Varying facial similarity: An automated lineup generator

Caitlin Grist

ACSENT Laboratory
Department of Psychology
University of Cape Town

Supervisor: Colin Tredoux

Word count:

Abstract: 210

Main Body: 8 511

Abstract

Lineup bias poses a large problem to the criminal justice system as innocent suspects may be at risk of being wrongly identified. The two methods used to select lineup foils have been shown to introduce lineup bias as they can result in different levels of identification accuracy. The lineup generator tool may serve as a third method for foil selection. The present study aimed to determine whether this is feasible. Experiment one investigated whether the images generated by the ID programme were perceived as realistic. Experiment two investigated which difference in facial similarity (DIFS) range was optimal in order to maximise the number of accurate identifications while minimising the number of false alarms. Experiment three investigated whether structural lineup bias was present in any of the target-present lineups used in experiment two. The results indicated that participants were accurate in rating the real faces as real however they were unable to determine whether the artificial faces were real or fake at a level greater then chance. The diagnosticity ratios indicated that at DIFS range five participants are 2.63 times more likely to identify a guilty suspect then an innocent suspect. The bias levels and average effective size values indicated that the lineups used in experiment two were fair and unbiased.

Keywords: artificial faces; lineups; difference in facial similarity; identification accuracy; foil selection; lineup bias

In court cases, the identification of a suspect by an eyewitness results in compelling evidence which often leads to a conviction. As such this evidence needs to be reliable. In a standard identification procedure a lineup is used which generally consists of the suspect believed to be the perpetrator and several other people who are known to be non-suspects. These non-suspects are referred to as lineup foils or lineup distractors (Tunnicliff & Clark, 2000). Foils are members of the lineup who are known to be innocent. The perpetrator is the culprit who committed the offense. Biased lineups are frequent and often result in the incorrect identification of an innocent person or a failure to identify the correct suspect. Lineup bias is a large problem as it can potentially increase injustice in the criminal justice system due to incorrect eyewitness identifications (Malpass & Devine, 1981). Lineup fairness has two proposed dimensions – a sufficient lineup size and the absence of lineup bias (Brigham & Brandt, 1992).

Eyewitness identification accuracy and lineup bias can be affected by various variables. One of the most important is system variables. System variables are those that are potentially controllable by the justice system and include lineup content. Lineup content involves the selection of foils and is one of the primary active issues that presents in the eyewitness identification literature. The ideal is achieved when the lineup foils do not enable the suspect to blend into the lineup or enable him/her to stand out. When lineup foils possess a high facial similarity to the suspect the lineup may be biased as this similarity will result in an increase in foil identifications (false alarms) as opposed to the correct identification of the suspect. Correct identification rates will therefore be decreased and false identification rates will increase. In the same vein, when lineup foils possess a low facial similarity to the suspect the lineup may also be biased. This will occur due to the foils not being realistic alternatives to the suspect and as such increasing the chance that the suspect will be identified. Wells, Rydell and Seelau (1993) propose that foils should be selected based on their ability to disguise the suspect in cases when the suspect is not the perpetrator and their ability to make the suspect stand out in cases when the suspect is the perpetrator.

Background

The Selection of Foils and Lineup Fairness

Lineup bias. In order for a lineup to have an absence of bias the suspect should not possess a distinctive appearance compared to the foils. Discussions regarding lineup bias are frequent in the eyewitness literature and many authors have argued that there is a need to

measure this bias (Wells & Bradfield, 1999). A fair lineup is a lineup in which all the major physical characteristics of the lineup foils are similar to all the major physical characteristics of the suspect. A lineup is biased when the lineup foils have physical features that are not shared by the suspect (Luus & Wells, 1991). How lineup foils are selected has been shown to be a source of lineup bias as certain methods of selection often result in higher levels of accuracy.

Wells and Bradfield (1999) argue that the mock-witness procedure has long been accepted as a suitably effective way in which to assess structural lineup bias. The mock-witness method provides a number of participants with a verbal or written description of a suspect who they have never seen and then provides them with the original lineup used and requests them to select the suspect based on the description they were given. The basic assumption of the mock witness paradigm is that if a lineup is fair, then the mock witnesses should select the suspect with a probability no greater than chance (Malpass, Tredoux, & McQuiston-Surrett, 1995).

Lineup size. A lineup with sufficient size is one that is large enough to result in a low probability of a chance identification of an innocent suspect (Brigham, Ready, & Spier, 1990). The nominal size of a lineup is the total number of people in the lineup. The effective size of a lineup refers to the number of plausible lineup members, whereby the maximum is k (the nominal size) and the minimum is 1, which assumes that the mock-witness method is a forced-task. A plausible filler is one who is a realistic alternative to the suspect. A lineups' functional size is another measure of the number of plausible lineup members. In order to determine the functional size of a lineup, you look at the reciprocal value of the proportion of the mock witnesses who selected the suspect (Malpass, Tredoux, & McQuiston-Surrett, 2005). Several criticisms are raised against the notion of functional size and effective size is thus promoted as a better measure. Effective size is determined by making a subtraction from 1 for each lineup member who is selected at a rate that deviates from the chance expectation (Malpass, Tredoux, & McQuiston-Surrett, 2005).

In a study conducted to assess the various measures of lineup size and lineup bias, Brigham et al. (1990) found that effective size was demonstrated to show significant discriminability between fair and random lineups as well as significant sensitivity to the fairness of a lineup (Brigham, Ready & Spier, 1990). Malpass's (1981, as cited in Brigham, Ready, & Spier, 1990) suggestion that a fair lineup should have an effective size that is 83% as large as its nominal size is critiqued as being unrealistic to achieve.

Functions of lineup foils. Lineup foils serve four important functions in lineups and each of these functions has an underlying assumption (Luus & Wells, 1991). An eyewitness's recognition memory may be fallible. When lineups are used in real police practice there is the possibility that the witness will identify one of the lineup foils as being the suspect. Foils therefore allow for responses made by eyewitnesses to be known errors in real-life cases. In actual cases the police will never know with certainty whether their suspect is the perpetrator and therefore the only response made by an eyewitness that can be classified with certainty into the category of accurate or inaccurate is the identification of a lineup foil.

A second related function of lineup foils is the ability to control for chance (Luus & Wells, 1991). When the foils used in a lineup are known to be innocent of the offense they serve as a statistical control. If a guessing strategy is used by an eyewitness the chance that they will identify the suspect will be equal to 1/N (N referring to the number of lineup members).

Lineup foils also help to ensure that the eyewitness does not make use of deductive reasoning in order to determine which of the lineup members the suspect is. Before a lineup is used in police practice the eyewitness provides the police with a description of the perpetrator which the eyewitness will then use to determine which of the lineup members' possesses the characteristics they described (Luus & Wells, 1991). This function is therefore related to the quality or plausibility of the foils used. When a lineup is constructed, foils should be selected carefully in order to ensure that the structure of the lineup does not indicate to the eyewitness which member of the lineup the police would like them to pick out. When a lineup has been properly constructed it will serve as a test of the eyewitness's memory as opposed to a test of their deductive reasoning.

The last function of lineup foils is to ensure that the lineup serves as a test of recognition memory as opposed to a test of recall. Although related to the third function this function provides the potential for accurate identification rates to be increased or preserved through the specification of factors. When an eyewitness provides the police with a description of the perpetrator they are engaging in a recall task. The purpose of a lineup is to gain additional information regarding the perpetrator. Lineups are conducted on the assumption that the eyewitness possesses information in their recognition memory that goes beyond what was available during the recall task (Luus & Wells, 1991). Based on these four functions it can be seen that the selection of foils for a lineup requires careful consideration. The literature proposes two main methods for the selection of foils (Tunnicliff & Clark, 2000).

Methods for selecting lineup foils. The two methods used in the literature to select foils are the similarity-to-suspect strategy and the match-to-description strategy. The similarity-to-suspect strategy requires foils to be selected based on an analysis of the major physical characteristics of the suspect. These major characteristics include features such as age, height, weight, race and sex. Other features can include hair colour, hair style, eye colour, body build and the like. Foils selected using this method must be matched on these predetermined characteristics. Problems with the similarity-to-suspect strategy involve knowing when to stop the analysis of characteristics and whether foils should be matched to the suspect only on the major physical characteristics or to all possible characteristics.

The match-to-description strategy requires foils to be selected based on the description of the perpetrator given to the police by the eyewitness. Theoretically this method ensures that foils are selected based on their similarity to the perpetrator as opposed to their similarity to the suspect. Generally the suspect used in the lineup is selected according to their similarity to the description. It can be assumed that the suspect would not be a suspect if they did not at least in part fit the description of the perpetrator given by the eyewitness (Luus & Wells, 1991). Both of these methods for foil selection have been alternatively critiqued and supported in the literature.

Criticisms of the similarity-to-suspect strategy. One of the central arguments against the similarity-to-suspect strategy involves the similarity between the foils and the suspect. Wells, Rydell and Seelau (1993) argue that inflated rates of false identification are likely to arise when foils are selected using this method as unnecessary similarity between the foils and the suspect is created. When foils that resemble the suspect are used they may serve to cancel all reliable cues that would lead to recognition. This would result in a reduction in accurate identification rates (Luus & Wells, 1991). Another criticism of this method involves knowing how much resemblance between the suspect and foils is sufficient.

In an experiment designed to compare the similarity-to-suspect strategy and the match-to-description method Wells, Rydell and Seelau (1993) found that that the match-to-description strategy resulted in significantly higher rates of accurate identification than the resemble-suspect strategy. This result was shown to be stable with the use of seven different target faces. A similar study conducted by Lindsay, Martin, and Webber (1994, as cited in Tunnicliff & Clark, 2000) illustrates parallel results although the increase in correct identification rates of the match-to-description lineup was not as large.

Navon (1992) provides support for these results and argues that when foils are not selected using a match-description strategy the lineup becomes biased towards the suspect.

His main argument emphasizes how when foils are immediately matched to the suspect rather than the description first, the bias against the suspect is not removed and the resemble-suspect strategy is therefore not an effective method.

Luus and Wells (1991) argue that the match-to-description strategy should be favoured over the similarity-to-suspect strategy for three reasons. The match-to-description strategy is able to specify the physical characteristics that should be shared by the members in the lineup as well as those physical characteristics that should not be shared by the lineup members. This ensures that the witness can and must rely on their recognition memory.

When the similarity-to-suspect strategy is used witnesses may not be able to discriminate between the foils and the suspect. Although the witness may be able to recognise certain features that were not available to their recall memory whilst giving their description of the perpetrator they may not be able to use this to identify the suspect as these features will exist in all the foils as well (Luus & Wells, 1991). If a match-to-description strategy had been used the witness would have been able to make a positive identification.

Criticisms of this strategy fall into two main categories – no clear definitions of similarity between the foils and the suspect; and a reduction in the likelihood of the witness correctly identifying the suspect when s/he is guilty (Tunnicliff & Clark, 2000).

Criticisms of the match-to-description strategy. Lindsay, Martin and Webber (1994) argue that although several authors make compelling arguments against the use of the match-to-suspect strategy there may be several problems with the favoured match-to-description method. When an eyewitness provides a vague or limited description of the perpetrator the risk that an innocent suspect may be incorrectly identified increases. Even if all the foils match the description given by the eyewitness they may be noticeably different from the suspect in appearance. If the definition was vague, possibly with some important physical characteristics missing, there is a chance that the foils will match the description yet not the suspect on these characteristics. The suspect may stand out as being prototypical of the description that was provided. This problem will commonly arise when witnesses notice particular features yet fail to mention them.

One reason put forward for eyewitnesses neglecting to mention particular physical characteristics involves the presence of default values. In a racially homogenous community for example, a witness may neglect to mention the race of the perpetrator if s/he was a member of the dominant group (Lindsay, Martin, & Webber, 1994). Thus race would be viewed as a default value. When receiving the description of the perpetrator the police may not think to ask for clarification should they share the same expectations as the eyewitness.

Another example of a default value would be the sex of a perpetrator in the case of rape due to this type of offense being most frequently committed by one sex rather than the other.

A vague description is a very important problem to the construction of a lineup as it could result in an increase in the level of bias of a lineup. Foils may be noticeably different in appearance to the suspect if the suspect matches several default values not mentioned in the description and not shared by the foils in the lineup. As a result the lineup will be biased towards the suspect.

In a study designed to test the frequency of vague descriptions by eyewitnesses Lindsay, Martin and Webber (2004) found that participants most commonly focused only on the clothing, hair colour and height of the perpetrator and were very vague in their descriptions of these characteristics. Other more specific physical features were ignored. The characteristics of age, race and sex were described by less than half of the witnesses in the study, possibly due to them being viewed as default values.

One of the most important problems regarding the use of a match-to-description strategy appears to be the restriction of features to be matched by the foils to only those features that are described by the eyewitness during their free recall. It is suggested that future research should focus on identifying what degree of detail is required during the description of a perpetrator if a match-to-description strategy is going to be used for foil selection (Lindsay, Martin, & Webber, 2004). This will reduce the rates of incorrect identification.

Criticisms of comparing the two methods. Darling, Valentine and Memon (2008) dispute the above results and argue that there is in fact no difference between the match-to-description and similarity-to-suspect strategies. They state that earlier research has resulted in the similarity-to-suspect method being replaced with the match-to-description method and they argue that this conclusion is premature. Thus, it can possibly be argued that a third method for selecting foils is needed in order to correct for the criticisms of the existing two methods.

Facial similarity of lineup foils and suspect. A similar debate in the eyewitness literature relates to the similarity of the target face/suspect to the foils used in the lineup. It is widely argued that the greater the similarity between the suspect and the lineup foils, the poorer the recognition performance of the eyewitness will be (Laughery, Fessler, Lenorovitz, & Yoblick, 1974). Various techniques have been used in order to assess the relative facial similarity between faces. These include the matching of decoys to suspects on the basis of ratings of particular physical features and the subjective assessment of physical resemblance.

Previous research that has made use of these techniques, demonstrates a clear, consistent pattern which illustrates how lower similarity foils result in a significantly higher number of correct identifications. It has also been argued that the level of similarity between the suspect and the foils is more important than the number of foils used in the lineup (Davies, Shepherd, & Ellis, 1979).

Hierarchical cluster analysis (HCS) separates target and decoy faces into various clusters based on differences on four dimensions: hair, facial structure, age and eyes. Experiments based on this technique have found that participants confuse the target face with the decoy faces at a much more frequent level when the target and decoy faces are drawn from the same cluster (Davies, Shepherd, & Ellis, 1979). These results have been shown to remain stable when other factors, such as instructions given to participants highlighting the importance of careful selection and delayed testing, are also investigated (Laughery et al., 1974).

One potential solution for the problem created by the selection of foils and the facial similarity debate is the use of the ID programme. The ID software is a face composite generator with the ability to generate photo quality face reconstructions (Tredoux, Nunez, Oxtoby, & Prag, 2007). ID also provides the user with the ability to upload a facial image of a target face and then, using the lineup generator tool, generate a specified number of foils that are artificially created by the programme based on a specified difference in facial similarity (DIFS) range. DIFS ranges from 0 (identical) to 1 (non-identical).

Experiment One

The present study used the ID programme in order to generate lineups consisting of artificial foils. As such it was necessary to conduct a study in order to investigate whether participants perceive the artificially generated faces to be realistic.

Research Design and Setting

In order to investigate whether the faces generated using the ID programme were perceived to be realistic participants were shown 25 real and 25 artificial faces and asked to indicate which were real and which were fake. They were also asked three meta-cognitive questions in order to gain insight into the decision they made. The experiment was run on E-Prime and took place in the ACSENT laboratory in the Department of Psychology.

Participants

Fifty participants were recruited using SRPP. No inclusion or exclusion criteria were used.

Materials

Images. Twenty-five images of real faces were randomly selected from a pre-existing database of real faces attached to the ID programme. The real face images were ones that had been previously uploaded and mapped into the programme. The 25 artificial face images were generated by the ID programme using the lineup generator tool.

Procedure. Participants were shown 25 real faces and 25 artificial faces one at a time. After each picture they were asked whether they thought the face was real or fake, how natural they thought the face was, how real they thought the face was and how good the picture quality was. For the meta-cogntive questions participants made their decision from one (low) to nine (high) on a Likert-type scale. The order in which they saw the 50 faces was randomised as was the order in which they were asked the meta-cognitive questions in order to counterbalance the experiment.

Results

A one-sample *t*-test was performed in order to determine whether the participants' accuracy in identifying which faces were real and which were fake differed from chance expectation. All assumptions were upheld. For the real faces $t\ (48) = 2.53$, $r = 0.34$. For the fake faces $t\ (48) = -0.43$, $r = 0.06$. The calculated *t* value for the real faces is greater than the critical *t* value. Thus we do not accept the null hypothesis and we therefore conclude that there is a difference between the sample mean and the population mean. The calculated *t* value for the fake faces is less than the critical *t* value. Thus we do not reject the null hypothesis and we therefore conclude that there is no difference between the sample mean and the population mean. When the face was real participants had an accuracy level greater than chance. However when the face was artificial participants were unable to determine whether the face was real or fake. The associated effect sizes can be seen in table 1.

Table 1.

One-sample *t*-test results for real and fake accuracy

	Mean	Population mean	Standard dev	*df*	*t* crit	*t* calc	*r*
Real	0.67	0.5	0.47	48	2.0086	2.53	0.34
Fake	0.47	0.5	0.49	48	2.0086	-0.43	0.06

Note. Population mean = chance (0.5)

A paired samples *t*-test was performed in order to determine whether there was a significant difference between the participants' ratings of realness, picture quality and naturalness for the real faces and the fake faces. All assumptions were upheld. The results, seen in table 2, indicate that the differences are statistically significant. Pearson's correlation coefficient, *r*, was calculated in order to determine effect size.

On average, participants rated the real faces significantly higher in realness (M = 5.77) than the fake faces (M = 4.83), t (48) = 4.8, $p < .05$, r = .14. On average, participants rated the real faces significantly higher in picture quality (M = 5.71) than the fake faces (M = 4.77), t (48) = 5, $p < .05$, r = .14. On average, participants rated the real faces significantly higher in naturalness (M = 5.67) than the fake faces (M = 4.8), t (48) = 4.5, $p < .05$, r = .13. Although the results are significant, Pearson's r indicates that the effect sizes are small. Although a difference was found between the real and fake faces in terms of realness, naturalness and picture quality it was small. The participants rated the real faces and the fake faces similarly in terms of realness, naturalness and picture quality.

Table 2.

Mean values and *t*-statistics for meta-cognitive questions

	Mean		Standard Dev		*t* stat	*p*	*t* crit	*r*
	Real	Fake	Real	Fake				
Realness	5.77	4.83	2.35	2.29	4.8	< 0.05	1.96	0.14
Picture Quality	5.71	4.77	2.19	2.15	5	< 0.05	1.96	0.14
Naturalness	5.67	4.8	2.3	2.25	4.5	< 0.05	1.96	0.13

Discussion

The results indicate that the participants' had a greater than chance accuracy in rating the real faces as being real. This indicates that participants' were generally correct in rating the real faces as real. However when the faces were artificial participants' accuracy levels did not differ from chance expectation. This indicates that when the faces were artificial participants were unable to determine whether or not the faces were real or fake. It is likely that participants' employed a guessing strategy when choosing whether the artificial faces were real or fake as their accuracy was no different to chance. Although participants' may have noticed differences between the real faces and the artificial faces it did not lead them to the conclusion that the artificial faces were fake. This demonstrates that the artificial images

generated by the ID programme look realistic and are perceived to be real. This is a valuable finding which provides support for the ID programme.

Meta-cognitive data was collected and analysed in order to gain insight into how the participants perceived the real and fake faces in terms of realness, naturalness and picture quality. It was found that participants rated the real faces higher in terms of realness, naturalness and picture quality. Despite the significant findings the effect sizes for all three meta-cognitive questions were small indicating that the differences between the real and fake faces were not very large. The real faces and fake faces were rated fairly similarly in terms of realness, naturalness and picture quality. This provides further support for the artificial images generated by the ID programme as they were not rated as being hugely different, in terms of realness, naturalness and picture quality, compared to the real faces.

Experiment Two

Hits represent the accurate identifications made by participants. False alarms represent the incorrect identification of a foil. In an optimal situation hits are maximised while false alarms are minimised.

The majority of the experimental research conducted in this area makes use of a particular procedure during which participants view a photograph of a perpetrator followed by a lineup. This standard experimental procedure allows for the conditions of real criminal investigations to be simulated (Clark & Tunnicliff, 2001). Participants are either presented with a perpetrator-present lineup or a perpetrator –absent lineup. A perpetrator-absent lineup is one in which the perpetrator is replaced by a similar looking other. This simulates instances in real criminal investigations when police do not know whether the suspect is the perpetrator. The current experiment made use of both perpetrator-present and perpetrator-absent lineups.

Research design and setting

A quantitative design was used with identification accuracy (for the perpetrator-present lineups this will be the correct identification of the target face shown at time one and for the perpetrator-absent lineups this will be the correct rejection of the lineup) as the dependent measure. Participants were randomly assigned into one of thirty groups. Twenty groups were shown perpetrator-present lineups and ten groups were shown perpetrator-absent lineups. Each group viewed five different perpetrators at one of the five DIFS ranges. Participants each viewed a photograph. They then completed three minutes of Sudoku as a distractor task. After three minutes they were shown a photo lineup and asked to indicate the number of the lineup member they believed to be the perpetrator they saw in the photograph

at time one. This process was repeated five times (once per perpetrator). The experiment took place in the ACSENT laboratory in the Department of Psychology and was run on E-Prime.

Participants

Ninety participants were recruited using SRPP. No inclusion or exclusion criteria were used.

Materials

The independent variables were target face/perpetrator (one; two; three; four; or five) and DIFS range (one; two; three; four; or five). The dependent variable was identification accuracy. Independent variable one (target face) was a within groups variable. Independent variable two (DIFS range) was a between groups variable.

Photographs. All of the participants were shown five photographs, one at a time, representing the target faces/perpetrators (head and upper body shots). The use of five different perpetrators controlled for any confounds that could have arisen due to one of the faces being more distinctive then the others. This ensured that the results were not affected should one of the perpetrators have had a more recognizable face. The photographs of the five perpetrators were randomly selected from a pre-existing database of faces attached to the ID programme.

Lineup construction. Fifty lineups consisting of six lineup members each were constructed for the present study. Each lineup was constructed using the lineup generator tool of the ID programme. Five lineups were generated for each of the five perpetrators. Each of these lineups was constructed based on a specified DIFS range. DIFS range one ranged between 0.5 and 0.55 and was the most similar. DIFS range two ranged between 0.6 and 0.65. DIFS range three ranged between 0.7 and 0.75. DIFS range four ranged between 0.8 and 0.85. DIFS range five ranged between 0.9 and 0.95 and was the most different. Each target face had five lineups corresponding to one of the five DIFS ranges. Each target face had a perpetrator-present lineup and a corresponding perpetrator-absent lineup. The perpetrator-present lineups consisted of the perpetrator and five artificial computer-generated foils. The perpetrator-absent lineups consisted of a similar looking other and five artificial computer-generated foils. All photographs used in the lineups were head and upper body shots against a textured brick background. Lineup members were all presented at the same time (the perpetrator/similar looking other and the five foils).

Lineup response screen. After viewing each of the five lineups (one for each of the perpetrators at one of the five DIFS ranges) participants were asked to complete the lineup response screen. The lineup response screen consisted of the lineup and participants were

asked to press the number key corresponding to the person of their choice. They were also given the option to press number nought if they thought the person was not present or if they didn't know.

Procedure

Participants viewed an image of the perpetrator on the computer screen for two seconds. They then completed the Sudoku puzzles that were handed out to them at the beginning of the experiment for three minutes. After three minutes they were presented with a lineup and were asked to press the number key corresponding to the person they believed to be the man they saw in the picture. They were allowed to take as long as they needed to make a decision. Once they had pressed a number key an image of the second perpetrator was shown to them for two seconds. The exact procedure was repeated until all five perpetrators and all five corresponding lineups had been seen. The order in which they saw the perpetrators and the position of the perpetrator in the lineup was counterbalanced. After viewing each of the lineups the participants were asked three meta-cognitive questions investigating how confident they were about their decision, how easy it was for them to choose and how similar they thought the men in the lineup were. They were asked to rate their decision on a Likert-type scale from nought (not at all) to 100 (very much).

Results

The data was divided into perpetrator-present and perpetrator-absent as the identification accuracy decision is different for each type of lineup (Pozzulo & Lindsay, 1999, as cited in Pozzulo, Crescini, & Panton, 2008). For the perpetrator-present lineups, a correct response involved making the correct identification of the target whereas an incorrect response involved making a false alarm (choosing a foil) or a miss (incorrectly rejecting the lineup). In the perpetrator-absent lineups, a correct response involved the correct rejection of the lineup whereas an incorrect response was a false positive (choosing a foil). The frequency of hits and correct rejections were calculated for each of the DIFS ranges for the perpetrator-present and perpetrator-absent lineups and can be seen in table 3. From table 3 it can be seen that the greatest percentage of hits was made at DIFS range four while the most correct rejections (and fewest false alarms) were made at DIFS range five.

Table 3.

Frequency and percentage of hits and false alarms of PP and PA lineups per DIFS range

		PP		PA	
	DIFS	Frequency	Percentage	Frequency	Percentage
Hits	1	16	26.67		
	2	20	33.33		
	3	23	38.33		
	4	29	48.33		
	5	21	35		
Correct rejections	1			17	56.65
	2			13	43.33
	3			15	50
	4			16	53.33
	5			26	86.67

Note. PP = Perpetrator-present PA = Perpetrator-absent.

Diagnosticity was calculated for each DIFS range. Diagnosticity refers to the "probability that the suspect is guilty given a particular response, whereas accuracy refers to the probability of a particular response given that the suspect is guilty or innocent" (Clark & Wells, 2008, p. 407). Diagnosticity is a ratio of likelihoods representing the ratio of the probability that the suspect is identified given that s/he is guilty to the probability that the suspect is identified given that s/he is not guilty (Tredoux, 1998). Diagnosticity ratios were calculated for each of the DIFS ranges and can be seen in figure 1.

Figure 1 illustrates the diagnosticity for each of the five DIFS ranges. From the graph it can be seen that as the difference in facial similarity increases, diagnosticity increases. DIFS range five has the highest diagnosticity ($d = 2.63$) whereas DIFS range two has the lowest diagnosticity ($d = 0.59$).

The homogeneity of diagnosticity test allows for the difference between two or more diagnosticity ratios to be compared (Tredoux, 1998). The homogeneity of diagnosticity test was performed on the calculated diagnosticity ratios in order to determine whether a significant association existed between the DIFS range and diagnosticity. The results indicated that a significant association does exist between DIFS range and diagnosticity χ^2 (4) = 78.9, $p < .05$. This analysis suggests that the difference between the five diagnosticity ratios is not only due to random sampling variation but rather is a function of DIFS range. From figure 1 it is clear from the individual diagnosticity ratios that DIFS range five has the largest associated ratio ($d = 2.63$).

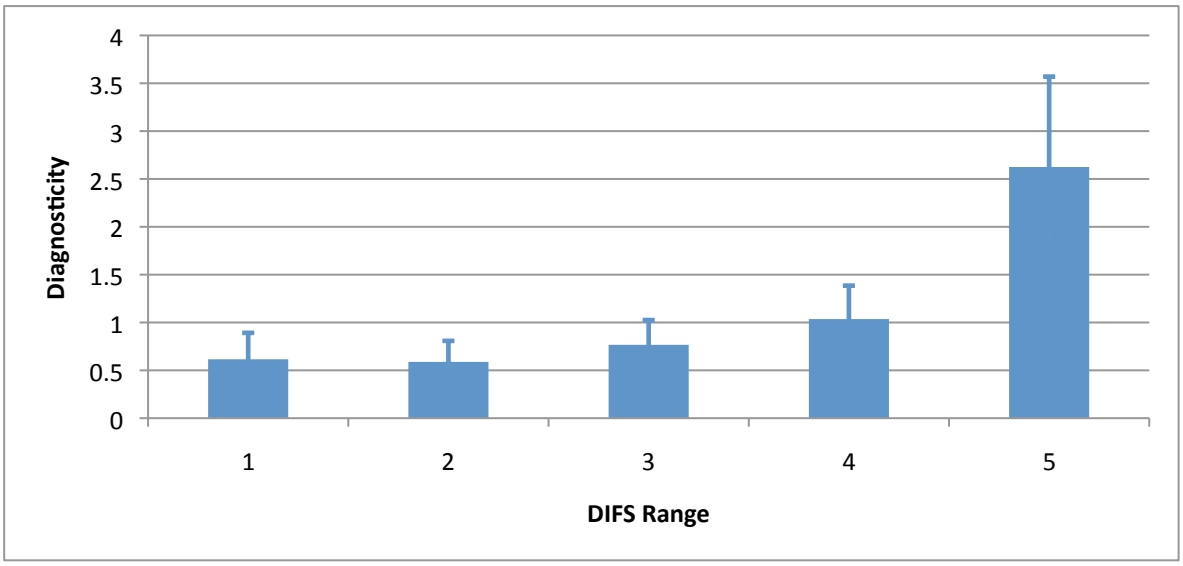

Figure 1. *Diagnosticity with 95% confidence intervals*

Perpetrator-present meta-cognitive questions. A simple one-way ANOVA was conducted in order to determine whether there were significant differences between the average confidence, similarity and how easy to choose ratings made by participants for each DIFS range. All assumptions were met. The results can be seen in table 4.

Confidence. There was a significant effect of DIFS range on confidence level, F (4, 295) = 6.53, $p < .05$, ω^2 =0.07. Tukey's HSD test was performed in order to determine where the differences in the means were. It was found that DIFS range one (M = 46.67) differed significantly from DIFS range two (M = 66), three (M = 69.3), four (M = 72.67) and five (M = 68.5). The results indicate that participants' level of confidence at DIFS one was significantly different to their level of confidence at DIFS five. Participants were significantly more confident when the foils and perpetrator were the most dissimilar. They were the least confident when the similarity between the foils and perpetrator was greatest.

Similarity. There was a significant effect of DIFS range on similarity, F (4, 295) = 6.7, $p < .05$, $\omega^2 = 0.07$. Tukey's HSD test was performed in order to determine where the differences in the means were. It was found that DIFS range one (M = 70.67) differed significantly from DIFS range four (M = 56.17) and five (M = 46), DIFS range two (M = 62.67) differed significantly from DIFS range five (M = 46) and DIFS range three (M = 59.67) differed significantly from DIFS range five (M = 46). The results indicate that participants perceived the lineup members in DIFS range one to be significantly more similar than the lineup members in DIFS range five.

How easy to choose. There was a significant effect of DIFS range on how easy participants' found it to choose, F (4, 295) = 2.43, $p < .05$, $\omega^2 = 0.02$. Tukey's HSD was performed in order to determine where the differences in the means were. It was found that DIFS range one ($M = 48.3$) differed significantly from DIFS range two ($M = 76.46$). Participants' found it significantly easier to make a decision at DIFS range two than at DIFS range one.

Table 4.

ANOVA results for meta-cognitive questions PP

	F	*p*	F crit	*r*	ω^2
Confidence	6.53	< .05	2.4	0.3	0.07
Similarity	6.7	< .05	2.4	0.3	0.07
How easy to choose	2.4	< .05	2.4	0.2	0.02

Perpetrator-absent meta-cognitive questions. From table 5 it can be seen that a simple one-way ANOVA was conducted in order to determine whether there were significant differences between the average confidence, similarity and how easy to choose ratings made by participants for each DIFS range. All assumptions were met.

Confidence. There was a significant effect of DIFS range on confidence level, F (4, 145) = 3.22, $p < .05$, $\omega^2 = 0.06$. Tukey's HSD test was performed in order to determine where the differences in the means were. It was found that DIFS range two ($M = 55.67$) differed significantly from DIFS range five ($M = 77.67$). As with the perpetrator-present lineups, participants were significantly more confident when the foils and the suspect were the most dissimilar.

Similarity. There was no significant effect of DIFS range on similarity, F (4, 145) = 2.01, $p > .05$, $\omega^2 = 0.03$. There were no significant differences between the average similarity ratings of DIFS one to five. This indicates that participants rated all the DIFS ranges similarly in terms of similarity.

How easy to choose. There was a significant effect of DIFS range on how easy it was for participants to choose, F (4, 145) = 3.18, $p < .05$, $\omega^2 = 0.05$. Tukey's HSD test was performed in order to determine where the differences in the means were. It was found that DIFS range four ($M = 49$) differed significantly from DIFS range five ($M = 70.33$). Participants found it significantly easier to make a decision at DIFS range five then DIFS range four.

Table 5.

ANOVA results for meta-cognitive questions PA

	F	*p*	*F* crit	*r*	ω^2
Confidence	3.22	0.01	2.43	0.28	0.06
Similarity	2.01	0.09	2.43	0.22	0.03
How easy to choose	3.18	0.01	2.43	0.28	0.05

Discussion

The results from table 3 indicated that identification accuracy did appear to be influenced by the difference in facial similarity (DIFS) range used in constructing the lineups. It can be seen from table 3 that the highest percentage of accurate identifications was made at DIFS range four. The results demonstrate a clear and consistent pattern illustrating a gradual increase in accurate identifications as the difference in facial similarity increased. This finding is consistent with previous research which has found that the similarity of foils to each other and the suspect has a strong influence over recognition performance (Davies, Shepherd, & Ellis, 1979).

It has been demonstrated consistently in the literature that lower similarity levels have resulted in significantly higher hit rates. As such it can be concluded that a greater level of similarity between the target and the foils results in a lower probability of the target being correctly identified (Laughery et al., 1974). The results indicate that participants were able to better discriminate between the foils and the target face/perpetrator when less similarity was present. When the foils were extremely similar (almost identical) to each other and to the perpetrator, participants were unable to discriminate effectively enough in order to accurately identify the perpetrator. This resulted in an increase in the percentage of foil identifications made. DIFS range one, consisting of lineups with foils who were almost identical to the perpetrators, demonstrated the lowest percentage of accurate identifications.

These findings were confirmed by the diagnosticity calculations. The diagnosticity ratios "indicate the probability that a guilty suspect is identified relative to the risk that an innocent suspect is identified. When diagnosticity is 1.0, the events are equally likely; departure from 1.0, on the other hand, reflects the degree to which events are not equally likely" (Tredoux, 1998, p. 232). The diagnosticity ratio of DIFS range five indicated that at DIFS range five the guilty suspect was 2.63 times more likely to be identified than the innocent suspect. The diagnosticity ratio of DIFS range one indicated that at DIFS range one

the guilty suspect was 0.62 times more likely to be identified than the innocent suspect. These results indicate that the perpetrators were more likely to be identified at DIFS range five. The homogeneity of diagnosticity test indicated a significant association between DIFS range and diagnosticity. This indicates that as the difference in facial similarity between the lineup foils and the perpetrator increases there is a greater probability that the perpetrator will be identified.

The results from the homogeneity of diagnosticity test illustrate that there is a significant difference between DIFS range one and DIFS range five in terms of the likelihood that a guilty suspect will be identified as opposed to an innocent suspect. The diagnosticity results provide support for the use of DIFS range five in order to achieve the maximum level of identification accuracy while minimising the number of false alarms. At DIFS range five there is a greater chance that an accurate identification will be made as opposed to an incorrect identification. Future lineups generated using the lineup generator tool should be generated at DIFS range five in order to maximise the likelihood that the guilty suspect will be identified as opposed to an innocent suspect.

The results for the perpetrator-present meta-cognitive data indicated that participants were significantly more confident in their decisions when they viewed lineups constructed at DIFS range five. As there was less similarity between the foils and the perpetrator the participants were better able to discriminate between the foils and target face resulting in an increase in their confidence levels. The participants also found the foils and perpetrator significantly more similar at DIFS range one then at DIFS range five. This provides support for the calibration of the DIFS ranges as participants did find DIFS range one the most similar and DIFS range two the most different.

Experiment Three

Experiment three aimed to determine whether structural lineup bias was present in any of the perpetrator-present lineups used in experiment two in order to determine whether the lineup generator tool of the ID programme was able to generate fair and unbiased lineups.

Research design and setting

A mock witness procedure was used to determine whether structural lineup bias was present. Participants (mock-witnesses) were given written descriptions of each of the five different perpetrators and then shown five photo lineups that had been constructed at one of the five DIFS ranges.

Participants

Two hundred participants were recruited from the UCT campus using convenience sampling. Anyone who was approached and agreed to participate was included.

Materials

Descriptions. Six participants, sampled using SRPP, were asked to describe the five perpetrators. From these six descriptions, five modal descriptions were generated (one for each perpetrator). These descriptions were used in place of the photographs shown in experiment one. They described the perpetrators major facial features.

Lineup construction. The lineups used in experiment three were identical to those used in experiment two, however only the perpetrator-present lineups were used as the aim was to determine whether the lineups were biased against or in favour of the perpetrators.

Lineup response sheet. The responses made by participants were recorded by the researcher on a lineup response sheet.

Procedure

Participants were given a written description of one of the perpetrators and an A4 copy of the lineup corresponding to that perpetrator. They were then asked to give the number of the person in the lineup who they thought the description was referring to. This exact procedure was repeated for each of the five perpetrators.

Results

Table 8 indicates the level of bias for each of the lineups used in experiment two. In an unbiased lineup the perpetrator should be chosen by the mock-witnesses at a level no greater than chance. A test of two proportions was conducted on each of the lineups in order to determine whether the bias level differed significantly from chance. It was found that the perpetrators' in three of the lineups used were chosen at a level significantly different to chance. As such these three lineups used in experiment two were biased against the perpetrator.

Table 6 indicates the average percentage bias levels per perpetrator. It can be seen that perpetrator three had an average bias level of 30% across the DIFS ranges. This suggests that perpetrator three was not a suitable target face for experiment two as he was chosen by the mock-witnesses more often. As such, perpetrator three was left out of the analysis of effective size.

Table 6.

Percentage bias per perpetrator

Perpetrator	Mean (x)
1	10.5
2	16
3	30
4	9.5
5	13

Table 7.

Percentage bias per DIFS range

DIFS range	Mean (x)
1	12
2	16
3	9.38
4	11.25
5	13.13

Note. Perpetrator 3 excluded

Table 7 shows the average percentage bias levels for each of the DIFS ranges with perpetrator three excluded.

The effective size was calculated for each lineup and then averaged for each of the DIFS ranges. The average effective size can be seen for each of the DIFS ranges in table 9. DIFS range five had the highest effective size (E′ = 3.94). DIFS range three had the smallest effective size (E′ = 3.32). Table 9 indicates the average number of plausible lineup members for each of the DIFS ranges.

Table 8.
Bias per lineup across all DIFS ranges

DIFS range	Perpetrator	Frequency Chosen	Percentage Chosen
1	1	3	7.5
	2	0*	0
	3	4	10
	4	1	2.5
	5	15	37.5
2	1	11	27.5
	2	8	20
	3	5	12.5
	4	3	7.5
	5	3	7.5
3	1	0*	0
	2	9	22.5
	3	25*	62.5
	4	4	10
	5	2	5
4	1	2	5
	2	6	15
	3	12	32.5
	4	6	15
	5	4	10
5	1	5	12.5
	2	9	22.5
	3	13	32.5
	4	5	12.5
	5	2	5

Note. Chance level = 1/*k* (16.667%), *k* = 6, * $p < .05$

Table 9.

E´ (Tredoux, 1998) for the lineups used in experiment 2

	Effective size (E´)				
DIFS range	1	2	3	4	5
Mean (x)	3.93	3.43	3.32	3.52	3.94

Note. Perpetrator 3 excluded

Discussion

Experiment three aimed to determine whether the perpetrator-present lineups used in experiment two were fair. Calculations of Tredoux's (1998) effective size (E') and bias were

derived from the mock witness identifications for each lineup in order to determine whether the lineup generator tool resulted in lineups with different levels of structural bias. It was found that DIFS range five had the highest average effective size followed by DIFS range one. This indicates that the lineups with the most plausible lineup fillers were generated at the greatest difference in facial similarity level. An effective size of three is considered to be acceptable in the eyewitness literature. The average effective size values indicate that across each DIFS range the majority of the foils used were plausible lineup fillers. DIFS ranges four and five had two of the greatest effective size values and the highest number of accurate identifications.

This finding is consistent with previous literature which has found lineup fairness to be a good postdictor of accuracy (Smith et al., 2000, as cited in Tredoux, Parker, & Nunez, 2007). Three out of the twenty-five lineups used in experiment two were found to be biased against and in favour of the perpetrator. It was found that perpetrator three had the largest individual bias level compared to the other perpetrators. A possible explanation for such a discrepancy could relate to the distinctiveness of perpetrator three. He may have had a more recognisable face which in turn resulted in more mock-witness choices. The large majority of the lineups used were found to be unbiased as the perpetrators were chosen at a level similar to chance. This result provides support for the ability of the lineup generator tool to generate unbiased lineups.

Previous literature has demonstrated that the higher the similarity between the foils and the suspect, the higher the bias levels will be (Tredoux, Parker, & Nunez, 2007). It has been stated previously that the similarity criterion, when taken to the extreme, would result in the lineup fillers being almost identical to the suspect (Wogalter, Marwitz, & Leonard, 1992). Resultantly it has been argued that if such a lineup could be constructed it would be extremely biased as it would simply be a show-up (the perpetrator shown by him/herself without lineup foils) in that the perpetrator would be seen several times. This argument was not supported as the results from experiment three indicated that only three of the lineups used in experiment two demonstrated bias levels above chance despite DIFS range one consisting of lineups in which the foils and perpetrator were almost identical.

It was demonstrated that the majority of the lineups were unbiased. The average effective size values also indicate that across each DIFS range the lineups were fair. One possible explanation for this discrepancy is that all the lineups were generated using the same programme which may have had an influence over the amount of bias demonstrated.

Conversely Shepherd, Ellis and Davis (1982, cited in Wogalter, Marwitz, & Leonard, 1992) have argued that a lineup consisting of foils who are almost identical to each other and the suspect will be extremely fair as the choices made by mock witnesses are more likely to be distributed amongst all the lineup members. The results from experiment three supported this argument as it was found that the majority of the lineups were unbiased and that DIFS range one had the second largest average effective size (E′ = 3.93). This indicates that the lineups generated at DIFS range 1 consisted of the most plausible lineup fillers. These results indicate an inverse relationship between identification accuracy and lineup fairness. The DIFS range one lineups resulted in the lowest percentage of accurate identifications yet the majority of the lineups were found to be fair. The results from experiment three provide support for the ability of the lineup generator tool to generate lineups with an acceptable level of plausible lineup fillers and very little bias.

Limitations and recommendations for future research

The present study made use of students as participants. As such the results might not generalize to the general population. Future research should aim to use a more representative sample. The calculated E′ values indicate fairly similar levels of fairness across each DIFS range. Several sets of different foils are able to be generated for the same target face which may result in different E′ values and levels of bias. Future research should aim to investigate this ability in order to determine whether the various sets of foils generated differ in their levels of fairness and bias.

One last potential limitation of the present study involves the calibration of the DIFS ranges used. Although the five DIFS ranges used were sufficient enough to result in adequate results the differences may have been greater had different DIFS ranges been used. It would be interesting for future research to calibrate the DIFS ranges differently in order to maximise the potential differences. The difference between DIFS one and DIFS five could be greater.

Conclusion

The results from the current study provide support for the use of an automated lineup generator tool as a possible third method for selecting lineup foils. Previous literature has critiqued the two most commonly used methods for selecting foils as they often result in the foils being too similar or too different to the suspect (Lindsay, Martin, and Webber, 1994; Luus & Wells, 1991). An automated lineup generator provides an attractive solution as provides the ability to specify the desired level of similarity between the foils and suspect. The current study found that the ID programme is able to generate realistic and natural looking foils. The diagnosticity ratios indicated that DIFS range five is optimal for achieving

the maximum number of accurate identifications and the minimum number of incorrect identifications. The results also support the ability of the lineup generator tool to generate fair and unbiased lineups.

References

Brigham, J., & Brandt, C. (1992). Measuring lineup fairness: Mock witness responses versus direct evaluations of lineups. *Law and Human Behaviour, 16,* 475-489.

Brigham, J., Ready, D., & Spier, S. (1990). Standards for evaluating the fairness of photograph lineups. *Basic and Applied Social Psychology, 1990,* 149-163.

Clark, S., & Tunnicliff, J. (2001). Selecting lineup foils in eyewitness identification experiments: Experimental control and real-world simulation. *Law and Human Behaviour, 25,* 199-216.

Clark, S., & Wells, G. (2008). On the Diagnosticity of Multiple-Witness Identifications. *Law and Human Behaviour, 32,* 406-422.

Darling, S., Valentine, T., & Memon, A. (2008). Selection of lineup foils in operational contexts. *Applied Cognitive Psychology, 22,* 159-169.

Davies, G., Shepherd, J., & Ellis, H. (1979). Similarity effects in face recognition. *American Journal of Psychology, 92,* 507-523.

Laughery, K., Fessler, P., Lenorovitz, D., & Yoblick, D. (1974). Time delay and similarity effects in facial recognition. *Journal of Applied Psychology, 59,* 490-496.

Lindsay, R., Martin, R., & Webber, L. (1994). Default values in eyewitness descriptions: A problem for the Match-To-Description lineup foil selection strategy. *Law and Human Behaviour, 18,* 527-541.

Luus, C., & Wells, G. (1991). Eyewitness identification and the selection of distractors for lineups. *Law and Human Behaviour, 15,* 43-57.

Malpass, R., & Devine, P. (1981). Eyewitness identification: lineup instructions and the absence of the offender. *Journal of Applied Psychology, 66,* 482-489.

Malpass, R., Tredoux, C., & McQuiston-Surrett, D. (2005). Lineup construction and lineup fairness. In R. Lindsay, D. Ross, J. D. Read & M. P. Toglia (Eds.), *Handbook of Eyewitness Psychology (Vol. 2): Memory for People* (pp. 155-178). Lawrence Erlbaum & Associates.

Navon, D. (1992). Selection of lineup foils by similarity to the suspect is likely to misfire. *Law and Human Behaviour, 16,* 575-593.

Pozzulo, J., Crescini, C., & Panton, T. (2008). Does methodology matter in eyewitness identification research?: The effect of live versus video exposure on eyewitness identification accuracy. *International Journal of Law and Psychiatry, 31,* 430-437.

Tredoux, C. (1998). Statistical Inference on Measures of Lineup Fairness. *Law and Human Behaviour, 22,* 217-236.

Tredoux, C., Nunez, D., Oxtoby, O., & Prag, B. (2007). An evaluation of ID: An eigenface based construction system. *South African Computer Journal, 37,* 90-96.

Tredoux, C., Parker, J., & Nunez, D. (2007). Predicting eyewitness identification accuracy with mock witness measures of lineup fairness: Quality of encoding interacts with lineup format. *South African Journal of Psychology, 37,* 207-222.

Tunnicliff, J., & Clark, S. (2000). Selecting foils for identification lineups: Matching suspects or descriptions? *Law and Human Behaviour, 24,* 231-258.

Wells, G., & Bradfield, A. (1999). Measuring the goodness of lineups: parameter estimation, question effects, and limits to the mock witness paradigm. *Applied Cognitive Psychology, 13,* 27-39.

Wells, G., Rydell, S., & Seelau, E. (1993). The selection of distractors for eyewitness lineups. *Journal of Applied Psychology, 78,* 835-844.

Wogalter, M., Marwitz, D., & Leonard, D. (1992). Suggestiveness in Photospread Lineups: Similarity Induces Distinctiveness. *Applied Cognitive Psychology, 6,* 443-453.

Appendix A

Experiment One Consent Form

Name: _____ Student Number: _____

Course code: _____

Real/Fake Study

The present study requires you to look at several faces and rate them according to the instructions given.

You will recieve 1 SRPP point (30 minutes) for participating in the present study.

No physical, mental or emotional stresss should be caused by the present study however should you wish to stop for any reason you are free to leave or notify the researcher.

Precautions will be taken to ensure that all of your personal information is safeguarded throughout the study.

Signature: _____ Date: _____

Appendix B

Experiment Two Consent Form

Eyewitness Study: Participant Number:_____

Participant Consent Form

Thank you for your participation in this study. The following demographic information is required. All the information you provide on this form will be kept confidential – only the results from the study will be reported in the research project. Your student number will only be used to allocate your SRPP point.

Student Number: _____

Sex: _____

Age: _____

Race: Black White Coloured Indian Other

You will recieve 1 SRPP point (30 minutes) for participating in the present study.

No physical, mental or emotional stresss should be caused by the present study however should you wish to stop for any reason you are free to leave or notify the researcher.

Precautions will be taken to ensure that all of your personal information is safeguarded throughout the study.

The study: The purpose of the present research study is to investigate whether the ID programme can be used as an alternate method for the selection of lineup foils/distractors (innocent non-suspects in a lineup). In the present study you will be required to look at several faces and several lineups.

If you have any questions, complaints or concerns regarding this research study you may contact the researcher, Caitlin Grist, via email, grscai001@uct.ac.za.

By signing this consent form, I hereby give my consent to participate in the study and my permission for any results obtained to be used in the research project. I acknowledge that I have read through this form and filled in the required information. I understand that no personal information will be used in any way.

Signature: _____

Date: _____

PLAGIARISM

This means that you present substantial portions or elements of another's work, ideas or data as your own, even if the original author is cited occasionally. A signed photocopy or other copy of the Declaration below must accompany every piece of work that you hand in.

DECLARATION

1. I know that Plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.

2. I have used the American Psychological Association formatting 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, 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.

DATE:

SIGNATURE:

STUDENT NUMBER: