A Log-Linear Model Approach for Eyewitness Identification Data

Amanda Luby
Advanced Data Analysis

Internal Advisor: Stephen Fienberg
External Advisor: Gary Wells, Iowa State Department of Psychology

January 19, 2016

Abstract

Although eyewitness identification is generally regarded as relatively inaccurate among cognitive psychologists and other experts, testimony from eyewitnesses continues to be prolific in the court system today. There is great interest among psychologists and the criminal justice system to reform eyewitness identification procedures to make the outcomes as accurate as possible. This involves both maximizing the true identification rate and minimizing the false identification rate. There has been a recent push to adopt Receiver Operating Characteristic (ROC) curve methodology to analyze lineup procedures, but has not been universally accepted in the field. This paper addresses some of the shortcomings of the ROC approach and proposes an analytical approach based on log-linear models as an alternative method to evaluate lineup procedures. We find that log-linear models can incorporate more information than previous approaches, and provide flexibility needed for data of this nature.

Contents

1 Introduction

Time and time again, studies have shown that eyewitness identification is unreliable. In a study of 300 convictions that have been overturned due to DNA exoneration, eyewitness identification was a contributing factor of false conviction in over 70 % of cases [4]. Although eyewitness identification does not carry much weight among those trained in cognitive psychology, juries and judges do not have the knowledge regarding memory that psychologists have, and still hold an eyewitness identification as strong evidence against a suspect. There has been a push in recent years to bridge this gap between scientific knowledge and commonly-held beliefs among the general public by determining how to construct police lineups to minimize false identifications without sacrificing true identifications.

There have been many different procedures proposed to improve upon eyewitness identification, including sequential instead of simultaneous presentation, different instructions given to the witness, implementing a double-blind procedure, and having a standardized way of choosing fillers to include in the lineup. These differences are generally measured in a lab setting, however, there is debate about how to best analyze these different procedures.

Some psychologists have turned to ROC curves (Receiver operating characteristic) to analyze how true and false positive rates change over different confidence ratings [15][6][13]. In machine learning, these curves are commonly used to assess binary classification systems across different threshold settings. The true positive rate is plotted on the y-axis with the false positive rate plotted on the x-axis. An ideal ROC curve lies high above the positive diagonal [5][7]. In the context of a binary classification problem in the Machine Learning discipline, where threshold values can be set to many different values and false positive and true positive classifications are clearly defined, it is a useful tool. However, there are fundamental differences between the eyewitness identification problem and a typical classification problem. These differences have led to a divide between psychologists regarding the correct way to interpret lineup experimental results, and it is extremely important to reconcile these differences and determine a statistically-sound procedure for analyzing this type of data.

For instance, when the question of sequential versus simultaneous was first addressed experimentally, the results showed that sequential procedures were 'better' in terms of both the true positive and false positive rate, although at first the decrease in false positive rate was dismissed as insignificant [2]. However, incorporating the confidence statements and constructing a ROC curve can lead to concluding that simultaneous procedures lead to better results. Additionally, after adding confidence bands to the ROC curves comparing sequential and simultaneous lineups, the difference between the two types of lineups was indistinguishable [4]. The estimated uncertainties made optimistic assumptions, and removing these assumptions would lead to even larger confidence bands.

This initial attempt to add statistical rigor to the confidence statement-based analysis of lineups illustrates the importance to better understand the assumptions and uncertainties associated with the chosen analysis. ROC analysis may

not be the appropriate tool to analyze the current data. Typical ROC analysis is an evaluation of a single decision-maker across different thresholds. It is frequently used to evaluate radiologists' ability to detect malignant growths in images; but they recognize the need for a different ROC curve for each radiologist. This helps control for the uncertainty in an accuracy statement, as the threshold for 'X-percent' certain is likely to remain nearly constant for a given radiologist. (CITE) This problem has not been addressed in the eyewitness identification literature, and a single ROC curve is assumed to be representative of the decision-making threshold for the population of eyewitnesses.

At the heart of the current debate regarding the use of ROC analysis in eyewitness identification research is the 3×2 versus 2×2 classification scheme, and how filler identifications should be addressed [11][12][14]. To create an ROC curve, a 2×2 classification scheme must be used, and how this 2×2 classification table is formed using 3×2 classification has a significant effect on the outcome of the analysis.

As an alternative, we propose analyzing the data through the use of a contingency table and log-linear model. The theory behind this categorical data analysis method has been well developed in the statistics literature, and has yet to be applied to the eyewitness identification literature as a means of data analysis. Not only does a log-linear model provide flexibility and robustness that is lacking in the ROC approach to eyewitness identification analysis, it is also able to maintain the natural 3×2 classification structure of the eyewitness identification task.

2 Data

We proceed using data collected in an eyewitness identification experiment performed by Wells and Brewer in 2006 [9]. This experiment tested the difference between biased and unbiased instructions, as well as record confidence statements from participants in target-absent and target-present lineups (Table 11). Unlike published data tables of other studies, we were able to gain access to this data in very fine detail which has allowed for the application of a wider range of analysis methods.

This data was collected by having participants watch a film in which a crime occurred. In groups of 2-4, they watched the video in which the thief entered a restaurant and waited in the background while a customer was leaving his credit card on a counter for a waiter to process. When the customer left, the thief asked the waiter a question which caused him to turn around, when the thief then took the credit card from the counter. After watching the video, participants were given puzzles to work on for fifteen minutes. The participant was then given a target-present or target-absent lineup for the thief; followed by the other option (target-present or target-absent) for the waiter. After participants made a selection, they were asked to report their confidence level.

To avoid the issue of correlation between observations taken from the same subject, we have restricted the results in this paper to the data collected from

the waiter identification task. We chose to use the waiter identification rather than the thief identification for illustrative purposes.

3 Receiver Operating Characterics

Receiver Operating Characteristics (ROC) Curves are often used in 2×2 classification tasks. Each point along a ROC curve represents the Hit Rate (HR) and False Alarm Rate (FAR) at a certain point of confidence, where

$$\text{HR} = \frac{\text{True Positives}}{\text{Target-Present Lineups}} \qquad \text{FAR} = \frac{\text{False Positives}}{\text{Target-Absent Lineups}}.$$

An ideal ROC curve lies high in the upper left corner and corresponds to low false alarm rates and high hit rates.

Wells and Smalarz [10] have argued that while ROC curves are designed for analysis of 2×2 classification outcomes, a line-up setting is actually a 2×3 classification outcome. When used to analyze eyewitness identification in a lineup setting, the case where the eyewitness identifies a filler is combined into the predict false category in the 2×2 case, under the understanding that a filler that is IDed will not be prosecuted, which has led to a skewed perception of eyewitness identification performance. This difference is illustrated in Tables 1 and 2.

	Predict +	Predict -
Actual +	True Positive	False Negative
Actual -	False Positive	True Negative

Table 1: Standard 2×2 Classification Task

	ID Suspect	ID Filler	Reject Lineup
Suspect Guilty	True Positive	False Positive	False Negative
Suspect Innocent	False Positive	False Positive	True Negative

Table 2: 2×3 Classification Structure of Lineup Outcomes

Since collapsing the 2×3 classification into the 2×2 structure essentially means treating the filler identifications as either False Identifications or True/-False Rejections, we lose all information about the filler identifications through the use of a 2×2 structure and, by extension, ROC analysis. This is significant, Wells et. al. argue, because filler identifications can be diagnostic of innocence of the suspect [12] and obscures the filler siphoning effect [11]. Filler siphoning is the term used to explain why good lineup fillers draw some of the false identifications away from an innocent suspect when the actual culprit is not in the lineup.

Although a powerful visual comparison tool, particularly for a 2×3 classification problem, we identified the following statistical issues when comparing lineup procedures.

3.1 Incorporating Uncertainty

In traditional ROC analysis, a 2×2 classifier is evaluated. This classifier can be algorithm based - as in machine learning applications - or it can be a human classifier, as in radiology applications. This application of ROC analysis to radiology is often cited as a justification for use for lineup comparisons CITE . However, a major concern is how to incorporate uncertainty. In an algorithm-based classifier, there is no need for the addition of uncertainty. In radiology and other human classifier evaluation, uncertainty has not been introduced since it has measured a single classifier (human decision maker) across different trials. As ROC curves in the context of eyewitness identification are measuring the correct ID and false ID rate across many different witnesses, the addition of an uncertainty measurement is necessary. In the eyewitness identification literature, we have only recently seen uncertainty introduced to the ROC curves [4] [6].

ROC analysis is being used to compare two different lineup procedures, therefore, the data consists of many individual human classifiers across a single trial (rather than the other way around). Since the data is coming from different people, there is a need for error measurement since we only have data for a sample of people rather than the entire population. Additionally, we need more uncertainty than a confidence band in the usual sense for a line, since there is uncertainty associated in both the hit rate (Y direction) and the false alarm rate (X direction). This is illustrated in Figure 1 .

A further issue with analyzing eyewitness identification using ROC curves, which will be addressed in a later section, is that the threshold value typically used in ML literature is replaced with a confidence statement taken from the witness at the time of the lineup. This is justified by the belief that a witness confidence level is indicative of the decision-making threshold they used in the identification. In a report on eyewitness identification published by the National Research Council [4], the relationship between confidence and accuracy is discussed. The authors note that uncertainty exists in the (HR, FAR) pair, which was accounted for, as well as the Expressed Confidence Level (ECL) of the subjects, which was not. This suggests that even when confidence intervals are added to account for uncertainty in the hit rate and the false alarm rate, these confidence intervals are likely optimistic due to the variability and uncertainty in confidence statements taken from witnesses at the time of the lineup.

3.2 Calculation of False Alarm Rate

One of the strengths of ROC analysis compared to simple hypothesis testing is ROC's ability to compare two quantities at once. These quantities are the Hit Rate (HR) and the False Alarm Rate (FAR). Consider the 2×2 classification

task that ROC was designed for (Table 1). It is clear that the Hit Rate has a direct interpretation for the lineup classification task - when the guilty suspect is in the lineup, how often is he correctly identified. That is,

$$\text{HR} = \frac{\text{\# Suspect ID's}}{\text{\# Target-Present Lineups}}$$

However, the translation of the False Alarm Rate from the 2×2 classification to the lineup classification is not as clear. If we consider the 3×2 lineup classification (Table 2), any formulation of the False Alarm Rate should include the Target Absent lineups in which an innocent suspect is chosen (False Positives). The ambiguity comes in the form of the Filler ID's. In the current literature, these observations are often ignored entirely, under the justification that in an actual lineup situation, these fillers are known to be innocent and thus would not be prosecuted [15][6][3]. However, if we're considering consequences in a true lineup situation, a filler ID in a target-present lineup means that the guilty suspect is not identified and she could then possibly go free. In this sense, a filler ID (TP) outcome is equivalent to a False Negative and by leaving these observations out of the analysis, we are missing information on a consequential outcome. In a 2×2 classification, the False Negative observations are implicitly included in the ROC curve, since $FNR = 1 - HR$. In the lineup setting, we make no adjustment to the (HR, FAR) pair based on the Foil ID's, and so this information is lost.

A potential fix to this loss of information problem is to include Filler ID's in the calculation of the FAR. We would then use the alternative False Alarm Rate:

$$\text{FAR} = \frac{\text{\# False Positives} + \text{\# Filler ID's}}{\text{\# TA Lineups } + \text{\# Filler ID (TP)}}$$

3.3 Restricts Comparison to Two Quantities

The literature has primarily focused on evaluating lineup conditions using quantities involved in ROC analysis. These are the hit-rate and false-alarm rate. These are typically defined as in the previous section.

If there is no designed 'innocent suspect' in the TA lineup, this FAR rate is divided by the number of people in the lineup, which is typically six.

Suppose that we coerce the structure of a lineup into a 2×2 classification task. Consider the set-up of Table 2

In this format, it's clear that the hit rate is solely determined by the bottom row, and the false alarm rate is solely determined by the top row. This can be thought of as conditioning on whether or not the lineup is TA or TP. In other words:

$$\text{HR} = P(\text{Target Identified } | \text{Target in Lineup}) \text{ and}$$

$$\text{FAR} = P(\text{ False ID } | \text{ Target not in Lineup })$$

However, when putting this in the context of the real-world, it seems like a crucial evaluation metric may be missing. In a true lineup, it is unknown whether or not the target is in the lineup. It seems like there are additional quantities of interest, namely

$P(\text{Target guilty} \mid \text{Identification made})$ and $P(\text{Target innocent} \mid \text{Lineup is rejected})$

In the 2×2 classification terminology, these quantities are known as the Positive Predictive Value (PPV) and Negative Predictive Value (NPV), respectively. They are computed through the following:

$$\text{PPV} = \frac{\text{\# of Correct IDs}}{\text{\# of ID's made}} \text{ and } \text{NPV} = \frac{\text{\# of Correct Rejections}}{\text{\# of total rejections}}$$

In the 2×2 classification setting, although not directly represented, these quantities are retrievable through the ROC curve. The FNR (False Negative Rate) can be calculated using $1 - HR$, and the TNR (True Negative Rate) can be calculated with $1 - FAR$, at each threshold value. Then, provided we know the sample size, we can calculate the number of true positives, false positives, false negatives, and true negatives, and calculate the PPV and NPV as described above. However, in the 2×3 classification problem, we are again unable to deal with the filler problem. Then, calculation of the NPV and PPV through the reported ROC curve is impossible.

4 The Log-Linear Model approach to Eyewitness Identification

As we have seen, both the diagnosticity ratio and ROC curve analysis evaluate the performance of lineups through the use of one or two quantities. These quantities are selected from a collection of counts that are taken from the lineup results. We can formulate these lineup outcomes into a contingency table of counts, and rather than use only certain entries of the table for statistical conclusions, we can perform statistical inference on the table itself to draw conclusions about each lineup procedure. This allows us to utilize all of the data collected, and gather a fuller picture of lineup procedures.

If we formulate the lineup outcomes as a contingency table, we can implement a log-linear analysis of the data [1] . That is, the log of the counts in each of the cross-classified cells can be fit using a linear model, and the estimates for the expected values can be computed directly. We denote the observed frequencies in each cell as x_S and the expected value of each cell as m_S, where S is the collection of indices for each of the variables that is used to describe the entry. We use p_S to denote the probability of an outcome falling into the given cell. In the Results section, we implement a log-linear model using both a two-dimensional and four-dimensional model.

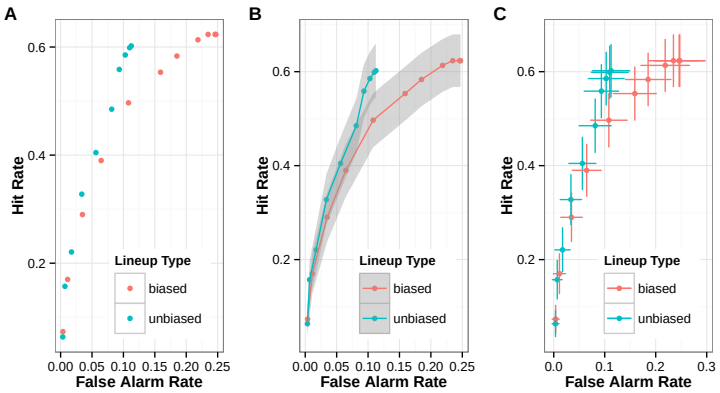


Figure 1: Plot (A) shows the default ROC curve with no uncertainty included, plot (B) shows a binomial confidence interval in the Y direction, and plot (C) incorporates uncertainty in both the X and Y directions.

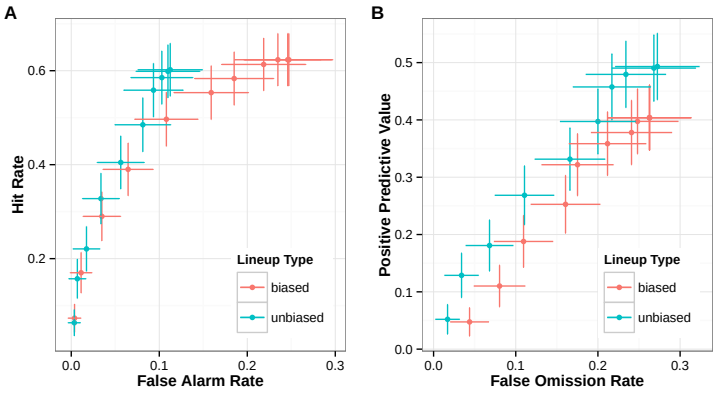


Figure 2: Plot (A) is the usual ROC curve with uncertainty bars. Plot (B) shows a PROC (Predictive Receiver Operating Characteristic) curve, where Positive Predictive Value $= \frac{\text{Correct Identifications}}{\text{Correct Identifications + Foil Choices}}$ and False Omission Rate $= \frac{\text{Incorrect Rejections}}{\text{Incorrect Rejections + True Rejections}}$. In an actual lineup setting, these predictive quantities may be more important than ROC quantities.

Unlike a typical linear model, the parameters are defined using the 'grand mean', u, and deviations from that mean according to variable values. For instance, in a 2×2 contingency table, the saturated model is given by

$$\log p_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$$

where $u_{1(i)}$ represents the deviation from u for the observations which take value i in variable 1, $u_{2(j)}$ represents the deviation from u for observations which take value j in variable 2, and $u_{12(ij)}$ is the deviation from u for variables which take both value i in variable 1, and value j in variable 2. This formulation of the model can be extended into higher dimensions. The model selection process determines which terms can be excluded from the model and still retain a valid fit.

4.1 Fixed vs Random Zeros

In log-linear models, there are two types of zeros that can appear in the tables. The first is due to chance - these are called random zeros and are due to sampling. Theoretically, if we were to observe the entire population, we would expect at least one observation to fall in these cells. Fixed zeros, on the other hand, are due to the nature of the data and regardless of sample size, we would not expect any observations to land in those cells. We see both kinds in our data.

Recall the 2×3 class structure (Table 2). We can then create this table for each level of confidence and transform the data into a $3 \times 2 \times 11$ array. In this fashion, we can create a table with a dimension for each observed variable in a dataset. However, high dimensional tables will be associated with more random zeros than a low dimensional table with the same number of observations. In the eyewitness identification data, it is possible that we would observe a random zero in the cell associated with 'Target Absent', 'Filler Identification', 'Unbiased instructions' and 'ECL = 20', for example. An observation of zero in this cell does not mean that this combination of variables is impossible, but that our sample was not large enough to capture the observations.

However, in many experimental designs, there is no designated innocent suspect in the Target-Absent lineups. A group of six fillers makes up the lineup, and only 'Filler identification' or 'reject the lineup' is recorded. A fixed-zero in the table would thus arise whenever the observation can be cross-classified as 'Target-Absent' and 'Suspect Chosen'. When analyzing the data, we would want to ensure that any expected value for those cells in the table maintains a zero, as any nonzero observations in these experimental settings is impossible.

4.2 Iterative Proportional Fitting

Hierarchical models are those models in which an exclusion of a term implies the exclusion of all higher-order terms that would include that interaction that is excluded. By restricting to hierarchical models, we are able to determine the

Maximum Likelihood estimates for the expected values of the cells in the contingency table using the Iterative Proportional Fitting algorithm. This procedure is guaranteed to converge to the unique set of MLE's, and we are also ensured accuracy to any given degree of the cell estimates. It is flexible enough that we can account for both fixed and random zeros to obtain a desired model. The algorithm is outlined for a three-dimensional table below.

while $|fit(0) - fit(3)| > \delta$ **do**
  fit(1) = fit(0) ×(observedMarginal3/fittedMarginal3)
  fit(2) = fit(1) ×(observedMarginal2/fittedMarginal2)
  fit(3) = fit(2) ×(observedMarginal1/fittedMarginal1)
  fit(0) = fit(3)
end

Algorithm 1: Iterative Proportional Fitting Algorithm

This algorithm is guaranteed to converge to the MLE for three different sampling schemes: (1) A Poisson random variable for each cell, (2) a single multinomial sample, and (3) a set of multinomial sampling schemes. Since our data is drawn using (3), we can obtain the MLE's using the IPF algorithm.

4.3 Model Selection

We describe the graphical model selection process, restricting to two-factor interactions. We first proceed with the conditional edge exclusion test. In this process, we begin with the model including all two-factor interactions. The likelihood ratio statistic, G^2, is calculated for that table, and if the model provides a suitable fit, it serves as the reference model for the rest of the test. If it does not provide a suitable G^2 statistic, we must examine the three-factor interaction terms. We then remove one of the edges, calculate the G^2 statistic for that model fit, and if the fit is no longer adequete, that edge is added to the list of necessary edges. We repeat that process for each two-factor edge, and the result of the unconditional edge exclusion test is the set of edges which were necessary to maintain a good fit.

We then proceed with the conditional edge inclusion test, which starts with the model obtained from the previous process. The G^2 statistic for that model is computed and serves as the reference statistic for the test. Each edge that was removed from the model is added back in, and the difference in G^2 is calculated. If adding the excluded edge results in a significant change in G^2, that edge is included in the final model. This stepwise process is repeated until we have found all significant edges.

5 Results

5.1 Two-Dimensional Cross-Classification

First, we collapsed all other variables and looked at how well the iterative proportional fitting procedure was able to fit the different classification schemes

discussed in the previous section.

We start with the data organized into a 2×3 contingency table (Table 3). We used the `loglin` implementation of iterative proportional fitting algorithm in `R` [8]. The traditional iterative proportional fitting procedure yields the expected values seen in Table 4 ($G^2 = 678.36$, df=2), while adjusting for the structural zero yields the expected values in Table 5 ($G^2 = .3149$, df=1). Thus we see a superior fit when adjusting for structural zeros in addition to evidence that the independence assumption is valid for this set-up of the data.

A natural question that arises from this result is whether or not this independence is due solely to the structural zero. It is also important to note that the structural zero only exists in experimental data in which there is no designated innocent suspect in the target-absent lineups. In order to suggest further use of this model, we would want to know if the assumptions still hold for a different experimental design. We thus transform the data and treat it as it would have been during the original ROC analysis - that $\frac{1}{6}$ of the foil identifications in target-absent lineups should be considered as innocent suspect identifications, and thus the remaining $\frac{5}{6}$ of those foil identifications should continue to be treated as foils. The resulting contingency table is shown in Table 6. The IPF procedure produces expected values shown in Table 7 ($G^2 = 359.86$, df=2) and we see that the independence assumption no longer holds. This suggests that witness choice depends on whether the lineup is Target Absent or Target Present, which we would expect.

5.2 Four-Dimensional Cross-Classification

One of the major benefits of ROC analysis as a comparison tool is that it combines information about four different variables. In one graph, we are able to visualize

1. the Hit Rate
2. False Alarm Rate
3. Biased or Unbiased instruction and
4. Expressed Confidence Level (ECL)

Thus a successful alternative method should have the ability to include, at minimum, the same information. We have already included (a) and (b) in the log-linear model, as well as including the analysis of other lineup outcomes. To include (c) and (d), we must add more dimensions to the analysis.

In the following tables, we use the following numeric notation:

1. To represent witness choice (possible values: suspect ID, foil ID, reject lineup)
2. To represent target status (target absent or target present)
3. To represent lineup condition (biased or unbiased instruction)

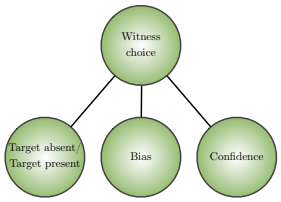


Figure 3: Final graphical model results

4. To represent ECL (0,10,20,...,90,100)

We approach the problem from a graphical model standpoint, and we restrict possible models to two-term interaction and lower. To perform model selection, we first implement an unconditional edge exclusion tests (Table 8). This test uses a G^2 goodness-of-fit test on the graphical model containing all but one of the edges. Models that produce significant p-values suggest that any possible model excluding that specific edge leads to a poor fit. Therefore, all edges that fail the unconditional edge exclusion test must be included in the final model. The entries in the table with significant p-values correspond to $(1,2),(1,3),(1,4)$ edges.

We then perform a conditional edge inclusion test. This method also uses a G^2 goodness-of-fit test. We start with the resulting model from the unconditional edge exclusion test and compare it to a model including one of the other possible edges. We then look at the difference in G^2 between the two models and determine if adding that edge results in a significantly better fit. As seen in Table 9, none of the differences are found to be significant and we don't add any more edges. The resulting graphical model is shown in Figure 3

5.3 Robustness to Expressed Confidence Level

One of the identified shortcomings of ROC methodology applied to eyewitness identification is the use of Expressed Confidence Level (ECL) as the threshold for decision making. As discussed previously, each point on the ROC curve represents the hit rate and false alarm rate for a given ECL. The concern for this design comes from the variability in the expressed confidence level of the witness. We would expect variability both between and within witnesses. That is, we expect different witnesses to have different ECL for the same 'true' confidence in a given lineup, and we also would expect a singular witness to have variability in ECL across many different trials.

Through adding uncertainty intervals to the hit rate and false alarm rate in a given ROC curve, we address the issue of between-witness variability. We now turn to within-witness variability and attempt to compare log-linear models to ROC curves through a simulation study.

We first assume a distribution for the variability of confidence across a given witness. Since ECL's in our case take values 0-100 in increments of ten, we have assumed these ECL distributions are within ± 20 of the observed value. We

	ID Suspect	ID Foil	Reject Lineup	
Target-Absent	0.00	329.00	272.00	601
Target-Present	367.00	132.00	100.00	599
	367.00	461.00	372.00	1200.00

Table 3: Observed counts under no innocent suspect designation.

	Suspect	Foil	Reject
TA	183.81	230.88	186.31
TP	183.19	230.12	185.69

Table 4: Expected values under independence with no correction for fixed-zero cells. $\chi^2 = 530.71; G^2 = 678.36; df = 2$

	Suspect	Foil	Reject
TA	0.00	332.61	268.39
TP	366.86	128.47	103.67

Table 5: Expected values under independence with correction for fixed-zero cells. $\chi^2 = 0.3143; G^2 = .3149; df = 1$

	Suspect	Foil	Reject
TA	55.00	274.00	272.00
TP	367.00	132.00	100.00

Table 6: Observed counts under alternative organization and innocent suspect designation

	Suspect	Foil	Reject
TA	211.35	203.34	186.31
TP	210.65	202.66	185.69

Table 7: Expected values under independence. $\chi^2 = 359.86; G^2 = 391.73; df = 2$

Excluded	χ^2	G^2	p	df
(3,4)	14.62	14.98	0.66	18.00
(2,4)	12.15	12.73	0.81	18.00
(2,3)	12.19	12.62	0.76	17.00
(1,4)	93.46	95.30	0.00	20.00
(1,3)	57.33	58.25	0.00	18.00
(1,2)	315.91	330.25	0.00	18.00

Table 8: Unconditional Exclusion Test Results. Resulting model yields the following goodness of fit results: $G^2 = 15.57$, $df = 21$, $p = 0.79$

Edge	ΔG	Δdf	P-value
(2,3)	0.16	1.00	0.69
(2,4)	0.26	2.00	0.88
(3,4)	2.51	2.00	0.28

Table 9: Conditional Edge Inclusion Test. We conclude that the resulting model from unconditional edge exclusion test does not change. The final model is showin in Figure 3

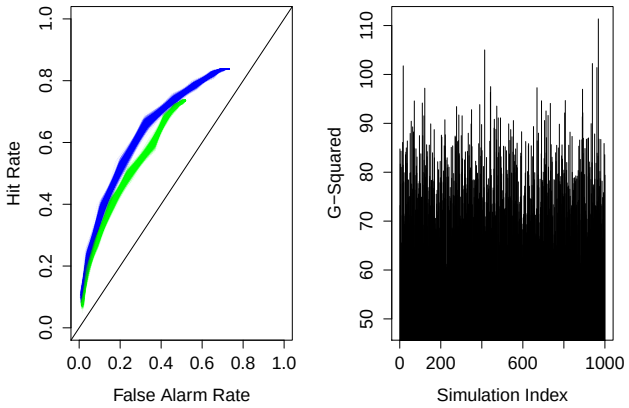


Figure 4: 'Best' case scenario simulation results

then simulate a new ECL for each witness according to that distribution, plot the ROC curve, fit the log-linear model from the previous result and calculate a G^2 value to summarize the fit. We repeat this simulation 1000 times for each assumed distribution.

We tested a range of distributions, with the most optimistic distribution simulating the same ECL 80% of the time, and simulating the ECL - 10 and ECL + 10 ten percent of the time each. The least optimistic distribution takes each of the ECL values ±20 of the observed value 20% of the time. We include the graphical results for each of these situations in Figure 4 and Figure 5 below. We also tested assumed distributions that are both left-skewed and right-skewed. In all cases, the ROC curves overlapped on at least a range of ECL values, making the procedure unable to discriminate between the two lineup conditions. However, even in the least optimistic case, the log-linear model fit the data over 99% of the time. The numeric results from all experiments are shown in Table 10. We conclude that log-linear models provide robustness for variability in ECL in a way that is not feasible in ROC curve methodology.

-20	-10	0	+10	+20	$\bar{G^2}$	$\hat{\text{se}}(G^2)$	Reject
0	.1	.8	.1	0	71.44	8.94	0
0	0.25	0.5	0.25	0	73.71	10.59	0
.1	.2	.4	.2	.1	75.20	11.92	1
.2	.2	.2	.2	.2	76.27	11.51	2
0	0	.7	.2	.1	69.78	9.68	0
.1	.2	.7	0	0	74.38	10.59	0

Table 10: Simulation results for different assumed ECL distributions

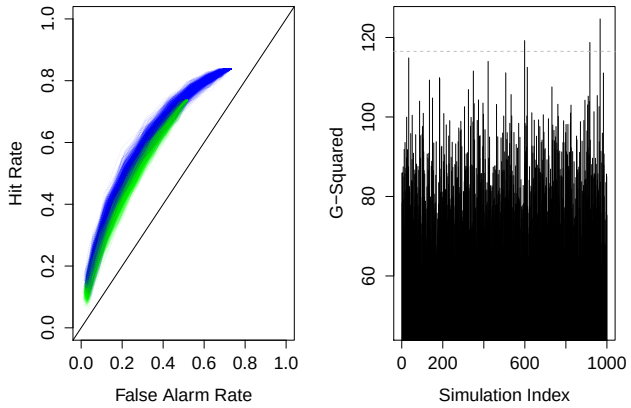


Figure 5: 'Worst' case scenario simulation results

5.4 Flexibility for different experimental assumptions

As discussed before, there are often experimental designs which do not designate an innocent suspect. In these cases, it is often assumed that each filler is identified by the witness uniformly, and the 'suspect ID' cell in the table is filled with $\frac{1}{\text{Size of Lineup}} \cdot$ Filler ID's . This has important implications in the analysis of the data, and there is a need for analysis methods that can handle the difference between designs. This difference in cross-classification is illustrated below.

1. Ideal Setting

	ID Suspect	ID filler	Reject Lineup
Target Present	True Positive	Filler ID (TP)	False Negatives
Target Absent	Innocent ID	Filler ID (TA)	True Negatives

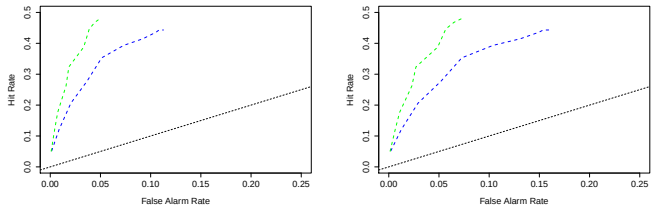
2. Commonly implemented in practice

	ID Suspect	ID Filler	Reject Lineup
Target Present	True Positive	Filler ID (TP)	False Negatives
Target Absent	X	Filler ID (TA)	True Negatives

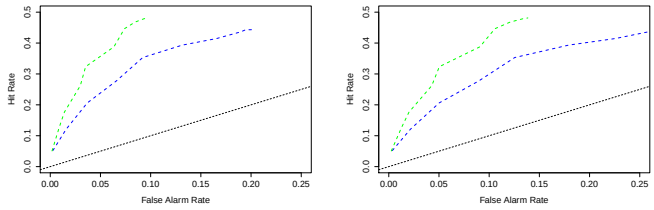
3. How this experimental design is analyzed

	ID Suspect	ID Filler	Reject Lineup
Target Present	True Positive	Filler ID (TP)	False Negatives
Target Absent	$\frac{1}{6}$Filler ID (TA)	$\frac{5}{6}$Filler ID (TA)	True Negatives

One question we aim to answer is how dependent the analysis methods are on the assumption that each filler is chosen uniformly at random. To test this, we arbitrarily chose different fractions. For instance, if we choose $\frac{1}{3}$, we are



(a) FAR=$\frac{1}{6}\times$Filler ID, $G^2 = 64.38$(b) FAR=$\frac{1}{4}\times$Filler ID, $G^2 = 64.69$



(c) FAR=$\frac{1}{3}\times$Filler ID, $G^2 = 64.33$(d) FAR=$\frac{1}{2}\times$Filler ID, $G^2 = 60.55$

Figure 6: Comparison of different innocent suspect designation fractions

testing the scenario that a designated innocent suspect is chosen $\frac{1}{3}$ of the time, while the other fillers are chosen $\frac{2}{3}$ of the time.

We created a new cross-classified dataset for each of the fractions $\frac{1}{6}$, $\frac{1}{4}$, $\frac{1}{3}$ and $\frac{1}{2}$. We then plotted the ROC curves based on this new data, then fit the log-linear model from the previous sections and calculated the G^2 statistic for that fit. As illustrated in Figure 6, the ROC curves stretch out over more of the FAR axis the larger the fraction is. Although this doesn't change which curve is on top, it does have implications for calculating standard errors. We see that each fraction produces a G^2 statistic from the log linear model that is well within the acceptable range.

This suggests that while ROC analysis is impacted by a change in the innocent suspect identification calculation, the log-linear model allows for more flexibility in the proportion of innocent suspect identifications to filler identifications in target-absent lineups. Although the expected values may change, the original graphical model structure still fits the data quite well.

6 Conclusions

We have identified shortcomings of ROC analysis in the context of eyewitness identification experiments. Although a useful tool for evaluating 2×2 classifiers, ROC analysis is not the right tool for the complex classification structure associated with lineup outcomes. We have shown that depending on how the

false alarm rate is calculated, ROC analysis can lead to concluding that either biased or unbiased lineups produce better results. As the definition of 'False Alarm Rate' is not well-defined in the field, this is a troubling issue as different research groups may make opposite conclusions based on the same results. A further statistical issue we have examined is that of quantifying uncertainty. Once uncertainty is added, using even the most optimistic assumptions, ROC does not detect a difference between procedures. Since other statistical procedures have detected differences in lineup conditions after accounting for uncertainty, we would expect a new method of analysis to retain the power to detect these differences. We have also identified other quantities of interest, such as the positive and negative predictive value, which are obscured when using solely ROC curve to analyze experiments.

As an alternative, we have proposed treating lineup outcomes as a contingency table and utilizing log-linear anlaysis, which has been well-established in the statistical community for other categorical data analyses. In the four-dimensional cross-classification, log-linear analysis leads to the exclusion of two-way interaction terms between Target status, Lineup condition, and Confidence statement. Any further exclusions of interaction terms in log-linear analysis leads to a poor model fit; this suggests that biased instructions interacts with Witness choice and has an effect on lineup outcomes. We have shown that log-linear analysis solves the statistical issues associated with ROC analysis, and is flexible enough to be used for different lineup experiments and assumptions.

We have provided what we believe is the first attempt to address the uncertainty associated with expressed confidence levels taken from the witness at the time of identification. Through a simulation study, we have illustrated both the robustness of the log-linear model approach and the variability in ROC curves once this uncertainty is taken into account. We have also tested the log-linear model under different simulated experimental conditions, where it continues to perform well when modeling the data.

Future problems include combining results from multiple experiments into a single log-linear model to better understand the interaction between different lineup conditions. This analysis of eyewitness identification has also led us to the broader issue of the gap between psychology lab studies and implementation in the criminal justice system. More complex experimental design and data collection is necessary to fully understand the effect and interaction of different lineup conditions. There is a need for data collected from the actual application area (in this case, the police departments conducting the lineups) in order to justify the extrapolation from laboratory results to real-world settings.

7 Appendix

7.1 Data

References

[1] Yvonne M. Bishop, Steven E. Fienberg, and Paul W. Holland. *Discrete Multivariate Analysis*. Springer.

[2] Steven E. Clark. Costs and benefits of eyewitness identification reform: Psychological science and public policy. *Perspectives on Psychological Science*, 7(238), 2012.

[3] Steven E. Clark, Ryan T. Howell, and Sherrie L. Davey. Regularities in eyewitness identification. *Law and Human Behavior*, 32:187–218, 2008.

[4] National Research Council. *Identifying the Culprit: Assessing Eyewitness Identification*. The National Academies Press, 2014.

[5] Tom Fawcett. ROC Graphs: Notes and practical considerations for researchers. *HP Laboratories*, 2004.

[6] Scott D. Gronlund, John T. Wixted, and Laura Mickes. Evaluating eyewitness identification procedures using receiver operating characteristic analysis. *Current Directions in Psychological Science*, 23(3), 2014.

[7] Margaret Sullivan Pepe. Receiver operating characteristic methodology. *Journal of the American Statistical Association*, 95(449):308–311, 2000.

[8] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. ISBN 3-900051-07-0.

[9] Gary L. Wells and Neil Brewer. The confidence-accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied*, 12(1):11–30, 2006.

[10] Gary L. Wells and Laura Smalarz. ROC analysis of lineups is not a measure of discriminability. *under editorial review*, 2014.

[11] Gary L Wells, Laura Smalarz, and Andrew M. Smith. Roc analysis of lineups does not measure underlying discriminability and has limited value. *Journal of Applied Research in Memory and Cognition*, 2015.

[12] Gary L Wells, Laura Smalarz, and Andrew M. Smith. Roc analysis of lineups obscures information that is critical for both theoretical understanding and applied purposes. *Journal of Applied Research in Memory and Cognition*, 2015.

[13] John T. Wixted and Laura Mickes. A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review*, 121(2):262–276, 2014.

[14] John T Wixted and Laura Mickes. Evaluating eyewitness identification procedures: ROC analysis and its misconceptions. *Journal of Applied Research in Memory and Cognition*, 2015.

[15] John T. Wixted, Laura Mickes, and Heather D. Flowe. Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous versus sequential lineups. *Journal of Experimental Psychology: Applied*, 18(4):361–376, 2012.

Confidence level (%)	0-20	20-40	40-60	60-80	80-100
Thief-Correct ID	9	21	50	88	54
Thief-Foil ID	12	23	36	24	10
Thief-False ID	13	40	73	61	10
Waiter-Correct ID	8	48	96	117	98
Waiter-Foil ID	12	34	44	31	11
Waiter-False ID	34	73	131	74	17
Thief-Correct Rejection	15	32	110	161	84
Thief-Incorrect Rejection	23	48	85	76	42
Waiter-Correct Rejection	26	45	76	79	46
Waiter-Incorrect Rejection	11	13	28	29	19

Table 11: Confidence statement distributions across different lineup possibilities. [9]