Spring 5-25-2023

An Archival Exploration of Lineup Fairness in Eyewitness Research

Phoebe Kane
CUNY John Jay College, kane.phoebe.f@gmail.com

An Archival Exploration of Lineup Fairness in Eyewitness Research

A Thesis Presented in Partial Fulfillment of the Requirements for the Degree of

Master of Arts in Forensic Psychology

John Jay College of Criminal Justice

City University of New York

Phoebe Kane

May 2023

An Archival Exploration of Lineup Fairness in Eyewitness Research

Phoebe Kane

Thesis Committee

Thesis Advisor: Steven Penrod

Second Reader: Margaret Bull Kovera

External Reader: Jamal K. Mansour

Table of Contents

Acknowledgments

I would like to express my sincere gratitude to my advisor, Dr. Steve Penrod, for his invaluable guidance, abundance of knowledge, and belief in me during this process. His expertise and mentorship encouraged my exponential academic growth during my time at John Jay, and paved the way for this project's success.

Special thanks to Dr. Margaret Kovera and Dr. Jamal Mansour for serving on my thesis committee. Your guidance and feedback were instrumental to the formation of this project. I'm very appreciative to have this wonderful team of mentors.

My deepest thanks go to my friends and family for their love and support during this journey. I couldn't have finished this process without their understanding and encouragement. Thank you to my partner, Jullia, and our cat, KC, for brightening my life outside of academia.

Thank you to the members of the Penrod and Kovera labs. For the past two years, they have been a major part of my support system, have served as a sounding board for my academic and professional ideas, have given me grace to try and to fail – and then helped me back on my feet. I'm so grateful for this community.

Major thanks to the researchers and data contributors who kindly granted us access to their work. You made this project possible. Thank you for your time and contribution.

Melisa Akan
Christopher Altman
Andrea Arndorfer
Stephen Badham
Mario Baldassari
Jennifer Beaudry
Aaron Benjamin
Neil Brewer
Curt Carlson
Steve Charman
Melissa Colloff
Laura Crane
Mitch Eisen
William Erickson
Kimberly Fenn
Jason Finley
Ryan Fitzgerald
Roy Groncki
Alistair Harvey
Lucy Henry
Clay Holroyd
Ruth Horry
Shaela Jalava
Nate Kornell
Margaret Kovera
Andrew Lampinen
Michael Leippe
Stephen Lindsay
Carmen Lucas
Simona Mackovichova
Jamal Mansour
Terence McElvaney
Laura Mickes
Chris Oriet
Matthew Palmer
Kathy Pezdek
Maria Robinson
Henry Roediger
James Sauer
Chelsea Sheahan
Michelle Stepan
Deryn Strange
Laura Smalarz
Andrew Smith
Jim Tanaka
Colin Tredoux
Kalif Vaughn
Kimberley Wade
Rachel Wilcock
John Wixted
Alex Wooten

Abstract

In this study, we were interested in investigating if the Betaface facial analysis program reliably predicts eyewitness lineup choosing behavior. If face analysis programs are as good or better than human judgements, using them could be a reliably more efficient, reproducible, and equitable basis for choosing fillers and evaluating lineup fairness. We collected 27 datasets from eyewitness researchers and analyzed them to produce Betaface similarity values, which measured the similarity between all the photos in each array. We compared these Betaface data to the identification data from the original studies. Our analysis of the arrays via Betaface yielded data with a fairly high degree of GT-IT, GT-filler, and IT-filler similarity across arrays, which implies that the arrays are quite fair. There is no evidence to show that Betaface can reliably predict identification choosing behavior. To find a clearer relationship in Betaface values and identification rates, we would require data from studies that are attentive to systematically manipulating similarities in the selection of the fillers and IT. Manipulating these variables independently would yield non-correlated measures; without these manipulations, the lineup construction variables in the current dataset display too little variability to permit detection of possible Betaface-identification relationships.

Introduction

Eyewitness identification is the process of viewing either one person or an array of multiple people for the purpose of identifying a suspect in a criminal case. Lineups, or arrays, are two names for the sample of people that the eyewitness may make an identification from. There are two different types of arrays: target-present (TP) and target-absent (TA). A TP array is a lineup in which the "guilty" target (GT), who the eyewitness has seen before, is shown alongside a number of fillers (the individuals in the array who are known to be innocent). A TA array may be one of two compositions: a lineup in which there are only fillers; or a lineup in which there is an "innocent" target (IT), who the eyewitness has never seen but who bears a resemblance to the "guilty" target, shown alongside fillers. Other than these basic definitions, there are few requirements for lineup assembly or presenting a lineup to an eyewitness. As the lineup construction process and the eyewitness identification process have varying and nonspecific guidelines, research exploring the intricacies of eyewitness identification is crucial to promote fairness and effectiveness in all steps of the process.

Eyewitness Misidentification

Misidentifications from eyewitness lineups are a consequential and dangerous issue in our justice system. Though eyewitness identifications can be valuable evidence, there are many cases of misidentification that lead to wrongful convictions. Eyewitness misidentification is responsible for more false convictions than any other factor (Innocence Project, 2022; Wells et al., 1998). The Innocence Project discloses that eyewitness misidentification is a factor in approximately 69% of cases repealed with post-conviction DNA evidence, and the National Registry of Exonerations lists eyewitness misidentification as a contributing factor in 28% of all

exoneration cases listed from 1989 to 2022 (Innocence Project, 2022; The National Registry of Exonerations, 2022).

Misidentifications could be decreased if lineups shown to eyewitnesses were less suggestive towards a specific target. Suggestive lineups typically refer to the phenomenon in which a suspect stands out from the lineup's fillers so that an eyewitness will be more inclined to identify the suspect. Academic researchers have explored the influence of lineup fairness on eyewitness performance and have sought an optimal level of lineup fairness to maximize the utility of eyewitness identifications (Colloff et al., 2021; Lucas & Brewer, 2022). In fact, the volume of literature on lineup fairness has grown exponentially since the 2000s. A Google Scholar search with lineup fairness or filler similarity as a keyword suggests only 89 articles were published in the 1990s while 204 articles were published in the 2000s, 378 articles in the 2010s, and 168 articles from 2020 to 2022 (Lee et al., 2023). However, suggestive arrays are still a problem in arrays constructed by law enforcement officers. Steblay and Wells (2020) found between 33% and 68% lineups used by police were suggestive in favor of the suspect, according to witness proportion scores. These biases may result in part because there is no standardized operational definition of an unnecessarily suggestive eyewitness lineup.

"Unnecessarily Suggestive" Lineups

Suggestive lineup construction encourages eyewitnesses to identify the target(s) by choosing faces so that the suspect's appearance is distinct from that of the fillers. This can have several unwanted effects, such as increasing positive identifications (or misidentifications) of the suspect, decreasing the probative value of the eyewitness testimony, and leading legal officials to conclude that an eyewitness has a better memory than they actually do. Arrays can be made more or less suggestive by altering the fillers' similarity to the target(s).

Suggestiveness typically indicates that the fillers are dissimilar enough for the target to stand out (e.g., the fillers have a different hair color than the target). However, arrays may also be unfair if the fillers are so similar that there is little ability for an eyewitness to make a distinction (e.g., the fillers and the target are siblings with similar features). There are many ways in which fillers could be suggestively dissimilar, but few scenarios that fillers would be similar enough to be suggestive. For this reason, it is much more common that suggestiveness is used to describe arrays with dissimilar fillers instead of fillers that are too similar. For the purposes of the present study, suggestiveness will address the more common phenomenon of fillers that are dissimilar to the target(s).

In *Stovall v. Denno* (1967), the Supreme Court declared that eyewitness lineups should not be unnecessarily suggestive enough to encourage an eyewitness to make a certain identification. It ruled that these unnecessarily suggestive lineups deny due process of the law (*Stovall v. Denno*, 1967). The *Stovall* case has been cited in over 2,300 federal and 4,100 state cases (the phrase "unnecessarily suggestive" was used in 2,900 and 6,200 cases respectively) according to a July 2021 LexisNexis search. Many of these *Stovall* citations were used when examining the suggestiveness of filler selections (Wells et al., 2015).

The United States legal system has attempted to improve lineup fairness by discouraging suggestive lineup practices. Former Deputy Attorney General, Sally Q. Yates wrote in a memo that a suspect should not stand out from fillers, but the fillers should also not be too similar as to reduce accurate eyewitness identifications (Yates, 2017). Legal officials have not supplied recommendations on avoiding suggestive lineups, so fair lineup construction is open for interpretation. Unfair practices may be widespread not only in law enforcement practice, but

also in psychological research. Without an operational definition of lineup suggestiveness, law enforcement and researchers cannot enforce fair and standardized lineup practices.

Lineup Construction Recommendations

In lieu of practical lineup construction guidelines from legal officials, eyewitness identification researchers have produced their own recommendations from analyses of empirical evidence. Wells et al. (1998) proposed that the following rules be enforced by the legal system:

1. Eyewitnesses should be informed that the offender may or may not be in the lineup.
2. Fillers should be similar to the suspect in general physical appearance.
3. The lineup administrator should not know which individual is the suspect of interest.
4. Confidence levels should be collected during the identification process.

Dangers of eyewitness misidentification could also be reduced if the United States Supreme Court specified best practices of lineup construction. Wells et al. (1998) called for the Supreme Court to further standardize lineups with detailed recommendations in their court rulings. Later research supported and reproduced these conclusions (Wells et al., 1998; Wells et al. 2020).

Lineup Filler Similarity

Lineup construction recommendations regarding filler similarity have not been standardized. Research on this topic has yielded divided conclusions about filler selection strategy. Depending on the study and methods, researchers may support high-, moderate-, or low-similarity fillers as the fairest option for lineups.

High-similarity fillers in lineups yield fewer identifications of targets, but the accurate identifications may be more valid, as reflected in the ratio of correct to incorrect IDs. Lindsay and Wells (1980) compared eyewitness accuracy between test lineups with either high- or low-similarity fillers. High-similarity filler arrays are less suggestive toward both guilty and innocent

targets. Compared to the low-similarity filler arrays, high-similarity filler arrays yielded fewer identifications of both guilty and innocent targets. Innocent targets were identified significantly less often from these arrays (Lindsay & Wells, 1980). Similar results were found by Fitzgerald et al. (2013) who compared identifications from lineups with low-, moderate-, and high-similarity fillers. The arrays with low-similarity fillers had the most identifications of both the guilty and innocent targets because the targets stood out (Fitzgerald et al., 2013). Though there are fewer identifications of the targets in lineups with higher filler similarity, the accurate identifications are more valid (that is, a higher proportion of suspect identifications are correct). These data supports that lineups with high-similarity fillers create more accurate eyewitness testimony.

On the other hand, high-similarity fillers may make it difficult for eyewitnesses to distinguish between the faces shown in a lineup, potentially increasing misidentifications of fillers, which might undermine the reliability of the witness. Findings from Carlson et al. (2019) supports the view that high filler similarity decreases discriminability. This study revealed that test lineups with the highest filler similarity yielded the lowest accuracy. The authors concluded that empirical discriminability decreases as the filler similarity increases (Carlson et al., 2019). Lucas and Brewer (2021) revealed similar conclusions in their exploration of the fairest level of filler similarity. This research operationalized similarity by using facial manipulation software to morph the fillers' faces to the target. The three groups of fillers were as follows: unmorphed, 33% morphed to the target, and 50% morphed to the target. Accuracy and the confidence-accuracy relationship became worse as the similarity to the target increased. Lucas and Brewer (2021) recommended that fillers should be at the lowest possible similarity levels to the target, while still matching the description of the offender. The topic of filler similarity should be further explored and standardized to yield best recommendations in practical lineup construction.

With recent advances in technology, research may turn to facial analytic software to select the best fillers for lineups.

Measuring Facial Similarity Using Betaface

Betaface is a software that measures facial similarities between photos. This software allows photos of faces to be uploaded and compared to produce a measure of similarity. Betaface can detect up to 101 facial classifications, including content like age, gender, ethnicity, hair and eye color, and face shape (Betaface, 2015). We used Betaface to produce similarity data between photos in our sampled lineups.

Each target's face can be used to generate similarity ratings with all of the filler faces. A full lineup of six or twelve photographs can be compared against each target (and other fillers). This function can be used to measure differences in features between targets and fillers. These ratings can then be tested as a potential measure of the suggestiveness of the array. If these data do reflect the similarities that drive witness choosing, then a lineup with low GT-filler similarity ratings would likely be suggestive vis-a-vis the target. Alternatively, an array with high Betaface similarity ratings between the GT and fillers would be less suggestive.

Lee et al. (2023) compared four facial recognition programs (Betaface, Azure, Amazon, and Face++) to determine if the software similarity ratings were predictive of eyewitness choosing decisions, using a set of arrays (N = 40) from their own research. Betaface was able to significantly predict filler identification rates and rejection rates, and was the only software able to significantly predict eyewitness identification choices in this study. These results suggest that Betaface may outperform the other three programs in predicting lineup choosing behavior (Lee et al., 2023). These findings are promising for Betaface as a reasonable predictor of identification behavior.

If Betaface is determined to be reliable in evaluating facial similarity, and predicting witness choosing behavior, this program may be used to assist in lineup construction. Albright & Rakoff (2020) suggest that the use of face analysis software could make it possible to select fillers for a target based on a set similarity range to the GT and variance of similarity among the fillers. Software could determine the ideal target-filler similarity, which would produce the most equitable eyewitness performance. Facial recognition software – if algorithms are shown to work well and to predict choosing behavior – could be an accurate and replicable measure of target-filler similarity. Law enforcement and researchers will be able to tell if arrays are suggestive, and could use this algorithm to select fillers from their database to match a target fairly.

Present Study

The present study uses the Betaface software to explore the suggestiveness of lineups used in eyewitness studies and to evaluate the potential of using Betaface as a measure of suggestiveness. We collected eyewitness memory datasets from past research via publicly available studies and data garnered from other researchers. This data sample included TP and TA arrays and identification decisions. Betaface produced similarity measures for these arrays, which was the proposed measure of suggestiveness.

Our research questions were as follows:

1. Are the arrays used in eyewitness research measurably fair?

   1A. To what extent, if any, are these arrays biased?

   1B. Is there a measurable difference in bias between TP and TA lineups?

2. Do Betaface similarity measures reliably predict eyewitness choosing behavior?

Our hypotheses were as follows:

H1. The lineups used in eyewitness research will be measurably biased as evidenced by Betaface similarity scores.

H2. Both TP and TA lineups will be suggestive, but TP lineups will be more suggestive than TA lineups because the GT will stand out more from the fillers than the IT will.

H3. Betaface similarity measures will reliably predict eyewitness choosing behavior.

There is a measurement gap in the literature on bias within eyewitness lineup research. This study aims to fill this research gap by reliably measuring similarity-based lineup bias in experimental arrays. These data may also reveal how the arrays are suggestive – for example, if they are significantly more suggestive in a target present array than a target absent array. Identification decision data was compared to the Betaface data as a measure of how the participants were affected by the potential suggestiveness – in other words, do Betaface similarity measures predict witnesses choosing behavior?

The suggestiveness of lineups used in research may be revealed via the results of this study. By revealing biases of arrays used in research and evaluating the effectiveness of Betaface to predict human similarity judgements, results can be used to correct flawed lineup processes and improve future research. This study aims to compare eyewitness and computer-based similarity judgments – do they produce similar results? If algorithms are as good or better than human judgments, using them could be a reliably more efficient, reproducible, and equitable basis for choosing fillers and evaluating lineup fairness. More precise and informed lineup practices in research will also lead to more accurate results and information distributed to law enforcement. As research becomes more informative, we can more effectively decrease eyewitness misidentifications in our justice system.

Method

Materials

The data used in this meta-analysis were collected from other studies on eyewitness identifications. Some datasets were gathered via public access online (using sites like the Open Science Framework, PsycInfo, ResearchGate, and other similar sources). Search results for publicly available datasets were filtered by studies that have files available online and were then checked manually to ensure that the dataset included the required data. Only seven publicly available datasets fit these requirements. The majority of the datasets were gathered by soliciting and obtaining permission from researchers involved in each study. These solicited researchers sent their datasets via email or sharing sites (Dropbox, Google Drive, and other similar sites).

Data collection began in February 2022 and concluded in March 2023. We required that datasets included identification data and both TP and TA arrays. After data collection ended, we excluded datasets based on the following criteria:

- Fewer than five fillers and more than 12 fillers,
- Non-face-based arrays (word/non-human image based),
- Deliberately unfair arrays (of which there was only one dataset),
- No designated GT,
- Poor quality images,
- Duplicated lineup - already included in other dataset(s),
- Statistical outliers.

This sample should be relatively representative of eyewitness research; the list of researchers was compiled by searching sites like Google Scholar, PsycInfo, and the Open Science Framework (using terms like: “eyewitness,” “lineup,” “array,” “target-present AND target-absent,” and “present AND absent”). By searching these sites, the solicitation list included

the most recent and most cited researchers; the list of researchers had a high level of inclusivity to the field. However, there is a possible bias in the sampling process considering that this study was only able to use data from the researchers who responded to the solicitation emails. There was a high rejection and non-response rate, which skewed the sample toward data produced by the participating researchers. Of 97 researchers solicited, 29 shared their datasets. We collected 46 datasets in total. However, some of these datasets were excluded from analysis due to our exclusion criteria. Our final sample was 27 datasets which included 189 arrays.

We collected identification data for 68 arrays. We only collected data from arrays with designated ITs, because this offers a more complete set of data. Of the 189 arrays, 98 (52%) had a designated IT. So, about half of the TA arrays were constructed solely of fillers. There were also a portion of these designated IT arrays that did not offer choosing data with enough detail (e.g., only identified if eyewitnesses were "correct" or "incorrect," did not distinguish IT and filler identifications) and were excluded for the purposes of this paper.

This study respects the privacy of collaborating researchers and the models used in the sampled lineups. None of the lineups or faces will be published or otherwise identified; the lineup models' original expectations of privacy will be upheld. All data has been de-identified to protect privacy of both the researcher who conducted the original study and the participants from the photo lineups.

Measures

The first of our two main measures was the use of Betaface to analyze the face arrays and produce similarity ratings between the faces. These data revealed the degree to which the faces are similar, according to the Betaface algorithm. We compared these ratings via a three-way analysis between the GT and IT, the GT and the fillers, and the IT and the fillers. This analysis

was used to determine how suggestive the lineups are in respect to the Betaface ratings. For our purposes, we produced the Betaface similarity values by uploading a full array to the software, then selecting either the GT or IT image as the base face for analysis. We then ran the "compare faces" function, which commands Betaface to compare the base face to the other faces in the group. When using the "compare faces" function in Betaface, each target was always a 100% match to itself. The additional faces will produce a percentage out of 100% similarity. These values are the similarity data used for our analysis.

The second measure we used was the choosing data from the eyewitnesses from the original experiments. The identification data reveals how participants in the experiments reacted to the array. Exploring these identification decisions could enhance understanding about the level of suggestiveness of the lineup. We collected and analyzed choosing data only from studies with a designated IT, as these studies may reveal more detailed data about choosing behavior. In future work on this dataset, we plan to collect choosing data from all of the arrays; however, there was not enough time and resources to produce all choosing data at this time. We compared these data to the similarity values to determine if identification rates were related to Betaface's similarity algorithm.

Results

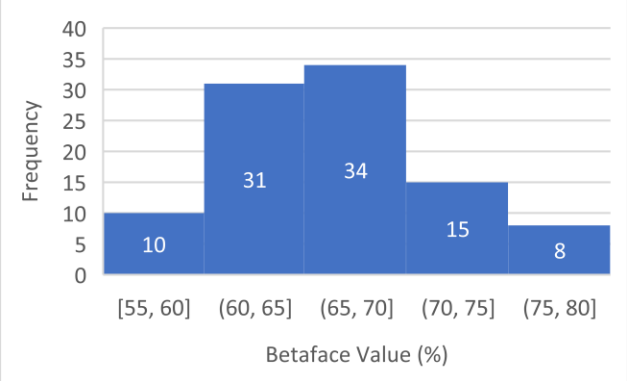
Betaface Similarity

The average GT in our data was a white male between 30 and 50 years old. The GTs in the sample were 33% women and 67% men. The GTs were 73% white, 13% Black, 5% Asian, and 10% Latinx. 4% of GTs were between 13-19 years old, 31% were 20-29, 65% were 30-50, and .5% were over 50 years old.

We ran Betaface's "compare faces" test between the GI and IT, the GT and fillers, and the IT and fillers. As shown in Figure 1, the GT-IT values (N = 98, M = 66.6, SD = 5.2, [56.4-80]) have a normal distribution with a peak in frequency at the 65-70% interval.

Figure 1.

Betaface Similarity Frequency of All GT-IT Values



The GT-filler values (N = 1030, M = 66.4, SD = 5.1, [53.8-87]) peak in frequency between 60-70% and have a positive skew, as evidenced in Figure 2.

Figure 2.

Betaface Similarity Frequency of All GT-Filler Values

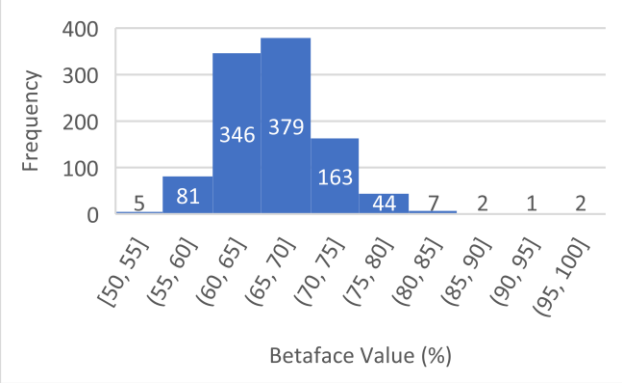
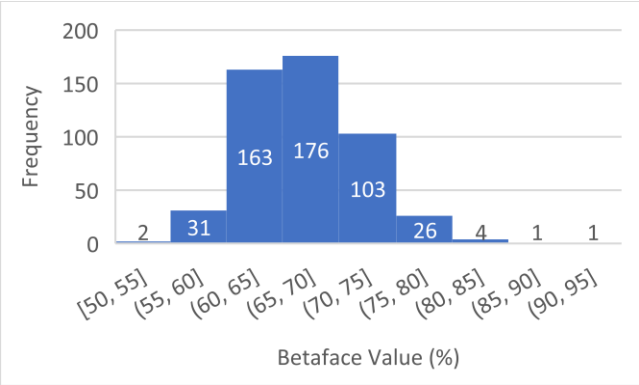


Figure 3 shows that the IT-filler values (N = 507, M = 66.9, SD = 5.2, [52.7-90.8]) are similarly distributed to the GT-filler values.

Figure 3.

Betaface Similarity Frequency of All IT-Filler Values



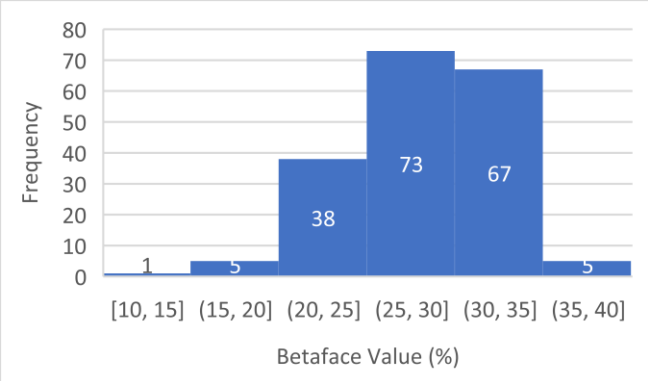
The average similarity of fillers to the targets was 66%. There was only one dataset that we excluded based on statistical outliers – though it is also excluded because of its small array size, as there were only two fillers in each of the three arrays. These arrays produced the highest GT-filler Betaface values from our sample, with a range of 82.4-95.5% similarity. From these data we noted that, though Betaface does not tend to yield very low similarity values, the high similarity values can get close to 100%. The mean filler similarity ratings between the TP and TA arrays are moderately correlated (r = .432, p < .001). This is consistent with the fact that the arrays in this sample tend to be constructed by matching to the GT face. The fillers are very similar to each other, which means that they not only have similar features, but they actually look alike. Additionally, this is consistent with the frequent report that the IT is chosen because the face was the filler with the highest similarity to the GT. Thus, the IT image run via Betaface's "compare faces" function would likely be very similar. We found that there is a moderate

correlation (r = .363, p < .001) between GT-filler similarity and GT-IT similarity across arrays. This means that when the fillers are similar to the GT, the IT also tends to be similar to the GT.

We produced two variables that would measure the similarity difference between the GT and IT and the fillers. When using the "compare faces" function in Betaface, each target was always a 100% match to itself. So, both of these variables measured the maximum filler similarity, respective to each target, subtracted from 100%. The difference from the maximum similarity filler to the GT (N = 189, M = 28.1, SD = 4.3, [13-38.4]) and the IT (N = 98, M = 27.5%, SD = 4.2, [9.2-35.8]) were both relatively small, with means of 28%. It is clear in Figure 4 that the large majority (94%) of the GT-to-maximum-filler differences fell between 20-35%. This grouping emphasizes both the fairly high degree of similarity from the fillers to the GT, and the likenesses in GT-filler similarity across arrays in this sample.

Figure 4.

Difference between the GT (Betaface Value = 100) and the Maximum GT-Filler Value from Each Array

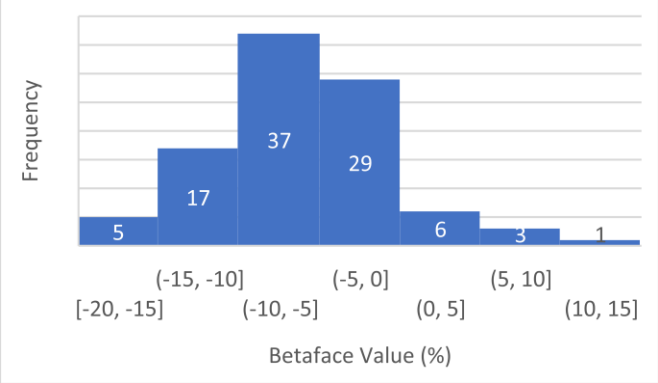


Along these lines, we questioned if the IT was the highest similarity face, in comparison to the GT. Many researchers indicated that the IT had been selected because it was the filler with the most similarity to the GT. We created another variable that measured the difference between the

GT-IT similarity and the maximum filler (N = 98, M = -6.2, SD = 6.0, [-23.5-10.1]). Whatever the basis for choosing the IT, there was often at least one filler in the array with a higher Betaface rating than the IT. Of the 98 arrays measured, 86 (88%) had a maximum filler value higher than the IT. There were 22 arrays with a filler that was 10% higher than the IT. In only 12 arrays, the IT's Betaface value was the same or higher than the maximum filler. As demonstrated in Figure 5, these values peaked between -10% and -5%, so the maximum filler Betaface value tended 5-10% higher than the corresponding IT value.

Figure 5.

Difference between the GT-IT and Maximum GT-Filler Value from Each Designated-IT Array



From this data, the hypothesis that the IT was chosen because it is the most similar face to the GT was not supported, according to the Betaface data. We found that there was a large correlation (r = .523, p < .001) between the difference of GT to the maximum similarity filler and the difference of GT-IT to the maximum similarity filler. This means that as the maximum similarity filler obtains higher GT-filler similarity, it tends to be closer in similarity than the IT to the GT. We ran a bivariate analysis between the mean TP GT-filler similarity values from the designated IT arrays and the non-designated IT arrays and found that there was no significant relationship (N = 98, r = .168, p = .112). This indicates that there is no relationship between the

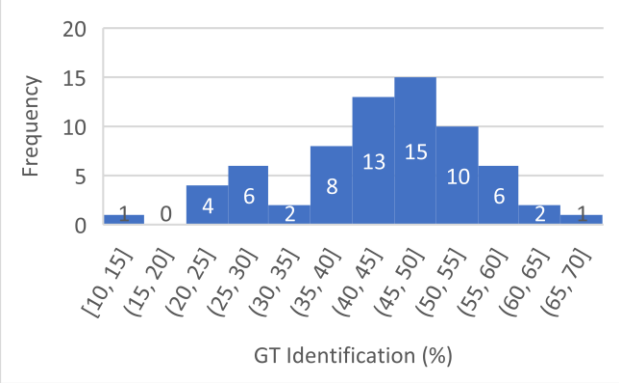
designated and non-designated IT arrays based on GT-filler similarity, so these two methods of construction are not shown to produce similarly fair arrays. Additionally, there is no significant relationship between all of the TP GT-filler similarity values from the designated IT and non-designated IT arrays (N = 621 (designated IT), 504 (non-designated IT), r = -.012, p = .790).

Eyewitness Identification

The frequency of GT identification (M = 43.6, SD = 11.0, [14.6-66.7]) peaks at the 45-50% identification interval, as shown in Figure 6.

Figure 6.

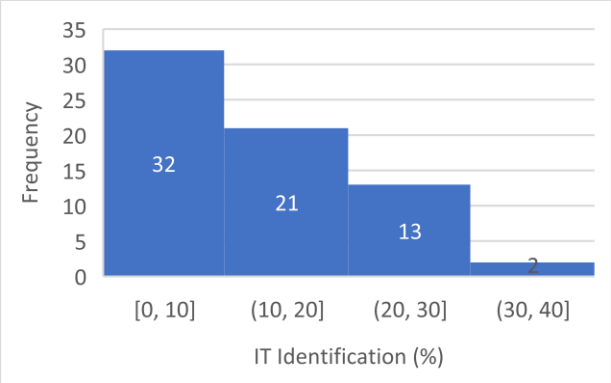
Eyewitness Identification Rates of the GT



As demonstrated in Figure 7, the IT identification rates (*M* = 12.6, *SD* = 8.6, [0.0-38.2]) have a triangular distribution, decreasing in frequency as identification rates increase. The peak frequency of IT identification rates is between 0-10%.

Figure 7.

Eyewitness Identification Rates of the IT



There is a significant relationship between the GT identification rates and the GT-filler similarity (*r* = .202, *p* = .011) so that the GT identification rates increase as GT-filler similarity increases. This relationship may seem counter-intuitive, but we theorize that higher similarity faces may lead to a more focused decision-making process. If all faces in a lineup match the description and look similar, then a witness may be more inclined to closely examine which face matches their memory. There were no significant relationships between GT identification rates to GT-maximum filler similarity difference (*r* = -.160, *p* = .192) or IT identification rates (*r* = .045, *p* = .715). Comparing the IT identification rates to the GT-IT Betaface similarity (*r* = .162, *p* = .187), IT-filler similarity (*r* = .104, *p* = .397), and IT to maximum filler similarity difference (*r* = -.033, *p* = .786) all yielded non-significant relationships. These non-significant relationships could relate to our findings that high IT similarity to the GT is correlated with high filler similarity to the GT. ITs and fillers are similarly high in Betaface similarity. The resulting

lack of variance undercuts the possibility of observing Betaface-identification relationships. It may be true that Betaface values and identification behavior are strongly related, but these relationships can only be observed if ITs and fillers vary significantly across arrays. Finding relationships may also only be possible if the GT and IT are high in similarity (according to Betaface, the ITs did not tend to be the highest similarity faces in the arrays). These findings of dependent variables suggests the need for studies in which GT-IT and GT-filler similarity are systematically varied.

It is disappointing that the identification data resulted in a lack of clear conclusions. Unfortunately, as a result of the study's design and the data we collected, we could not effectively study variables independently. The selection of ITs and fillers was closely tied to the appearance of the GTs. Though this GT dependence is standard for lineup construction, it results in these factors being linked. To find a clearer relationship in Betaface values and identification rates, we would require data from studies that are attentive to systematically manipulating similarities in the selection of the fillers and IT. Manipulating these variables independently would yield non-correlated measures; without these manipulations, the lineup construction variables are somewhat dependent on each other.

Discussion

Key Findings

We hypothesized that the arrays in our sample would be biased vis-à-vis the GT, based on Betaface values. We also hypothesized that Betaface similarity ratings would significantly predict identification decisions. GT-IT, GT-filler, and IT-filler Betaface values had the same mean (66%) and standard deviation (5). They also had similar ranges, with minimum values between 52-56% and maximum values between 80-90%. There was a significant relationship

between the GT identification rates and the GT-filler similarity, from which we concluded that high similarity faces in an array yield more focus, and more correct GT identification rates (though whether this relationship would persist over a wider range of similarity values is an open question). There is a significant correlation between GT-IT similarity values and GT-filler similarity values, so ITs and fillers are correspondingly similar to the GT. These data imply one or both of the following:

1. Betaface produces comparably high-similarity values, even when the faces are not highly similar.
2. The faces are in fact very similar.

We did not find any evidence to show that Betaface is successful and reliable in predicting eyewitness decision-making. This conclusion could be a result of the actual relationships between the variables, or an error due to the study's limitations.

Limitations

Materials Analyzed

This meta-analysis used materials from many other eyewitness identification studies. The original researchers did not gather their data for the intention of using it in this study. As a result, when used for this study, the materials limited parts of our dataset.

For the purposes of this study, we wished to have detailed eyewitness identification data. However, for many datasets in our sample, the choosing data did not detail exactly what decision the eyewitness had made. There were some datasets that only noted if the participant was correct or incorrect. Other studies grouped the fillers together, so we were unable to distinguish which filler a participant had chosen. Only three arrays included individual filler identification data.

Our analysis was limited because only about half of the sampled arrays had a designated IT. This missing data led to us being unable to produce Betaface ratings for the IT and the IT to TA filler Betaface ratings. Additionally, we were most interested in obtaining identification data for designated IT datasets; we could not collect choosing data for non-designated IT datasets, which decreased the power of our identification data. There were also some datasets which had a designated IT, but the materials did not identify which face the IT was. As a result, we had to treat these datasets as if they did not have a designated IT.

The analysis was also limited because almost all the arrays presented the same fillers for both TP and TA arrays. There was only one dataset with one array that presented different fillers between TP and TA arrays.

There may be a bias in our sample because most of our data was shared from individual researchers. A bias may stem from the type of researcher who granted us access to data – these researchers may be more experienced in conducting eyewitness identification studies, or may be more active in the psychological science community. These types of contributors may also have higher quality and more uniformly fair arrays. Additionally, the data that each researcher shared may be biased. For example, researchers may have shared datasets that have less suggestive lineups than their other work. The sample may not be accurately representative of eyewitness identification research.

Betaface Algorithm

We were not able to establish the accuracy of the Betaface algorithm. It is possible that the Betaface similarity ratings are predictive of human similarity judgements from identification arrays – but the fairly uniform fairness of the arrays studied in this project make it difficult to reach any conclusions.

Research-Based Identification Measures

The datasets analyzed in this study relied on research-based eyewitness identification data. Participants in an experiment likely have lower reliability than eyewitnesses for actual cases. They likely are less attentive to detail in an experimental environment, while an eyewitness may take a lineup administered by law enforcement more seriously. Regarding the practical application of research eyewitness identifications, there is some error in the data based on the reliability of the participants' memory quality. Some participants may not have had good memories, or they may have made their identifications on a guess, which could have affected our findings regarding the accuracy of Betaface.

Future Research

In future research, we will collect witness similarity ratings of the arrays in this sample to explore if they are predictive of choosing and correlated with Betaface similarity ratings. Though, as a response to the results from this study, a new project may be to run new studies with the purpose of varying face similarities more dramatically, in order to have useful similarity variance for analytic purposes.

Some of our future research questions include the following:

1. Which is a better prediction of eyewitness identification choices – human similarity ratings or Betaface similarity ratings?
2. Do AI algorithms or humans produce the least suggestive arrays? Are the arrays from this sample, or arrays that Betaface produces from a face database, less suggestive?
3. Do the designated IT datasets and non-designated IT datasets yield different relationships between Betaface similarity and eyewitness identification?

This paper and the data we have collected thus far will be used as the basis of our research, and we will expand on the findings over the next few years.

References

Albright, T.D. & Rakoff, J.S. (2020). The impact of the National Academy of Sciences Report on Eyewitness Identification. *Duke University Judicature, 104*(1), 21—29.

Betaface. (2021). *Betaface API* (Version No. 2.0). https://www.betafaceapi.com/wpa/

Brewer, N., Weber, N., & Guerin, N. (2020). Police lineups of the future? *American Psychologist, 75*(1), 76—91. https://doi.org/10.1037/amp0000465

Carlson, C. A., Jones, A. R., Whittington, J. E., Lockamyeir, R. F., Carlson, M. A., & Wooten, A. R. (2019). Lineup fairness: Propitious heterogeneity and the diagnostic feature-detection hypothesis. *Cognitive Research: Principles and Implications, 4*(2), 1—16. https://doi.org/10.1186/s41235-019-0172-5

Fitzgerald, R. J., Price, H. L., Oriet, C., & Charman, S. D. (2013). The effect of suspect-filler similarity on eyewitness identification decisions: A meta-analysis. *Psychology, Public Policy, and Law, 19*(2), 1—14. https://doi.org/10.1037/a0030618

Innocence Project. (2022). *Eyewitness identification reform*. https://innocenceproject.org/eyewitness-identification-reform/

Krawitz, A. (2019). *An explorable explanation of signal detection theory*. d'etectable. https://decidables.github.io/detectable/index.html

Lee J., Mansour, J. K., & Penrod, S. D. (2023). How to assess lineup fairness: Concurrent and predictive validity of lineup-fairness measures. Manuscript in preparation.

Lindsay, R. C. L., & Wells, G. L. (1980). What price justice? Exploring the relationship of lineup fairness to identification accuracy. *Law and Human Behavior, 4*(4), 303—313. https://doi.org/10.1007/BF01040622

Lucas, C. A., & Brewer, N. (2021). Could precise and replicable manipulations of suspect-filler similarity optimize eyewitness identification performance? *Psychology, Public Policy, and Law, 28*(1), 108—122. https://doi.org/10.1037/law0000329

Neil v. Biggers, 409 U.S. 188 (1972).

Sporer, S. L., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence-accuracy relation in eyewitness identification studies. *Psychological Bulletin, 118*(3), 315—327. https://doi.org/10.1037/0033-2909

Steblay, N. K., & Wells, G. L. (2020). Assessment of bias in police lineups. *Psychology, Public Policy, and Law, 26*(4), 393—412. https://doi.org/10.1037/law0000287

Stovall v. Denno, 388 U.S. 293 (1967).

The National Registry of Exonerations. (2022, March). *Exonerations by contributing factor and type of crime*. https://www.law.umich.edu/special/exoneration/Pages/ExonerationsContribFactorsByCrime.aspx

Wells, G. L., Kovera, M. B, Douglass, A. B, Brewer, N., Meissner, C. A., & Wixted, J. T. (2020). Policy and procedure recommendations for the collection and preservation of eyewitness identification evidence. *Law and Human Behavior, 44*(1), 3–36. http://dx.doi.org/10.1037/lhb0000359

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M, & Brimacombe, C. A. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior, 22*(6), 603—647. http://doi.org/10.1023/A:1025750605807

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. https://doi.org/10.1037/lhb0000096

Yates, S. Q. (2017). *Memorandum for heads of department law enforcement components all department prosecutors*. Office of the Deputy Attorney General.

Appendix A

Additional Figures

Figure A1.

Demographic – Sex of GT Across Arrays

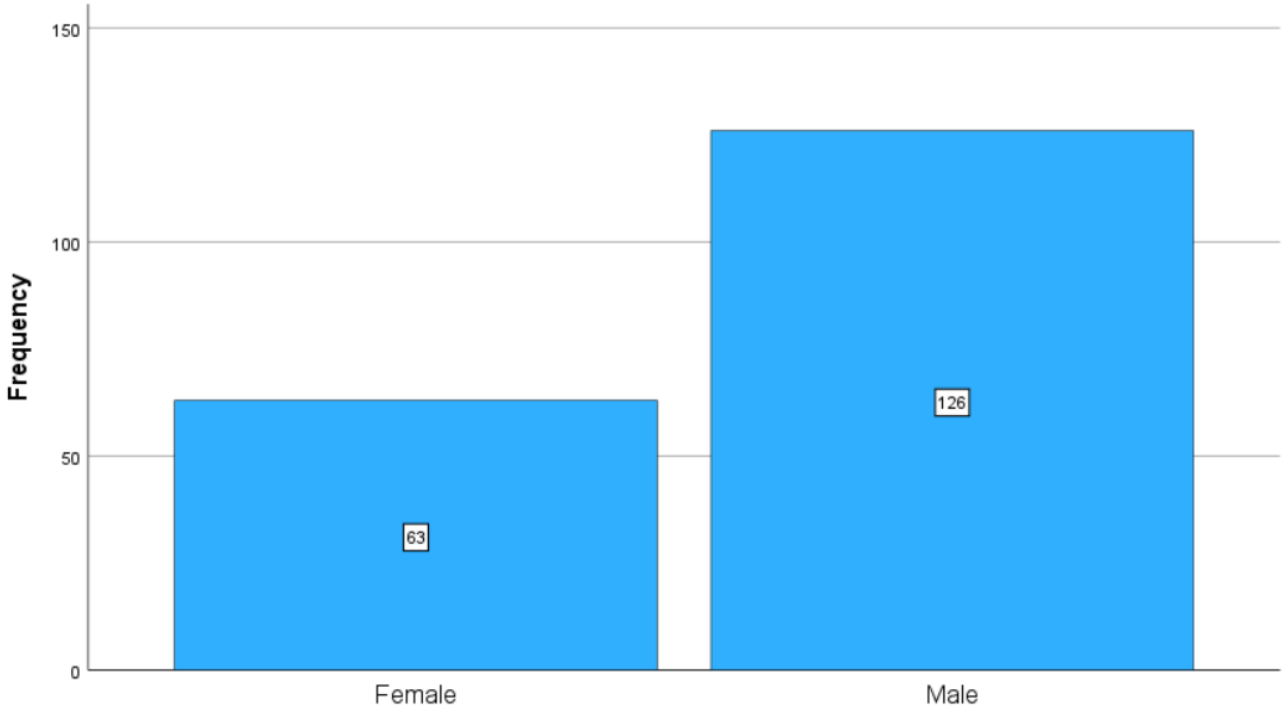


Figure A2.

Demographic – Race of GT Across Arrays

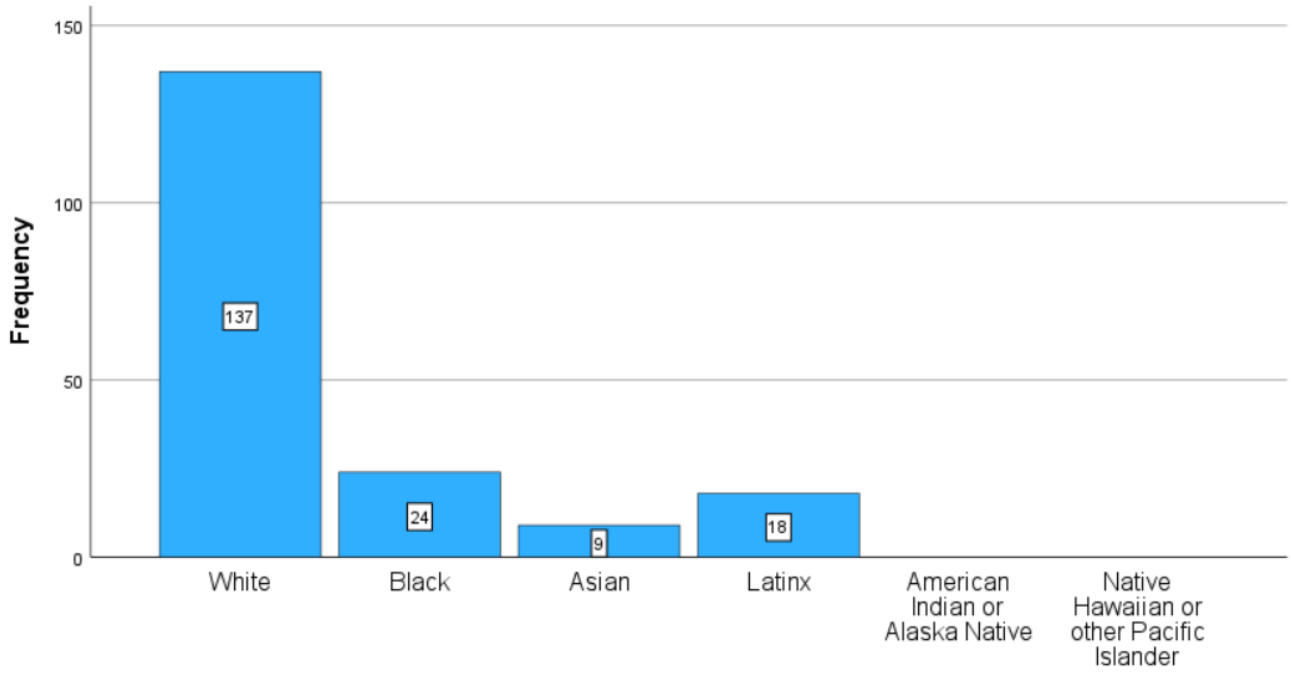


Figure A3.

Demographic – Age (Years) of GT Across Arrays

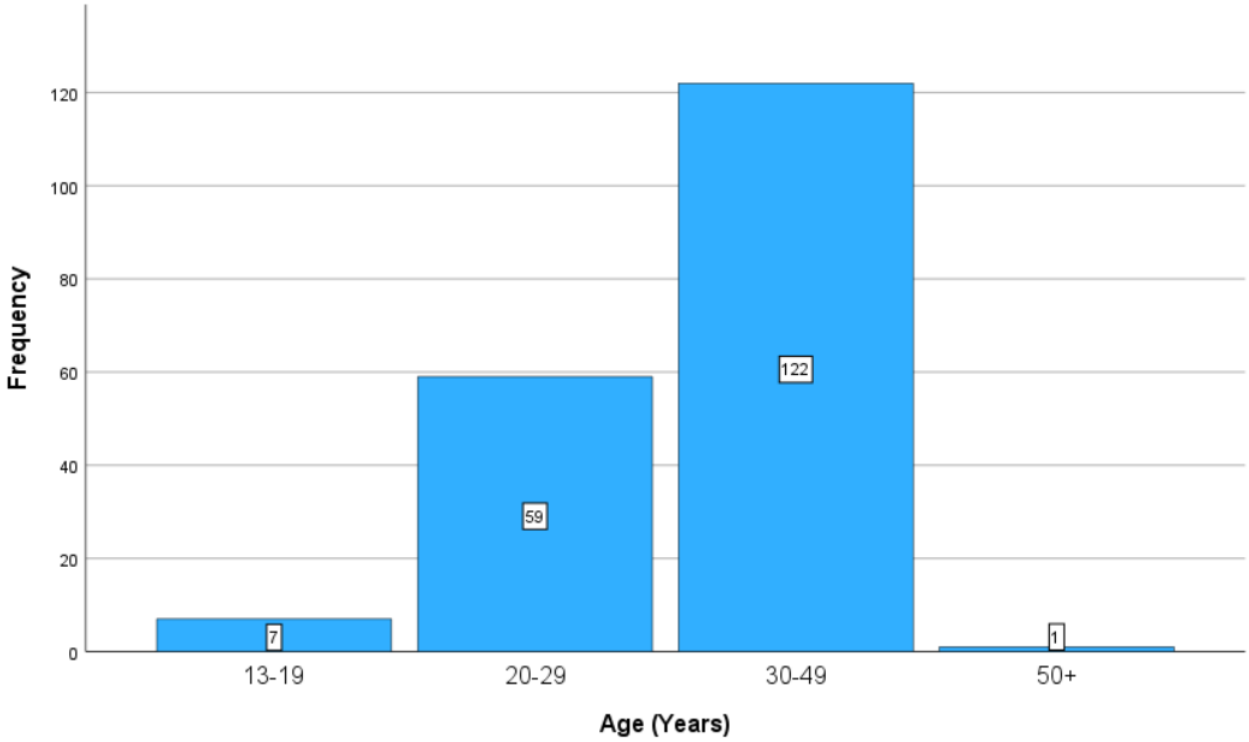
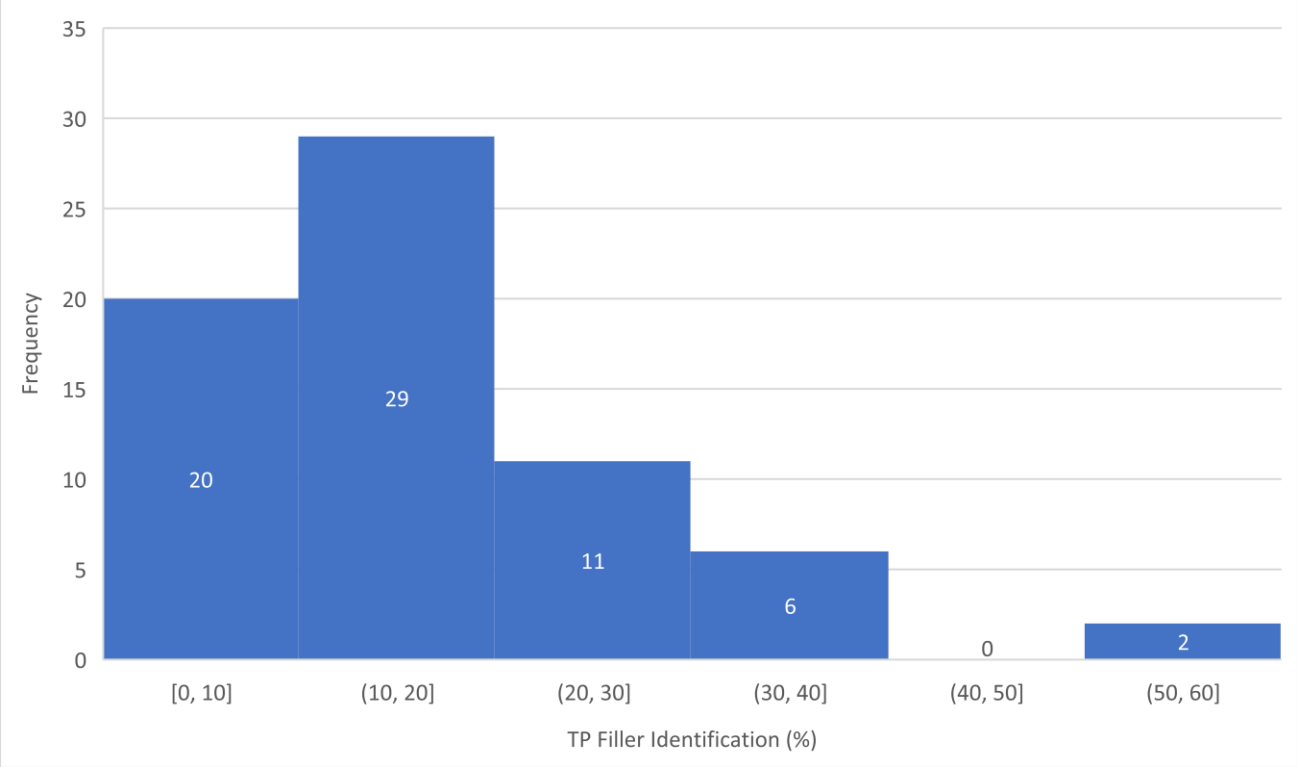


Figure A4.

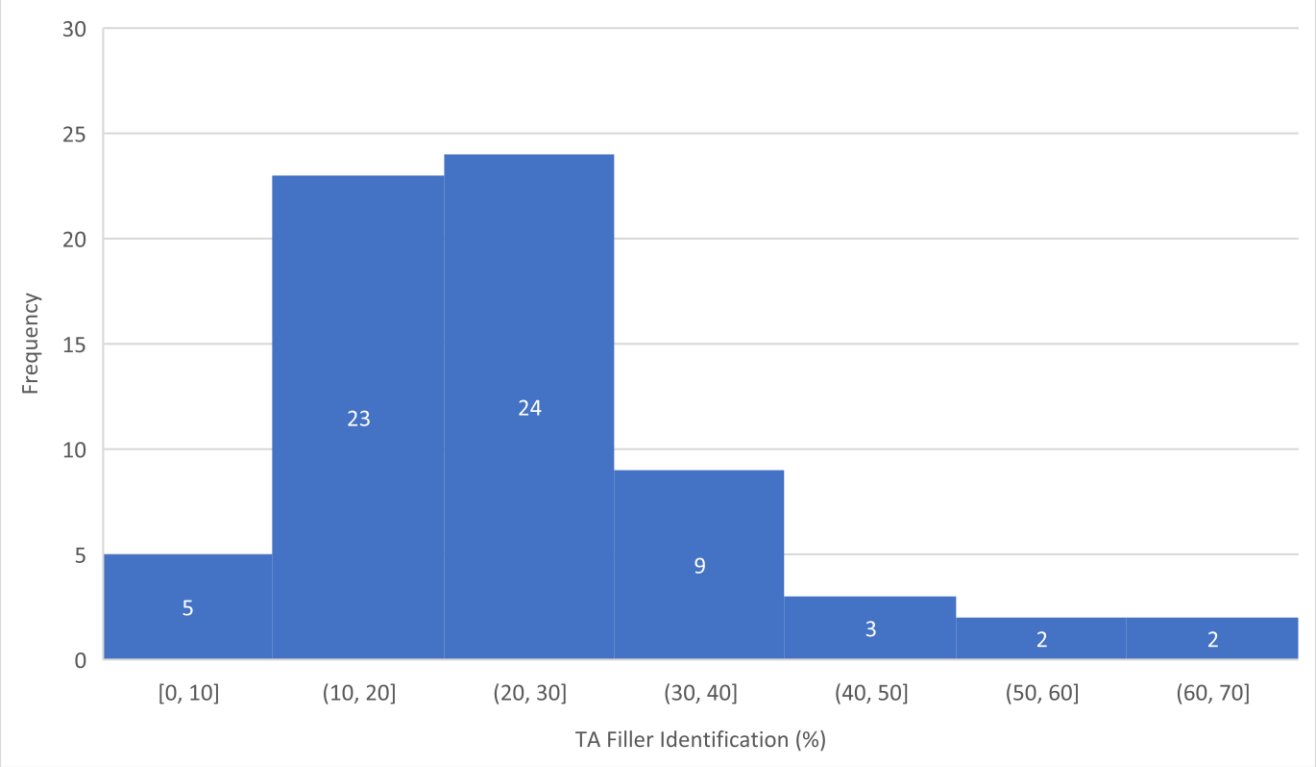
Eyewitness Identification Rates of the TP Fillers



Note. M = 17.3, SD = 10.7, [4.4-56.1]

Figure A5.

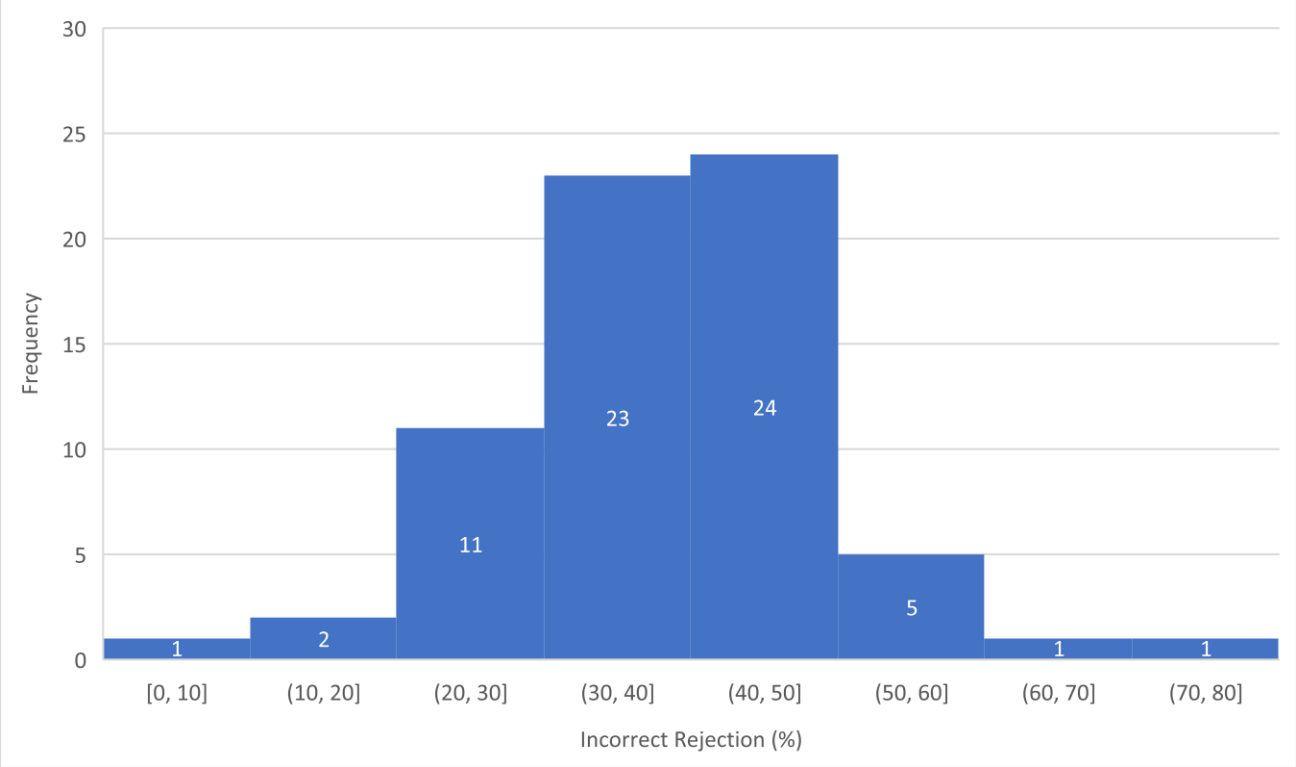
Eyewitness Identification Rates of the TA Fillers



Note. M = 24.3, SD = 12.1, [5.66-60.8]

Figure A6.

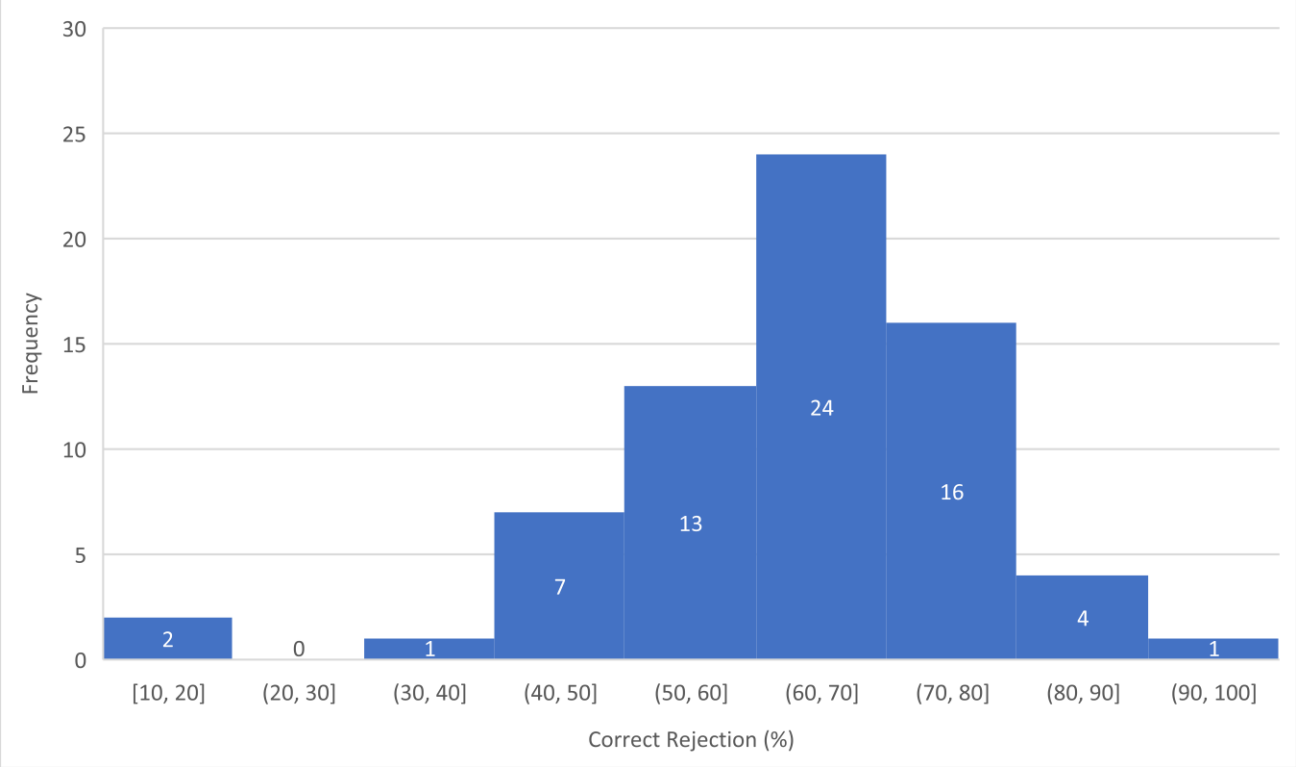
Eyewitness Incorrect Rejection Rates



Note. M = 39.1, SD = 11.4, [7.8-79.2]

Figure A7.

Eyewitness Correct Rejection Rates



Note. M = 63.1, SD = 13.9, [18.2-92.5]

Appendix B

Betaface Upload Limit

Betaface may be accessed for free and without signing up for an account. However, this free access has an upload limit of 50 images/day. We found that Betaface prevents more than 50 uploads from a device using the same network IP address. So, a device would not be able to surpass this limit even if it switched to a different Wi-Fi. Likewise, two different devices on the same Wi-Fi would not be able to surpass the limit. This decreases accessibility and ease of use. This limitation could decrease effectiveness of Betaface as a lineup construction aid.

Betaface offers subscription plans, which allows users to upload more images. The four subscription plans offered are

- Freemium, in which a user may upload 500 images/day, and would pay €0.035 for each extra image;
- Basic (€199/month fee), in which a user may upload 40,000 images/month, and would pay €0.025 for each extra image;
- Premium (€399/month fee), in which a user may upload 100,000 images/month, and would pay €0.02 for each extra image; and
- Ultra (€1299/month fee), in which a user may upload 300,000 images/month, and would pay €0.015 for each extra image (Betaface, 2021).

For the purposes of this study, we got around this limitation by using a virtual private network (VPN) while running the Betaface analysis. This allowed for the same device to be uploading over 50 faces/day, on the same Wi-Fi, because the VPN changes IP addresses.