Descriptions Given by Eyewitnesses Affect Lineup Fairness

Danielle Chevalier

Augsburg University

Abstract

Eyewitness descriptions are essential when constructing a police lineup, as an eyewitness's description is the basis for selecting the members of the lineup. Unfortunately, eyewitnesses' descriptions are often vague, because eyewitnesses may fail to mention features of the culprit. We believe that vague descriptions create unfair lineups, as the lineup would fail to match for features not mentioned by the witness. To test this hypothesis, participants viewed 20 photos of suspects pulled from lineups with vague descriptions, and described the race, sex, age, size (height, weight, build), hair, and face of each suspect. Then, new framework descriptions were created using the collected data. Seventy-three participants in a mock-witness procedure then provided a basis for measuring the fairness of each lineup, viewing both the old description lineups and the new framework lineups in a within subject design. We predict that the old descriptions will have higher bias rates than the new framework descriptions. Our hypothesis was supported for the proportion score, functional size, effective size, number of rejected lineup members, and mean rank bias measures. Features not mentioned in vague descriptions can lead to unfair lineups.

Eyewitness Descriptions: Default Values Affect Lineup Fairness

In many cases of police lineups, the eyewitness makes an error, choosing an innocent person from the lineup (Findley, 2013). As of 2011, more than 250 people had been exonerated by DNA tests, and more than 75% of these people were victims of mistaken eyewitness identification (Garrett, 2011, p. 48). Even if eyewitnesses are certain of their decisions, they can be wrong and send innocent men and women to prison. Studies show that juries are readily persuaded by eyewitness identification testimony, and eyewitness testimony plays a significant role in the conviction of a defendant (Wells, Leippe & Ostrom, 1979). Mistaken identifications can destroy the life of an innocent person who is falsely convicted. Innocent people have even been sentenced to death and executed after they were falsely identified from a lineup (Innocence and the Death Penalty Capital Punishment in Context, n.d.). Research can provide insight into what can be done to improve eyewitness identification procedures, to reduce the risks of false identifications in lineups, and to increase the identification of true guilty suspects.

This study is concerned with the completeness of eyewitness descriptions, and how incomplete descriptions can potentially lead to problems in identification procedures and eyewitness decisions. First, it is important to consider what eyewitness science understands about sources of eyewitness error, police procedures that exacerbate error, and the role of eyewitness description of the offender in this process.

Sources of Error

Eyewitness error begins at the time of the crime. An eyewitness's attention, perception, and encoding of memory can be inhibited by many factors, including aspects of the crime event, characteristics of the offender, and characteristics of the eyewitnesses themselves (Wells, 1978).

These *estimator* variables are aspects of a crime that can influence the memory of the eyewitness. Poor lighting, brief duration of the crime, presence of a weapon that draws attention, and distance between the perpetrator and the witness can exacerbate error. Characteristics of the offender that can lead to error include a facial covering or disguise, being a stranger, or being a race different from the eyewitness. Characteristics of the eyewitness can also lead to errors; fear, stress, or intoxication add to the likelihood of the eyewitness making an error (Steblay, 2015). Estimator variables may exacerbate eyewitness error. Unfortunately there is little that can help diminish these factors, as they happen during the crime and are therefore out of the control of law enforcement.

On the other hand, *system* variables are factors that also influence eyewitness decisions, but are potentially within law enforcement control because they occur during the police investigation (Wells, 1978; Wells, Memon, & Penrod, 2006). Examples of these factors include the time between the crime and the police interview, interaction with other witnesses or the police, leading interview questions, the witness's description of the culprit, and the structure and process of identification lineups. System variables greatly affect an eyewitness's memory and decision-making. For example, witnesses are often motivated to pick someone from the lineup, even if they are unsure, because they believe that the police have good reason for organizing the lineup (Wells, 1978). Additionally, if a witness can determine which member of the lineup the police suspects, that lineup member might be more likely to be picked from the lineup by the witness (Wells, 1978). Consequently, research on system variables may help to prevent eyewitness error from occurring.

Lineup Construction

Lineup construction is the foundation of the current study. A police lineup serves several functions: it tests the police hypothesis that the suspect is the guilty offender, tests witness memory and alerts the police when the witness's memory is unreliable, and protects an innocent suspect. In order to accomplish these functions, rules for building a lineup have been established. First, the lineup should always contain one, and only one, suspect. This is the police suspect of interest. The witness's response to this suspect will offer evidence for or against the police hypothesis about the suspect's guilt.

The lineup must also be able to reveal unreliable witness memory and protect an innocent suspect. Therefore, a second rule is that the remainder of the lineup members should be known-innocent fillers (Wells, Memon, & Penrod, 2006). If a witness picks a filler from a lineup, the police are able to determine that the witness's memory is unreliable. The presence of fillers also protects an innocent suspect by reducing the likelihood that this suspect will be chosen by chance.

A third rule is that the suspect must not stand out among the lineup members. All members must be similar in appearance to the suspect. The suspect's photo should not be different in size, background, or clothing compared to the fillers. Characteristics of the lineup that make a suspect stand out and increase the chances that the suspect is identified are called structural biases. These biases can put innocent people at risk of being falsely identified (Wells & Bradfield, 1999). A fair lineup will protect suspects, guilty or innocent, from standing out compared to the other lineup members and will therefore be free of structural bias. Studies have shown that even if a suspect is guilty, unbiased lineups do not negatively affect rates of correctly choosing the suspect (Wells & Bradfield, 1999).

Match-to-description vs. match-to-suspect. Lindsay and Wells (1980) found that an innocent suspect is about two and a half times more likely to be falsely identified if that suspect is the only one to fit the witness's description compared to lineups with at least two other fillers matching the description. Therefore, fair lineup construction must begin with the appearance of the culprit. Lineup construction can be based on whether the lineup fillers are chosen to match the description of the culprit provided by the witness (match-to-description) or if they are chosen to match the police suspect (match-to-suspect). The match-to-suspect method means that the police select fillers who look similar to their suspect. Match-to-suspect methods of constructing lineups are not the preferred method of lineup construction. Luus and Wells (1991) and Navon (1992) argue that the match-to-suspect methods are impossible to assess, as there are no parameters indicating how similar the fillers should be. The match-to-suspect model also prompts the "backfire effect." Wogalter, Marwitz, and Leonard (1992) found that the suspect might stand out if he or she was the basis for choosing the fillers (match-to-suspect), as the suspect then represents the origin of the lineup and the central tendency for the group of photos.

For the match-to-description method, the police select fillers that match the description of a culprit given by an eyewitness. Wells, Rydell, and Seelau (1993) argued the superiority of match-to-description; their study found that selecting fillers based on eyewitness descriptions protected innocent suspects without negatively affecting hit rates (correctly choosing the guilty suspect) in culprit-present lineups. Luus and Wells (1991) added to these findings, claiming three strengths of the match-to-description method: the ability to specify a priori the physical features that should be shared by all lineup members, to specify a priori the physical features that lineup members should not share, and to help assure that an eyewitness both can and must use

recognition memory to perform the task. The recommendation of these researchers is that the eyewitness's description of the culprit should be used for selecting fillers for lineups.

There are, however, limitations to the match-to-description method of selection of fillers. This method may be dangerous for innocent suspects when the eyewitness description of the culprit is vague (ex: "black male"); the less detailed the description given by the eyewitness, the greater possibility of a biased lineup construction with fillers who differ substantially from the appearance of the accused (Lindsay & Webber, 1994). This problem is tied to the natural limitation of the match-to-description method: free-recall. A witness's free recall of characteristics is often vague, and the witness may only provide limited descriptors of a culprit (Lindsay & Webber, 1994). Despite its shortcomings, the match-to-description method of lineup construction is preferred over match-to-suspect.

How Eyewitnesses Make Decisions

Errors in eyewitness identification can also arise from the witnesses themselves, particularly in their decision-making processes during the lineup procedure. Dunning and Stern (1994) describe two different decision-making processes used by witnesses when selecting a lineup member as the culprit. The first is deliberative decision-making, when a witness slowly and thoughtfully comes to a decision, often expressed in the form of a process of elimination (Dunning & Stern, 1994). This form of decision-making has been shown to lead to inaccurate decisions when witnesses attempt to recognize a perpetrator (Johnson, Hashtroudi, & Lindsay, 1993). More accurate decisions come from automatic decision processes. Automatic decision-making involves the witness recognizing the perpetrator quickly without deliberating (Dunning & Stern, 1994). The eyewitness literature has consistently found a negative relationship between

latency and accuracy: the more time it takes for a witness to come to a decision, the less accurate he or she will be (Webber, Brewer, Wells, Semmler & Keast, 2004). Thus, automatic decision-making in witnesses yields more accurate lineup selections.

Similar to automatic and deliberative decision-making, witnesses also use two other forms of decision-making: absolute and relative judgment. When eyewitnesses view a simultaneous lineup (all members displayed side by side), they often use relative judgment to come to a conclusion (Wells, 1993). The eyewitness selects a person from a lineup who is more similar to memory of the culprit than any of the others in the lineup. Research on relative judgment has shown that there is an inverse relationship between the quality of the witnesses' memory and their tendency to rely on relative judgments (Wells, 1984). Relative judgment may work well if the guilty culprit is in the lineup. However, an innocent suspect may become closest to memory when the true culprit is absent.

A sequential lineup (members displayed one at a time) typically leads witnesses to use absolute judgment. Absolute judgment is a more accurate strategy, because the eyewitnesses must recognize the culprit immediately without comparing the members within a lineup (Wells, 1993). Law enforcement officials may therefore prefer sequential lineup to be used as a means to reduce eyewitness error (Klobuchar, Steblay & Caligiuri, 2006; Wells, Steblay, & Dysart, 2015).

Mock Witness Procedure

It is well established that lineup bias increases the chances of false identifications; so, there is a need to measure these biases in order to assess their presence in real lineups (Wells & Bradfield, 1999) as well as in future laboratory lineup studies. One way in which current and

future studies testing lineup fairness can be conducted is by using the mock- witness procedure. The mock-witness paradigm, in essence, is meant to test the fairness of a lineup. It is run under the assumption that if people have not witnessed the crime, nor have any knowledge of it, they will only choose the suspect from a lineup at a rate no better than chance. A mock witness is armed only with the real witness's verbal description of the culprit. If a lineup is fair, then the suspect should not be chosen more than chance by mock witnesses with no memory of the crime (Wells & Bradfield, 1999). In this study, the mock-witness procedure will be used to evaluate fairness or bias levels in actual police lineups.

Metrics. Results from the mock-witness paradigm can be assessed using multiple metrics. The primary measure used in early mock-witness procedures was, and still is, simply the proportion of mock witnesses who choose the suspect ("Proportion Score"). Using this metric, Doob and Kirshenbaum (1973) defined a biased lineup as one in which mock witnesses choose the suspect more than chance (chance being 1/n, n being the number of people in the lineup). Wells, Leippe, and Ostrom (1979) additionally advocate the use of functional size. Functional size is the number of participants in the mock-witness procedure divided by the number of participants that select the suspect. The closer the functional size is to the number of people in the lineup (nominal size), the less biased the lineup is (Wells, Leippe, and Ostrom, 1979). Take for example a lineup consisting of six people, viewed by 36 mock-witnesses. A good, unbiased lineup would have a functional size of 6, the nominal size of the lineup. Each lineup member should only be picked 1/6 of the time. Therefore, if 6 participants picked the suspect, then the functional size would be 36/6=6.

Sometimes, however, the suspect may be picked less than 1/6 of the time. This would yield a functional size larger than 6. Take for example a lineup consisting of six members, viewed by 36 people. If only one person selected the suspect, then the functional size would be 36/1=36. This gets confusing because the closer the functional size is to the number of people in the lineup, in this case six, the less biased it is (Wells, Leippe, and Ostrom, 1979). A lineup of 36 would far exceed the nominal size of the lineup, but it is also an unrealistic number in practice. In order to reduce confusion in this study, we will cap the measurement of functional size at six, as there will be six members in the lineup.

Another metric calculated from a mock-witness procedure is the effective size. The effective size of a lineup represents a comparison of the proportion each member was picked to chance (Malpass, 1981). The members that do not meet chance level, or nominal chance expectation, are subtracted from the nominal size of the lineup, thus yielding an effective size (Malpass, 1981). Malpass defines nominal chance expectation as 1/N(n), where N is the number of people in the lineup (nominal size), and n is the total number of mock witnesses. Take for example a six-person lineup with 120 mock witnesses. The nominal chance expectation equals 1/6 (120)=20. Thus, following the mock-witness procedure, any member of the lineup identified less than 20 times does not fulfill expectation and is then less effective within the lineup. The smaller the effective size of a lineup, the greater degree to which members of the lineup fail to reach nominal chance expectation.

Additionally, mean rank can be calculated as a bias indicator in the mock-witness procedure. The mean rank of a lineup averages the rankings the suspect receives from the each participant in the mock-witness procedure. In a fair lineup, the suspect would receive a mean

rank of 3.5. The lower the ranking is than 3.5 the more biased the lineup. Finally another metric that is used as a bias indicator is the amount of mock-witnesses who reject no lineup members. During the mock-witness procedure, participants are asked to indicate if any lineup member could not be the suspect. In a fair lineup the mock-witness would not select anyone, as every lineup member should be similar in appearance to the suspect. Thus, fair lineups should have higher percentages of participants rejecting no lineup members than biased lineups.

Questions to ask mock witnesses. When conducting a mock-witness procedure, the experimenter must consider which question to ask the mock witness. Just as instructions to real witnesses can influence decision-making, so can instructions to mock witnesses. It has been debated whether "which person is the accused?" or "which person matches the description?" is the superior question to ask a mock witness. Wells and Bradfield (1999) tested to see which of these two questions is better by running a mock-witness procedure to compare outcomes from the two different questions. They concluded that the question asked of mock witnesses should be "which person is the accused?" This question yielded a smaller functional size (1.6) than did the question "which person was the witness describing?" (3.0.), indicating a sharper detection of biased lineups. "Which person is the accused?" allows the eyewitness to use other information, such as whether the suspect stood out to them for other reasons (defining feature, tattoo, etc.), beyond the description given.

Importance of Witness Descriptions

A culprit description given by an eyewitness is the basis for lineup construction. Fillers are selected based on the description by the witness, and those fillers are intended to protect the suspect by possessing the same attributes. However, the eyewitness literature has shown that

eyewitness descriptions are often vague, leaving out aspects about the culprit (Lindsay, Martin & Webber, 1994). If the witnesses leave out details about the culprit when prompted for a description, it is possible that only one member of the subsequent lineup has the feature(s) the witnesses remembered but failed to mention. Thus, the suspect becomes noticeably different from other lineup members, and the lineup and does not protect the suspect from weak or unreliable witness memory. Even if the suspect is guilty, one of the functions of a lineup is to be fair, protecting the suspect from standing out. Biases must be prevented within lineup construction, and the descriptions given by eyewitnesses can be a good place to prevent biases.

Issues of brief descriptions. The descriptions given by witnesses will be used in police investigations, used to create lineups, and used to test the witness's memory of the culprit. A brief description can make it easy to create a lineup, but difficult to create a lineup matching for features not mentioned. Take for example a description of "black male." The police can choose any six black men off the street and have them "match" this description. However, the chosen six black men may look substantially different, as they will likely differ in age, hair, skin tone, facial features, height, weight, and other features beyond gender and race. The witness may be able to select the suspect only because the fillers are quite divergent from the features the witness is expecting. **Default values.** Eyewitnesses give descriptions of the perpetrator of the crime to aid the investigation. The police then create a lineup and add fillers. However, problems with eyewitness descriptions involve the limitations of default values and free recall. Descriptors not mentioned at the time the police acquire a description can alter the fairness of a lineup. Lindsay, Martin & Webber (1994) explain that a failure to mention a feature during free recall may indicate "that there is a *default value* rather than that the witness

failed to remember the information." (p. 529). These default values are descriptors of the perpetrator the eyewitness has not mentioned or did not think to mention, perhaps assuming the police already know. The police may fail to ask for clarification because they share the same expectations as the eyewitness.

Default values are a problem because they can lead to the creation of an unfair lineup, even though the police match the verbal description given by the witness. Take for example a witness who is asked by the police to describe the culprit. The witness may not know what features are important to mention, so the witness gives a vague description of "black male", even though he or she knows more defining features. The police then create a lineup of six black males, a seemingly fair lineup as all match the description of "black male." However, when the witness is shown the lineup, he or she may notice features previously not mentioned (such as age, height, build, weight, etc.) in only one of the suspects. Thus, these default values have created an unfair lineup, because not all of the lineup members match the features that the eyewitness remembers.

Free recall Vs. leading questions.

When investigating a case, police ask questions of the witness to describe the culprit, and the witness uses free recall to answer. Free recall is often accurate, because it simply requires the witness to recall what he or she remembers (Lindsay, Ross, Read & Toglia, 2013). However, free recall may provide an unsatisfactory or incomplete description. Police might ask leading questions to solicit more descriptors from the eyewitness. However, if police ask leading questions, they may taint the information received from the eyewitness. Leading questions are specific questions asked by police (i.e. Did the man have long red hair with side burns?) that point to who they believe the suspect is. This form of questioning can influence the eyewitness's

decision and consequently lead to an eyewitness selection of an innocent suspect. In order to avoid leading questions and enhance free recall, police must find a guideline of features to solicit accurate information from the eyewitness by asking specific, closed ended but not leading questions, such as “Do you remember the hair color? Weight? Height?” What is still left, however, is the question of what features police should prompt eyewitnesses for, a central issue of this study.

Lindsay, et al, (1994) speculated more than two decades ago that lineup construction would produce greater lineup fairness if a framework of descriptors was asked of each witness, without leading questions. The witness can be prompted for features he or she might not otherwise mention. Past research on eyewitness descriptions lends direction to determine what descriptors should be prompted of eyewitnesses, Lindsay, Martin, and Webber (1994) found that mock witnesses (without prompting) described hair color, height, and clothing most frequently, followed by race, sex, and age. Real witnesses to crimes (descriptors of 105 criminals published in a local newspaper in Kingston, Ontario) described hair color (38.1%), height (46.6%), and clothing (60%) the most (Lindsay, Martin, & Webber, 1994). Both the laboratory witnesses and real witnesses were least likely to describe the face and size of the suspect, and race, sex, and age were mentioned by less than half of the mock witnesses. Frequency of descriptors aside, both studies had the same basic framework of descriptors mentioned: hair, height, face, weight, race, sex, and age. If the witness is asked for these features during the lineup construction, then he or she is likely to produce a more complete description of the culprit than through free recall. Lindsay, Ross, Read, and Toglia (2013) revealed a similar pattern of features and frequencies. The study involved 100 police protocols in Seattle, WA. In descending order of frequency, the

following features were described by the witnesses: gender, age, height, build, race, weight, complexion, and hair color. While there were low frequencies for some of the features mentioned, they were mentioned in both studies (Lindsay, Martin, & Webber, 1994; Lindsay, Ross, Read, & Toglia, 2013); therefore, a pattern of common descriptors arises. The following framework of features should be prompted from the raters in our study: race, sex, age, size (height, weight, build), hair, and face. For simplicity, height, build, and weight will be grouped together and be named "size", and facial features will be grouped into a "face" descriptor.

Current Study

Our study is interested in testing the speculation of Lindsay et al (1994), that a problem of default values (offender attributes not mentioned by witnesses) can be turned into a positive recommendation for lineup construction policy. That is, the problem of vague eyewitness descriptions can be handled as a system variable: a police protocol that will prompt eyewitnesses to report default values and thereby aid the police in construction of a fair lineup. We begin by examining a set of lineups that have been previously tested for lineup construction bias. All have been judged as "fair" through a mock-witness procedure. However, each lineup also only involved a vague witness description of the culprit (e.g., 'black male'). Police can easily structure what appears to be a fair lineup in that all members meet this vague description. Yet, it is possible that standard metrics of lineup fairness (e.g., functional size) will not detect lineup bias because lineup members may not match basic details the witness remembers about the culprit.

Our study aims to add knowledge of biases in lineup construction, particularly how the vagueness of eyewitness descriptions can lead to problems with creating fair lineups. Specifically, we are looking at lineups collected by Wells, Steblay, and Dysart (2015), from four police departments (Austin, TX, San Diego, CA, Tucson, AZ, and Charlotte, NC). We are interested in a subset of 494 of the lineups that were run in a double blind procedure (detectives not aware of the suspect's identity). Bias metrics were calculated using the mock-witness procedure for 120 of the lineups (Steblay & Wells, 2014).

The current study relies on bias metrics calculated as part of the NSF grant research. Thus far, 60% of the lineups are fair (based on the calculated metrics), and 40% are biased (Steblay & Wells, 2014). However, we believe that many of the lineups judged to be fair may not actually be fair; many of the fair lineups only have 2-4 descriptors (ex. Black male), and it is likely that there were descriptors left out by the eyewitness. The mock-witness procedure cannot function properly if these default values go unmentioned by the witnesses. A real witness potentially has more information about the suspect than the mock witness, details simply not mentioned. A mock witness has no way of knowing these left out details; so, the mock witness armed with a description as vague as "black male", and having never seen the culprit is essentially choosing a lineup member at random. Thus, the bias metrics may falsely indicate that a lineup with few descriptors is fair, when in reality the default values are not matched among the fillers.

We predict that lineups that have previously undergone the mock-witness procedure and have been judged to be fair, but only have 2-4 descriptors, will reveal more bias when a framework of descriptors of the suspect are added. The prediction is that lineup fairness metrics will be (artificially) inflated when based on a vague description compared to a description based

on a set of core attributes (called "framework descriptors"). To test this hypothesis, this study will be comprised of two phases. In the first phase, 20 lineups were selected that have undergone the mock-witness procedure, have a suspect picked less than 1/6 of the time, and no more than four descriptors in the witness's description. Then, a framework of descriptors to describe the suspects was determined. Based on past research, we asked the participants to describe the race, sex, age, size (height, weight, build), hair, and face are of all 20 suspects. From this data, we will synthesize the descriptions given by participants.

In Phase 2, we will replace the original description of the suspect with the newly synthesized, more complete description to produce our experimental manipulation. Thus, there will be 40 lineups in total: 20 with a new *framework descriptions*, and the same 20 with the old *original* descriptions. The subsequent study will be a two-group within-subjects design, using the mock-witness procedure. The independent variable is the description provided with a lineup: either the original description or the new framework description. Each mock witness will view all 20 lineups. For each witness, half of the lineups will include the new framework descriptions, half will include old original descriptions. The order of the lineups and assignment of lineup descriptions to lineups will be counter-balanced. The dependent measures are the indices of lineup bias. We predict that the lineups with the new framework descriptors will be more biased (e.g., a higher Proportion Score) than the old original lineups.

Method: Phase 1

Participants

For phase 1, 20 participants from Augsburg University psychology classes partook in the study, selected through the psychology department participant pool. The participants self-

reported as 8 females and 12 males with a mean age of 18.84 years. Self-reported race was as follows: Black (20%), White (45%), Hispanic (25%), and mixed race (10%). Participants were compensated with one research participation credit for their psychology (105 or 215) class.

Materials

The experimenters chose 20 real lineups that had undergone bias testing in an earlier study (NSF grant). The lineups were picked for this phase of the research if the reported description was vague (2-4 descriptors, mix of age, race, height, sex, or hair) and the lineup was judged to be fair (mock witnesses picked the suspect 1/6 or less of the time). Then, the experimenters pulled out the photos of the suspect in each lineup, and placed the photos in a computer file (power point). Two power points were created ("A" and "B"), with reversed order of the pictures on the slides as a counterbalance for sequencing effects. Participants were randomly assigned to either "A" or "B" and viewed the power point displayed on a computer. The pictures were displayed one at a time on the computer, and the participants had an unlimited amount of time to look at each individual photo. Participants were given a response sheet with the number of the lineup at the top, followed by a series of six questions asking them to describe the photo: age, race, sex, size (weight/height/build), hair, and face. There were two different response sheets; one corresponded with power point "A" and the other corresponded with power point "B."

Procedure

Participants were briefed about the procedure, the experimenters read directly from a script, and the participants were asked to read and sign a consent form. They were taken into a room with a computer and given their response sheet. The experimenter left the room and

allowed the participant to fill out the response sheet. When the participants completed the task, the experimenter debriefed the participants and thanked them for their participation.

Method: Phase 2

Participants

For phase 2, 73 participants partook in the study and were recruited from the Augsburg University Psychology department participant pool. There were 29 self-reported females and 19 self-reported males with a mean age of 19.5 years (*SD*=2.71). Self-reported race was as follows: Black (18.8%), White (54.2%), Hispanic (10.4%), and mixed race (4.2%). Twenty-five participants did not record gender, age, or race in this study. Participants were compensated with one research participation credit for their psychology (105 or 215) class.

Materials

First, to create the new framework descriptions, the experimenters synthesized the descriptions given by participants in phase one by marking the number of times a feature was mentioned by participants. In order to synthesize the descriptions, we created a set of rules to be consistent on which features were included in the framework descriptions. Age ranges were set to a 10 year age range, unless a smaller age range was indicated by more than 10 (50%) of participants. Weight and height ranges were set by 10 or more occurrences. Race was determined if there were 10 or more occurrences. If no consensus was made on race (i.e. no more than 10 occurrences of one race), then experimenters included the highest occurring race; however, if two races added up to more than 10 occurrences together (i.e. 6 occurrences of Hispanic, 7 occurrences of White, 3 occurrences of Hispanic/White) then both races were added to the framework description. Sex of the suspect was unanimous for every lineup. Every other feature

(face, build, hair) prompted was included in the framework description if it occurred 7 or more times (35% of participants).

Using the newly synthesized descriptions from phase one, the experimenters replaced the old descriptions with the new framework descriptions, creating a second condition. In phase 2, the experimenters tested both conditions (old and new framework) for the 20 lineups. Then, they created four forms to create a within subject study design. Each form contained 20 lineups: 10 were the lineups with the original description, and 10 were the lineups with the new framework descriptions. The position of lineups in the forms and the old/new descriptions were randomly assigned. The forms counterbalanced the order of the lineups in the form and the type of description used in the lineup (old or new).

The forms were created on Google Forms to run the mock-witness procedure, and were administered on computers in Augsburg University's Psychology Labs.

Participants were asked to rank who they believed was most likely to be the suspect on a scale from 1 (most likely) to 6 (least likely). The participants could not mark two members with the same ranking. The rankings provided by the mock witnesses were automatically added to a Google Spreadsheet, and these rankings were used to calculate bias metrics. Additionally, participants were asked to check boxes for any lineup member who they believed could not be the suspect. After the data were collected, one lineup was dropped from the study. This lineup was dropped because the wrong photo was selected as the suspect in phase 1. Thus, the experimenters examined 19 lineups in their analyses.

Procedure

Participants were randomly assigned to one of the four forms upon arrival in the lab. Participants were briefed about the procedure, and the experimenter explained directions for the mock-witness procedure using a script. Participants were asked to read and sign a consent form. Then, participants were taken into a room with a computer and shown an example of how to rank the lineup members. The experimenter then left the room and allowed the participants to complete the mock-witness task. When the participants completed the task, the experimenter debriefed them and thanked them for their participation.

Results

This first level of analysis was taken directly from participant responses. Each of the 19 lineups, with original or new framework descriptors, was viewed by 36 or 37 participants. The uneven number of participants was counterbalanced across lineups.

The first step in analysis was to place participant rankings for each lineup member in each lineup into a spreadsheet by condition (original description versus new framework description). This allowed us to calculate the frequency of mock-witness 1st- picks for each of the six lineup members in each of the 19 lineups, separately for original and framework descriptions. Additionally, we obtained frequencies for the percentage of mock-witnesses who rejected lineup members, and we were able to calculate mean rank scores for the suspect.

For each individual lineup the proportion score was the proportion of participants who ranked the suspect with a one (most likely to be the suspect). The functional size, and the functional size capped at six, was obtained by dividing the the total number of participants who viewed the lineup by the number of participants who ranked the suspect first. Effective size was determined by recording the frequencies of participant choices (ranking of one) for each lineup

member and entering the rankings into an online calculator. Mean rank was obtained by averaging the ranks the suspect received. The percentage of participants who did not reject any lineup member was assessed by frequency of mock-witnesses with no checked boxes for lineup members who could not be the suspect.

The second step of the analyses was to summarize these measures across lineups, calculate average proportion score, functional size, functional size capped at six, effective size, mean rank, and percentage of mock-witnesses who rejected no lineup members. These fairness measures were calculated across the 19 lineups, separately for the original and new framework descriptions.

The final step of the analysis was to transfer the aggregate fairness measures into an SPSS file in which the level of analysis was the lineup, each lineup with two conditions: an original description or a new framework description. The analyses compared the fairness measures obtained from the original descriptions to the fairness measures from the new framework descriptions using two-tailed *t* tests with alpha set at .05.

Proportion score. The proportion score of a lineup is simply the proportion of mock witnesses who choose the suspect. In a fair lineup containing six members, a suspect would have a proportion score of .167, thus being picked 1/6 of the time. As predicted, the proportion scores for the new framework descriptors was higher (M=.43, SD=.19) than the original descriptions (M=.11, SD=.07), t (36)=6.89, p<.0001, d=2.30. In the framework description lineups, the suspect was selected more than 1/6 (.167) of the time, indicating that the lineups reveal bias when framework descriptors are employed.

Functional size. Functional size is the number of participants in the mock-witness procedure divided by the number of participants that select the suspect. The closer the functional size is to the number of people in the lineup (6), the less biased it is, and the closer it is to 1 the more biased it is. As predicted, the functional size of the new framework descriptors was lower (M=3.04, SD=1.95) than the functional size of the original descriptions (M=14.78, SD=11.90), $t(36)=4.25$, $p<.0001$, d=1.42. The lineups with the new framework descriptors had functional sizes far less than 6, indicating that the lineups are unfair. In some cases, functional size exceeded six and so the experimenters also calculated functional size scores by capping the scores at six. With the cap of six, framework descriptor scores were still lower (M=2.87, SD=1.49) than the original descriptions (M=5.78, SD=.54), $t(36)=8.01$, $p<.0001$, d=2.67.

Effective size. The effective size of a lineup represents a comparison of the proportion each member was picked to chance. The members that do not meet chance level, or nominal chance expectation, are subtracted from the nominal size of the lineup, thus yielding an effective size. The smaller the effective size of a lineup, the greater degree to which members of the lineup fail to reach nominal chance expectation. As predicted, effective sizes were lower in the new framework descriptors (M=3.38, SD=1.20) than the original descriptions (M=4.07, SD=.91), $t(36)=2.01$, $p<.052$, d=.67. The lineup members in the new framework descriptions were less effective at protecting the suspect than the lineup members in the original description lineups.

Mean rank. The mean rank of a lineup averages the rankings the suspect receives from the mock-witness procedure. In a fair lineup, the suspect would receive a mean rank of 3.5. The lower the ranking is than 3.5 the more biased the lineup. As predicted, the suspects in the new framework descriptors received lower mean ranks (M=2.49, SD=.57) than suspects in the

original descriptions (M=3.8, SD=.60), $t(36)$=7.00, p<.0001, d=2.33. Suspects in the new framework lineups received mean ranks that indicated biased lineups.

Rejected lineup members. Additionally, we asked participants to indicate if there were any members of the lineup that could not be the suspect at all. In a fair lineup, no member would be rejected because they all should fit the description. We calculated the percentages of lineups in which no lineup member was rejected. As predicted, more participants rejected lineup members in the new framework lineups (M=.46, SD=.15) than in the original description lineups (M=.62, SD=.14), $t(36)$=3.20, p<.003, d=1.07. Fewer participants rejected no lineup members in the new framework descriptions than in the old descriptions.

Table 1 Bias Indicator Statistics

	Lineup Condition	Mean	Standard Deviation	Effect Size
Proportion Score	Original	.11	.07	2.30
	New	.43	.19	
Functional Size	Original	14.78	11.90	1.42
	New	3.04	1.95	
Functional Size of 6	Original	5.78	.54	2.67
	New	2.87	1.49	
Effective Size	Original	4.07	.91	.67
	New	3.38	1.20	
Mean Rank	Original	3.81	.60	2.33
	New	2.49	.57	
Rejected 0	Original	.62	.14	1.07
	New	.46	.15	

Discussion

The results indicate that lineups that may appear fair when undergoing the mock witness procedure can contain biases that are revealed when tested with more detailed descriptions. Our hypothesis was supported; the new framework descriptors, which were detailed, yielded higher bias measures than the original descriptions, which were vague. In the original lineups, the suspect was picked on average 11% of the time. In the new, framework descriptions, the suspect was picked 43% of the time.

We believe that these results occurred because the original descriptions are vague, and do not contain enough descriptors to match every lineup member to the suspect. If the police receive a vague description of "black male," it is likely that the lineup members will not be matched to other features such as hair, size, face, or age. Then, if the witness notices only one member with long hair, and remembers that the suspect had long hair, the lineup is biased because not every member has long hair. This phenomenon was the basis for our new framework descriptions, which flush out characteristics not mentioned in the original description. Using a framework of descriptors, features such as age, race, sex, size (height, weight, build), hair, and face were obtained for each suspect. Our overall results indicate that with these added features, suspects no longer matched the other members of the lineup, thus standing out and rendering the lineup unfair.

Analysis of individual lineups. There were three lineups in which the new framework description seemed to make little difference. The suspect was picked 22% of the time in one lineup, 16% in another, and 11% in another lineup. In these lineups, the foils the police chose effectively matched, and therefore protected, the essuspect.

Two lineups were found in which the suspect did not seem to match the original witness description at all. In one lineup the witnesses indicated that the culprit had blonde hair. However, the suspect the police chose had brown hair, as did most of the lineup foils. In the other lineup, the witness had described the culprit as stocky. The suspect that the police chose was instead skinny. These two lineups demonstrate a failure of the police to match the suspect to the description given by the witness, the first and most important step in lineup construction.

A set of individual descriptors seemed to really make the suspect stand out compared to the lineup foils, indicating that soliciting these features from the witness is essential to create a fair lineup. In many cases, the "face" descriptions seemed to substantially set the suspect apart from the foils. In one lineup, the suspect was picked by 65% of the participants in the framework description lineup, and was only picked by 3% of the participants in the original description lineup. We believe that this substantial increase in proportions picking the suspect could be due to the addition of "mole under the eye" in the new framework description. When looking at the lineup, only the suspect has a mole under his eye, making him stand out noticeably from the other lineup members.

Another feature that seemed to make the suspects stand out was the addition of "size" features. For example, "small, slim build" was added to the new framework description in one lineup and the proportion of participants selecting the suspect rose from 14% in the original description lineup to 51% in the new framework description lineup. The addition of "small, slim build" easily eliminates the majority of the lineup members, resulting in the suspect standing out.

In one lineup in particular, we yielded unusual results when the mock-witnesses described the race of the suspect. In the original description, the suspect was described as

"Hispanic", and he was picked by only 8% of the mock-witnesses. However, during phase one the majority of the participants indicated that this suspect's race was "White" and not Hispanic. Using the race as "White" in the description, the suspect was picked 81% of the time. This brings up a potentially important addition to the framework descriptions: complexion. The suspect had a very light complexion compared to the foils, making him look like a different race. It is hard to say that the addition of complexion would have changed the mock-witness's indication of race, but it could be important for real witnesses describing a culprit.

One potential issue with our results is the accuracy and plausibility of the application of our framework descriptors in the field; would we get these default values created from the framework descriptors if they were asked of witnesses in the field? It is hard to determine what witnesses may or may not notice when they come into contact with a perpetrator. For example, the majority of mock-witnesses describing the suspect in lineup 706 indicated that he had a mole under his eye. How plausible is it that a real witness would notice this mole when being exposed to the perpetrator? It is a lot easier for the mock-witnesses to describe in detail all the framework descriptors asked of them because they have an unlimited amount of time to study the suspect. Real witnesses, however, do not have this opportunity. Thus, it is very likely that real witnesses would not be able to create as complete and detailed descriptions as the mock-witnesses in our study. However, for other features such as hair, size, height, weight, build, and age, we believe it is very reasonable for witnesses to be able to provide at least some of these features. Asking these features, rather than simply being asked to "describe the perpetrator", gives witnesses a guideline to understand what police are looking for while helping them to jog their memory.

Future research. Future research on eyewitness descriptions, and research expanding upon the current study, can take a few different directions. One direction could be to better replicate a real witness experience when describing a suspect. Our participants had an unlimited time to view and describe the suspect, a luxury real witnesses lack. Thus, a similar study could be conducted, but by altering the phase 1 study. Instead of allowing mock-witnesses an unlimited time to view the suspect, only allow them 10 seconds and ask them to describe the suspect after the time is up. Another direction that research could take is to expand upon the "face" category in the framework descriptions. Mock-witnesses were often confused by the "face" prompt because it was vague and open-ended. So, for future research, the face category could be expanded to include "facial hair" and "distinguishing marks." In one lineup, the suspect had a distinguishing mark (a giant mole on his chin) that set him apart from the other foils. The mole he had was very noticeable, and because of how big it was it would be very likely a witness may notice it during a crime. Additionally, facial hair set the suspect apart from foils in many lineups. The new framework description included "beard" under the face category, and the suspect was the only one with a beard in the lineup.

Implications for policy and practice. Based on our findings, there are several recommendations that can be made for police procedures. First and foremost, the police should use a framework of features that they ask the witness to describe. Instead of simply asking to "describe the perpetrator", the police should ask the witness to " describe the age, sex, race, build, weight, size, height, hair, and face" of the perpetrator. Then, the police should create a lineup matching the description of the perpetrator. After the lineup is created, the police should then check for match to suspect to ensure that the suspect does not stand out. These

recommendations could help police practices by reducing the selection rates of innocent suspects. If police gather more detailed descriptions from the start, they can build better lineups to protect innocents from being incorrectly selected.

References

Doob, A. N., & Kirshenbaum, H. M. (1973). Bias in police lineups-Partial remembering. *Journal of Police Science and Administration, 1, 287-293.*

Dunning, D., & Stern, L. B. (1994). Distinguishing accurate from inaccurate eyewitness identification via inquiries about decision processes. *Journal of Personality and Social Psychology, 67, 818-835*

Findley, K. A. (2013). The Innocent Project. In J. A. Siegel, Saukko, P. J., & Houck, M. M. (Ed.), *Encyclopedia of Forensic Sciences* (pp. 449-453). Elsevier Ltd. . doi:http://doi.org/10.1016/B978-0-12-382165-2.00222-1.

Garrett, B. L. (2011). *Convicting the innocent: Where criminal prosecutions go wrong*. Cambridge, MA: Harvard University Press. .

Innocence and the Death Penalty | Capital Punishment in Context. (n.d.). Retrieved September 16, 2016, from http://www.capitalpunishmentinconte xt.org/issues/innocence

Klobuchar, A., N. Steblay, & H.L. Caligiuri (2006), "Improving Eyewitness Identifications: Hennipin County's Blind Sequential Lineup Pilot Project," *Cardozo Public Law, Policy, and Ethics Journal*

Johnson, M. K.., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114,* 3-28.

Lindsay, R. C., Ross, D. F., Read, J. D., & Toglia, M. P. (2013). *The Handbook of Eyewitness Psychology Memory for People* (Vol. 2). London: Taylor and Francis.

Lindsay, R. C., Martin, R., & Webber, L. (1994). Default Values in Eyewitness Descriptions: A Problem for the Match-to-Description Lineup Foil Selection Strategy. In R. Martin (Ed.), *Law and Human Behavior* (5th ed., Vol. 18, pp. 527-541). Plenum Publishing Corporation.

Lindsay, R. C. L. & Wells, G. L. (1980). What price justice? Exploring the relationship between lineup fairness and identification accuracy. Law and Human Behavior, 4, 303±314.

Luss, C. A. E., & Wells, G.L. (1991) Eyewitness identification and the selection of distractors for lineups. *Law and Human Behavior, 15,* 43-57.

Malpass, R. S. (1981). Effective size and defendant bias in eyewitness identification lineups. *Law and Human Behavior,5*(4), 299-309. doi:10.1007/bf01044945

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

Steblay, N. K. (2015). Scientific Advances in Eyewitness Identification Evidence. In *William Mitchell Law Review* (3rd ed., Vol. 42, 2015).

Steblay, N. K., Wells, G. L. (2014). Collaborative Research: Understanding and Predicting Eyewitness Identification Errors: Studies Using a Unique Set of Materials from Actual Lineups, NSF grant proposal #

Weber, N., Brewer, N., Wells, G. L., Semmler, C., & Keast, A. (2004). Eyewitness identification and response latency: The unruly 10-12 second rule. Journal of Experimental Psychology: Applied, 1 139-147.

Wells, G. L. (1978). Applied eyewitness-testimony research: System variables and estimator variables. In Journal of Personality and Social Psychology, Vol 36(12), Dec 1978, 1546-1557.http://dx.doi.org/10.1037/0022-3514.36.12.1546

Wells, G. L. (1984), The Psychology of Lineup Identifications. Journal of Applied Social Psychology, 14: 89–103. doi: 10.1111/j.1559-1816.1984.tb02223.x

Wells, G. L. (1993). What do we know about eyewitness identification? *American Psychologist,48*(5), 553-571. doi:10.1037//0003-066x.48.5.553

Wells, G. L. and Bradfield, A. L. (1998). `Good, you identified the suspect': Feedback to eyewitnesses distorts their reports of the witnessing experience. Journal of Applied Psychology, 83, 360±376.

Wells, G. L., & Bradfield, A. L. (1999). Measuring the Goodness of Lineups: Parameter Estimation, Question Effects, and Limits to the Mock Witness Paradigm. In *Applied Cognitive Psychology* (Vol. 13, S27-S39). John Wiley & Sons.

Wells, G. L., Memon, A., & Penrod, S. D. (2006). Eyewitness Evidence: Improving Its Probative Value. *Psychological Science in the Public Interest, 7*(2), 45-75. doi:10.1111/j.1529-1006.2006.00027.x

Wells, G. L., Leippe, M. R., & Ostrom, T. M. (1979). Guidelines for empirically assessing the fairness of a lineup. *Law and Human Behavior,3*(4), 285-293. doi:10.1007/bf01039807

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

Wells, G. L. Turtle, J. W. (1986). Eyewitness identification: The importance of lineup

models. Psychological Bulletin, Vol 99(3), May 1986, 320-329.http://dx.doi.org/10.1037/0033-2909.99.3.320

Wells, G. L., Steblay, N. K., & Dysart, J. E. (2015). Double-blind photo lineups using actual eyewitnesses: An experimental test of a sequential versus simultaneous lineup procedure. *Law and Human Behavior,39*(1), 1-14. doi:10.1037/lhb0000096

Wogalter, M.S., Marwitz, D.B., & Leonard, D.C. (1992). Suggestiveness in photospread lineups: Similarity induces distinctiveness. *Applied Cognitive Psychology, 6,* 443-453.