that might be subjected to more critical evaluations by jurors (Roane, 2005).

Another possible explanation for this null finding has to do with people’s perceptions of the difficulty of the facial composite task. If people think that creating a facial composite is difficult, they might be willing to disregard it when it is not consistent with the appearance of the defendant by reasoning that the composite construction process does not truly reflect the reliability of the eyewitness’s account. If people have this view, but the composite clearly resembles the defendant, this might instead be viewed as bolstering the witness’s credibility and, in-turn, the defendant’s culpability. Research on augmenting and discounting in attribution theory would support this contention based on the differential evaluation of evidence as a function of prior expectations about the difficulty people have with the task (Kelley, 1972). To learn more, we conducted a follow-up test in which we asked 80 students to indicate on a 7-point scale (1 = very easy, 7 = very difficult) how difficult they think it is to make a composite of a person they have seen only once for a brief time. Results indicated a mean of 5.53 (SD = 1.23), which differs significantly from the scale mid-point, t(79) = 7.44, p < .001, confirming our conjecture that people think this task is a difficult one. It could be that regardless of the other case evidence, jurors are unwilling to penalize the prosecution if the eyewitness creates a poor-resemblance composite, a conclusion supported by participants’ reported scepticism concerning the general reliability of composites as they relate to culpability (see Tables 1 and 2). Future research should explore specifically where jurors think the difficulty with composite construction lies, and whether their assessment of task difficulty plays a role in their evaluations of the degree to which composite evidence should be considered in a criminal trial.

Our second prediction that a facial composite bearing no resemblance to the defendant would be considered more exculpating than a poor match composite was also not confirmed. Participants’ ratings of the eyewitness’s accuracy, eyewitness’s view, defendant’s guilt, strength of the defence’s case, or believability of the alibi did not differ between these two conditions, despite the results of our manipulation check showing that participants did perceive a difference between the similarity of those two composites and the defendant. Hence, although participants identified the difference between the composites’ resemblance to the defendant, these judgments did not influence their broader opinions about the evidence and the case. Notably, while statistically significant, the actual mean difference between the similarity ratings for the
poor and mismatch composites vs. their target face was small. It is possible that, in comparing the target faces with their corresponding composites, a feature-by-feature analysis resulted in sufficient dissimilar characteristics for participants to categorize both composites as quite dissimilar from the target. More specifically, given that hairstyle can be a key physical marker used in characterizing and classifying faces (MacLin & Malpass, 2001), the dissimilarity between the hairstyle shown in the target faces as compared to both of their corresponding composites may have stood out to participants and, thus, similarly impacted the ratings of both composites.

Although the presence of facial composite evidence did not interact with the other case evidence to impact mock jurors' evaluations of the case, evaluations of the composite evidence itself were influenced by the other case evidence presented. Reading about incriminating evidence led participants to evaluate the composite as bearing a better resemblance to the defendant compared to reading about weak or moderate case evidence even though, overall, the quality of the composites was the same across evidence conditions. A similar result was found when participants were asked to rate the strength of the composite evidence: When the other case evidence was strong, participants thought the composite evidence was stronger than when the evidence was weak or moderate. Interestingly, although our manipulation check indicated that participants did perceive a difference in the resemblance between the poor match composites and the mismatch composites, this difference did not extend to their perception of the strength of that composite evidence between these two conditions. These data are consistent with the research suggesting that the perceived strength of one type of evidence can be influenced by the presence of other evidence (e.g. Bradfield & Wells, 2000; McAllister & Bregman, 1986).

**GENERAL DISCUSSION**

Two experiments examined the impact of facial composite evidence on jurors' judgments in a criminal trial. Experiment 1 demonstrated that mock jurors' judgments about the trial were more favourable to the prosecution when the facial composite matched the defendant vs. a composite that did not match him, but their judgments were comparable when exposed to a poor match composite vs. reading about no composite evidence. Experiment 2 further explored this finding and implemented a stronger test of the inconsistent testimony manipulation in assessing whether variations in the other case evidence against the defendant would facilitate use of the weak facial composite evidence (either poorly matching or mismatched). Again, no difference in evaluations of the eyewitness and defendant was found when participants received weak composite evidence vs. no composite evidence, regardless of the strength of the other case evidence, thus replicating the findings from Experiment 1. We particularly expected the exculpating nature of the weak composite evidence to be salient when it clearly contradicted the other case evidence, but this finding persisted whether the other evidence in the case exculpated the defendant, incriminated him, or was somewhere in between. Importantly, variations in the other evidence led mock jurors to react differently when evaluating the composites themselves: As the strength of the evidence against the defendant increased, so did participants' ratings of both the strength of the composite evidence and how well the composite resembled the defendant.

A significant result of the evidence variable was that mock jurors' evaluations of the composite evidence shifted so as to be consistent with the strength of the other
evidence in the case. Thus, even though the presence of weak composite evidence had no impact on participants’ ratings of the eyewitness’s accuracy and defendant’s culpability, the perceived strength of this evidence increased and composite resemblance ratings increased as the other case evidence became more incriminating.

It would be interesting to assess jurors’ reactions to strong facial composite evidence when varying the strength of the other case evidence, since our research only included weak facial composite evidence in that investigation. Would mock jurors confronted with evidence that exculpated the defendant adjust their views of a good match composite to be consistent with the other evidence and, thus, rate it as bearing only a slight resemblance?

Our finding that composite evidence is discounted under some conditions but factored into decisions under other conditions is consistent with the notion that people are selective about the evidence they consider which will confirm their ‘story’ that explains the majority of the evidence (Pennington & Hastie, 1986, 1992; see also Kuhn, 2001; Nickerson, 1998). According to this model, jurors organize incoming information in a trial and decide which story concerning the crime makes the most sense, at which point evidence that confirms their hypothesis of a particular story will be retained while evidence that contradicts it could be disregarded or modified to be consistent with the story. A handful of studies support this idea of a confirmatory bias in the consideration of evidence. For example, some research indicates that jurors underutilize eyewitness non-identifications when this information disconfirms an earlier hypothesis about defendant culpability (Leippe, 1985; McAllister & Bregman, 1986, 1989; see also Wells & Lindsay, 1980) and others argue that jurors’ consideration of corroborating vs. contradictory eyewitness testimony depends on whether the testimony confirms their beliefs about defendant guilt (Lindsay et al., 1986). Future research should continue to investigate jurors’ selective consideration of evidence, particularly under more ecologically valid conditions. How might jury deliberations impact evaluations of a facial composite that is inconsistent with the defendant’s physical appearance? How do jurors’ differing points of view regarding contradictory evidence converge to arrive at a verdict? How would jurors react to conflicting accounts given by multiple witnesses, such as the presence of several composites created by separate witnesses, or even a morph of several witnesses’ composite faces (e.g. Hasel & Wells, 2007)? The nature of cross-examination could also be examined to determine whether aggressively challenging the witness would sensitize jurors to inconsistencies in credibility (e.g. Bradfield & McQuiston, 2004).

Our results should be qualified by some limitations to our experimental methods. First, participants read a transcript of a fictitious trial rather than listening to an audiotape or watching a videotape of a trial that might seem more engaging or life-like to participants. Yet, anecdotally, several participants in both studies raised questions about the case out of their own curiosity to the experimenters, demonstrating that they read the materials carefully and had a genuine inquiry regarding what they read. Second, our participants evaluated the case materials individually, but jury deliberations represent an important element of legal decision making as mentioned above. Third, both studies here relied upon college students as participants rather than real jurors. Of course, mock jurors often are not representative of actual jurors in terms of age, education level, or experience, thus the sample used in the present study may not closely parallel the characteristics of actual jurors who sit on a jury; our laboratory simulation can only approximate real juror behaviour (Reifman, Gusick, & Ellsworth, 1992). However, the available research indicates that the behaviour of mock jurors in simulation studies tends not to differ from that of real jurors (see Bornstein, 1999). Nonetheless, given that this
case involved the evaluation of evidence that is sometimes sensationalized in the media (facial composites), it would be useful to examine whether our findings persist among a general population sample.

The experiments reported here provide additional evidence that mock jurors may ignore inconsistencies in an eyewitness’s account and, instead, selectively incorporate evidence into their judgments about the credibility of the eyewitness and the culpability of the defendant. Mock jurors responded to strong vs. weak facial composite evidence as one would expect: A good match composite was viewed as more favourable to the prosecution than a poor match composite. Weak facial composite evidence was largely ignored, however, participants’ assessment of the strength of this evidence was adjusted to be consistent with the other evidence presented. In light of the body of literature spanning the last few decades demonstrating real problems with the accuracy of facial composites (Davies & Valentine, 2006; Frowd et al., 2005) and possible subsequent impairments in memory (Wells et al., 2005), it would be interesting to assess how informing jurors of the challenges associated with facial composite production might affect their evaluations. As it currently stands, it is clear from the results here that facial composites that do not resemble the defendant will be of no assistance to defendants in convincing juries of their innocence.

Acknowledgements
Dawn McQuiston-Surrett, Division of Social and Behavioral Sciences, Arizona State University; Amy Bradfield Douglass, Department of Psychology, Bates College; Samantha G. Burkhardt, Division of Social and Behavioral Sciences, Arizona State University. We thank Roxan Alexander, Mitchell Bartholomew, Gina Erikson, Allyson Horgan, Imran Kirkland, Deanna Kyrimis and Danielle Pieters for assistance in data collection. Portions of these data were presented at the 2004 and 2006 Meetings of the American Psychology-Law Society. Experiment two reflects the third author’s undergraduate honours thesis.

References


People v. Maldonado, 2002 N.Y. Int. 40.


Received 31 January 2007; revised version received 27 April 2007