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

December 16, 2015

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 [5]. 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.

Many different procedures have been 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 [17][7][15]. 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 [6][9]. In the context of a binary classifier, ROC analysis 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. 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 [3]. 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 [5]. 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 [8]. 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 [13][14][16]. 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. We will show that 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 [11]. This experiment tested the difference between biased and unbiased instructions, as well as record confidence statements from participants in target-absent (TA) and target-present lineups (TP) (Table 6). 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 set 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. Each participant was then given either a target-present or target-absent lineup for the thief; followed by the other option (target-present or target-absent) for the waiter. After the participant made a selection, she was asked to report her confidence level.

The data set consists of 1200 observations taken from 600 subjects, as each subject participated in two different lineups. 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 Characteristics and its Shortcomings

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.

Although a powerful visual comparison tool for a 2 × 2 classification problem, we discuss four significant statistical issues when using ROC curves for comparing lineup procedures. The first is that lineup outcomes are actually a 2 × 3 classification, and using ROC analysis obscures information about the third class of outcomes. The second is that lineup experiments represent a sample of the population, and there is thus uncertainty associated with the calculated hit rate and false alarm rate for each experiment. There is also variability in the confidence statement taken from each witness, which is used as the 'threshold' for lineup ROC curves. In the eyewitness identification literature, ROC curves are often reported without any uncertainty included. The third issue we discuss is the calculation of the false alarm rate used in ROC analysis. As there are two distinct 'false positive' outcomes in a lineup task, identification of an innocent suspect and identification of a filler, there are different ways to calculate a false alarm rate. Depending on the definition used, ROC curves can lead to conflicting conclusions using the same data. The final issue we discuss regarding ROC curves is that it restricts to comparisons of two quantities. Any other quantities of interest, which may include positive predictive value and negative predictive value, are impossible to calculate using ROC methodology.

3.1 Lineup outcomes are not binary classification

Wells and Smalarz [12] 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. This difference is illustrated in Tables 1 and 2. The third class, which is not included in the 2 × 2 classification analysis of lineups, is filler identifications. A filler identification occurs when the eyewitness identifies a person in the lineup who is not the suspect. This outcome is distinctly different from a suspect identification or a lineup rejection, but is often lost when it comes to analysis. For ROC analysis, the 'filler identification' category is often combined into the predict false category in the 2 × 2 classification scheme (Table 1), with the understanding that a filler who is identified will not be prosecuted. This simplification of lineup outcomes has led to a skewed perception of lineup performance.

	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 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 [14] and obscures the filler siphoning effect [13]. 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.

3.2 Uncertainty needs to be incorporated

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 identification and false identification 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 [5] [7].

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 [5], 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.

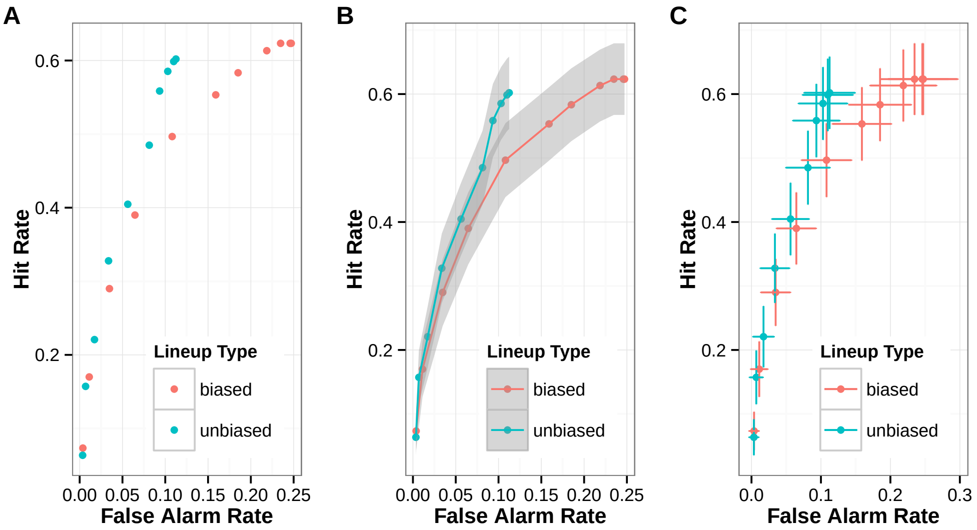


Figure 1: Plot (A) shows the default ROC curve with no uncertainty included, plot (B) shows a binomial confidence interval in the Y direction, and plot (C) incorporates uncertainty in both the X and Y directions.

3.3 Ambiguous 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, known as a target present lineup, how often is he correctly identified. That is,

$$\text{HR} = \frac{\#\text{ Suspect Identifications}}{\#\text{ Target-Present Lineups}}$$

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). In a target absent (TA) lineup, the actual perpetrator is not included in the lineup, and the suspect the police have included is innocent. The ambiguity comes in the form of the filler identifications. 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 [17][7][4]. 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 identification in a target present lineup 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 filler identifications, and so this information is lost.

A potential fix to this loss of information problem is to include filler identifications in the calculation of the false alarm rate. We would then use an alternative definition of the false alarm rate.

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

However, these two different representations of the False Alarm Rate produce contradictory ROC results (Figure 2). This is problematic, since the exact method used to calculate the False Alarm Rate is often not reported in the literature.

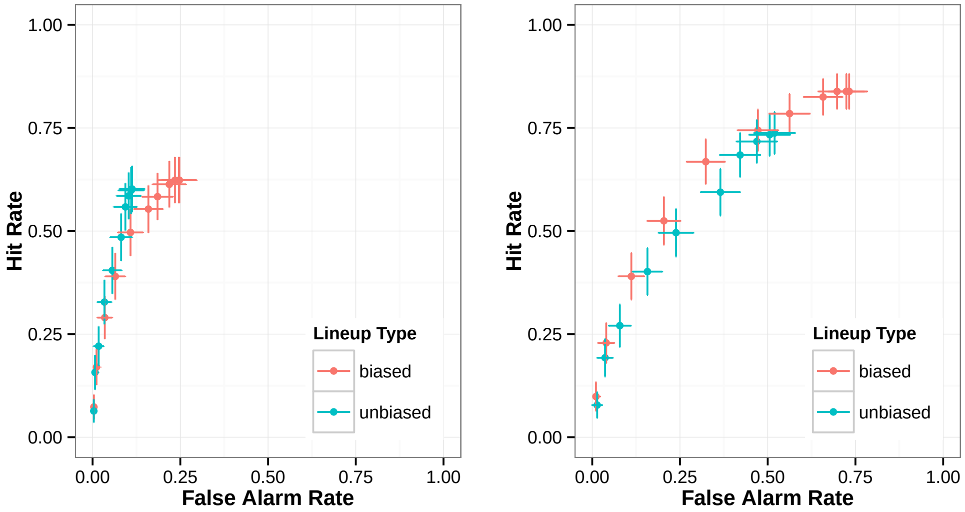


Figure 2: Plot (A) shows the default ROC curve that only includes innocent suspect identifications in the false alarm rate, while plot (B) shows the alternative ROC curve that also includes filler identifications in the false alarm rate. We see that the conclusions from the two plots are contradictory.

3.4 Restricts comparisons 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 target absent lineup, this false alarm 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 top row in the confusion matrix, and the false alarm rate is solely determined by the bottom row. This can be thought of as conditioning on whether or not the lineup is target present or target absent. 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. There may be additional quantities of interest, namely

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

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 positive predictive value and negative predictive value as described above. However, in the 2 × 3 classification problem, we are again unable to deal with the filler problem. Then, calculation of the predictive values through the reported ROC curve is impossible.

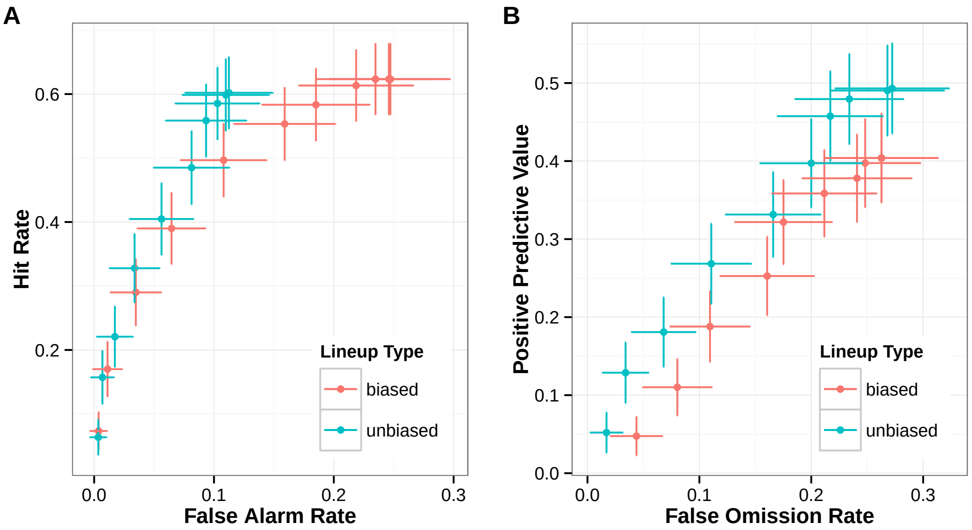


Figure 3: Plot (A) is the usual ROC curve with uncertainty bars. Plot (B) shows a PROC (Predictive Receiver Operating Characteristic) curve, where Positive Predictive Value = $\frac{\text{Correct Identifications}}{\text{Correct Identifications + Foil Choices}}$ and False Omission Rate = $\frac{\text{Incorrect Rejections}}{\text{Incorrect Rejections + True Rejections}}$. In an actual lineup setting, these predictive quantities may be more important than ROC quantities.

4 The Log-Linear Model Approach

As we have seen, ROC analysis evaluates the performance of lineups through the use of 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.

Another statistical procedure that has been proposed to analyze lineup data is logistic regression and its extensions. Logistic regression is a tool used to model binary or multinomial responses based on explanatory variables. When applied to lineup data, the response variable would be the outcome from the lineup procedure, and the explanatory variables describe the conditions of the lineup. Although the logistic regression method allows for multi-dimensional analysis in a way that ROC curves do not, we propose a different, but related, approach - a log linear model. Log linear models rely on much of the same theory as logistic and multinomial regression, but are more general in the sense that they allow for multiple response variables [1]. Since we have seen that lineup outcomes are naturally described through two variables - target present or target absent and witness choice - log linear analysis allows us to study the associations between these variables and explanatory variables in a way that logistic or multinomial regression does not.

If we formulate the lineup outcomes as a contingency table, we can implement a log-linear analysis of the data [2] . That is, the log of the counts in each of the cross-classified cells can be fit using a linear model, and the maximum likelihood estimates for the expected values of each cell 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.

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, which adds more deviation terms and indices to the equation. 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. Both types of zeros are observed in our data set.

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 data set. However, high dimensional tables will be associated with more random zeros than a low dimensional table with the same number of total 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', 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 any observations in this cell.

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 in this form, 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

To fit a log linear model to the lineup data, we restrict to hierarchical models. 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 (MLE) for the expected values of the cells in the contingency table. These estimates are found using the Iterative Proportional Fitting (IPF) algorithm. This procedure is guaranteed to converge to the unique set of maximum likelihood estimators, 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.

In a Poisson sampling scheme, we assume each cell follows a random Poisson distribution with a different mean. Lineup data from laboratory experiments does not follow this scheme, since the number of observations in each cell is restricted by the total number of participants in the study and other controlled variables. A single multinomial sampling scheme assumes the number of observations in each cell is restricted only by the total number of observations, and a set of multinomial sampling schemes allows for the number of observations in each cell to be restricted by multiple constraints. In our case, we want to not only restrict the observations to the total number of participants in the study, but also the number of target present versus target absent lineups, as well as the number of cases which received biased instructions versus unbiased instructions. Then, since our data is drawn using multiple multinomial sampling schemes, we 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. We decide if the model provides a suitable fit using the likelihood ratio statistic, G^2. If the G^2 statistic falls within an acceptable range, it serves as the reference point for the remaining steps in the model selection process. If it does not provide a suitable G^2 statistic, we must examine the three-factor interaction terms. Assuming the G^2 statistic is acceptable, we remove one of the edges, and calculate the G^2 statistic for that model fit. If the fit is no longer adequate, 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 an acceptable fit according to the G^2 statistic.

This resulting model is then built upon using the conditional edge inclusion test. The G^2 statistic from the unconditional edge exclusion test is computed and serves as the reference statistic for conditional edge inclusion test. Each edge that was removed from the model is added back in, one at a time, 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 process is repeated until we have found all significant edges.

5 Results

First, we discuss an initial implementation of a log-linear model for lineup analysis using the two-dimensional 3×2 classification structure discussed in section 3.1. We then include the remaining lineup variables in a four dimensional log-linear analysis to obtain our final model. Third, we compare the robustness to Expressed Confidence Level (ECL) of the resulting log-linear model compared to ROC analysis through a simulation study. Finally, we illustrate the relative flexibility of the log-linear model to different experimental designs.

5.1 Two-Dimensional Cross-Classification

Recall that our data consists of a series of people asked to identify a perpetrator from a lineup. We know whether they received biased or unbiased instructions, whether the actual perpetrator was in the lineup or not, and the expressed confidence level of the witness at the time of the lineup in addition to the lineup outcome. As discussed in section 3.1, lineup outcomes can be formulated as a 2×3 classification task consisting of whether the lineup was target present or target absent, and whether the witness picked the suspect, a filler, or rejected the lineup (Table 2). We began analysis by collapsing biased/unbiased instructions and ECL and looking at how well the iterative proportional fitting procedure was able to fit this two-dimensional classification scheme.

We start with the data organized into a 2×3 contingency table. We used the `loglin` implementation of iterative proportional fitting algorithm in `R` [10]. The traditional iterative proportional fitting procedure yields expected values that are sufficiently different from the observed values to conclude that they correspond to different models ($G^2 = 678.36$, df=2), while adjusting for the structural zero yields expected values close enough to the observed values that they could have been drawn from the same model ($G^2 = .3149$, 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 recommend 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 filler identifications in target-absent lineups should be considered as innocent suspect identifications, and the remaining $\frac{5}{6}$ of those filler identifications should continue to be treated as fillers. The IPF procedure

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 3: Unconditional Exclusion Test Results, indicating that the edges (1,4), (1,3), and (1,2) must be included in the model. 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 4: Conditional Edge Inclusion Test, indicating that additional edges do not produce a significant improvement in fit. We conclude that the resulting model from unconditional edge exclusion test does not change. The final model is showin in Figure 4

produces expected values sufficiently different than the observed values ($G^2 = 359.86$, df=2) and we see that the independence assumption no longer holds. This confirms the commonly-held belief that witness choice depends on whether the lineup is target absent or target present. It also suggests that more variables need to be included in the analysis to explain the data. We thus proceed by including ECL and lineup instructions in a similar analysis.

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 containing two curves, we are able to visualize

(a) the Hit Rate

(b) False Alarm Rate

(c) Biased or Unbiased instruction and

(d) 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 to represent each variable:

$\boxed{1}$ To represent witness choice (possible values: suspect ID, foil ID, reject lineup)

$\boxed{2}$ To represent target status (target absent or target present)

$\boxed{3}$ To represent lineup condition (biased or unbiased instruction)

$\boxed{4}$ 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 3). 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. As seen in Table 4, 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 4, along with the expected values and parametric bootstrap confidence intervals for each lineup outcome in Figure 5. We have thus included not only the four variables included in the ROC approach, but also information about the filler identifications and whether or not the perpetrator is in the lineup.

As we see from the model, witness choice interacts with all three of the other variables - target absent or target present, biased or unbiased instructions, and the expressed confidence level. We also see that

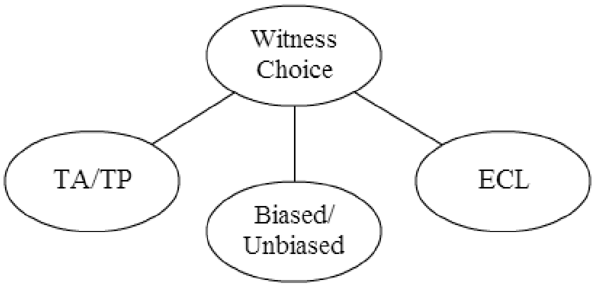


Figure 4: Final graphical model results. We see that Witness Choice interacts with the other three variables, but no other edges are needed. This suggests that instructions influence Expressed Confidence Level only conditionally through Witness Choice.

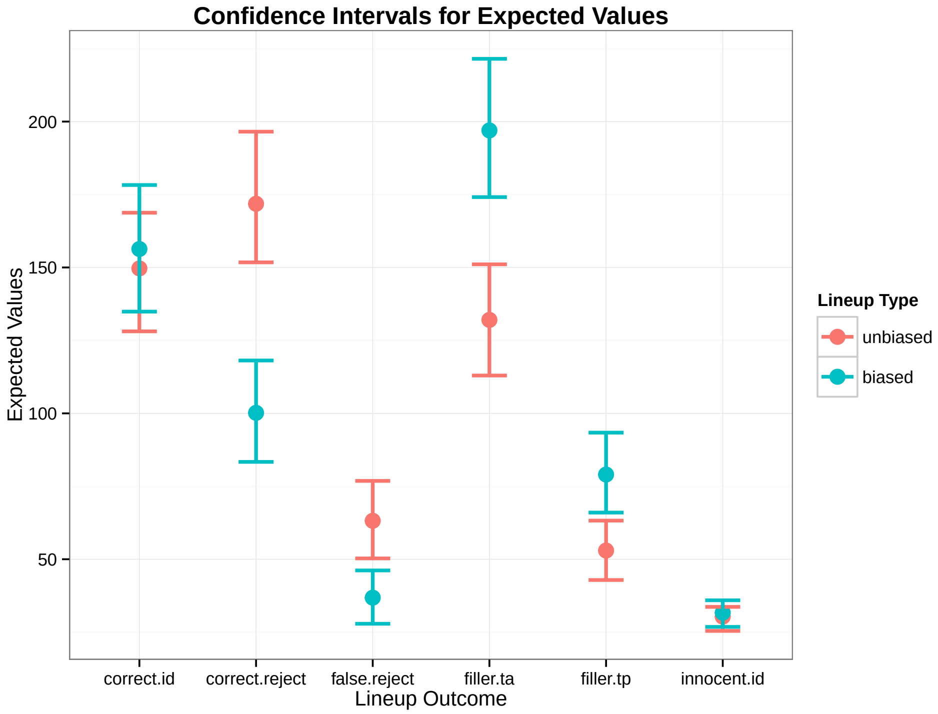


Figure 5: Expected values from the IPF algorithm using the model obtained in Section 5.2. The filler siphoning effect is illustrated through the significant differences in rejections and filler identifications between the biased and unbiased instructions group.

those are the only interaction terms included in the model. This suggests that the choice witnesses make - suspect identification, filler identification, or rejecting the lineup - depends on whether the lineup is target absent or target present. This result confirms the filler siphoning effect discussed in section 3.1. That is, when the actual perpetrator is not in the lineup, witnesses are more likely to pick fillers than when the perpetrator is included in the lineup. We also see that there is an interaction term between witness choice and instruction type. This suggests that whether the witness receives biased or unbiased instructions has an effect on their choice. The last interaction term, between witness choice and ECL, is also consistent with previous results. One of the most interesting things about the resulting model is the lack of interaction term between biased/unbiased instructions and ECL. Psychologists have often thought that instruction type would have an effect on witness' confidence at the time of the lineup, but this model suggests that once we account for witness choice, there is no interaction between instruction type and ECL.

By examining the expected values, we are able to see the filler siphoning effect explicitly. The significant differences in lineup outcomes are found in the rejections and filler identifications. The unbiased instruction group is more likely to reject the lineup, as they are given the explicit option that the perpetrator is not in the lineup. The biased instructions group, on the other hand, is more likely to identify fillers. We thus see that log-linear analysis produces results consistent with commonly-held beliefs in the literature. We now aim to compare performance to the ROC analysis discussed previously.

5.3 Log-Linear Model is robust to ECL

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 in section 3, 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 single 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 in section 3, we have addressed 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 do so by assuming each witness in the original study had the same lineup conditions and outcome, but their ECL is associated with some variability. If we assume that ECL variability follows a given distribution, and we repeated the exact same experiment many times with the same witness, we can then see what the analysis outcomes would be with this variability included.

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 then simulate a new ECL for each witness according to that distribution, plot the ROC curve, fit the log-linear model from the previous section (Figure 4) 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. That is, if a witness reported '60%' confidence in the experiment, we simulate a new ECL for that witness that is either '50%', '60%', or '70%' according to the assumed confidence distribution. The least optimistic distribution takes each of the ECL values ±20 of the observed value 20% of the time. Using the previous example, the witness originally reporting '60%' confidence would have a new ECL simulated that was either '40%', '50%', '60%', '70%', or '80%'. Since this simulation is the most likely to have simulated ECL different than the original, we call it the least optimistic distribution. We include the graphical results for each of these situations in Figure 6 and Figure 7 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 5. We conclude that log-linear models provide robustness for variability in ECL in a way that is not feasible in ROC curve methodology.

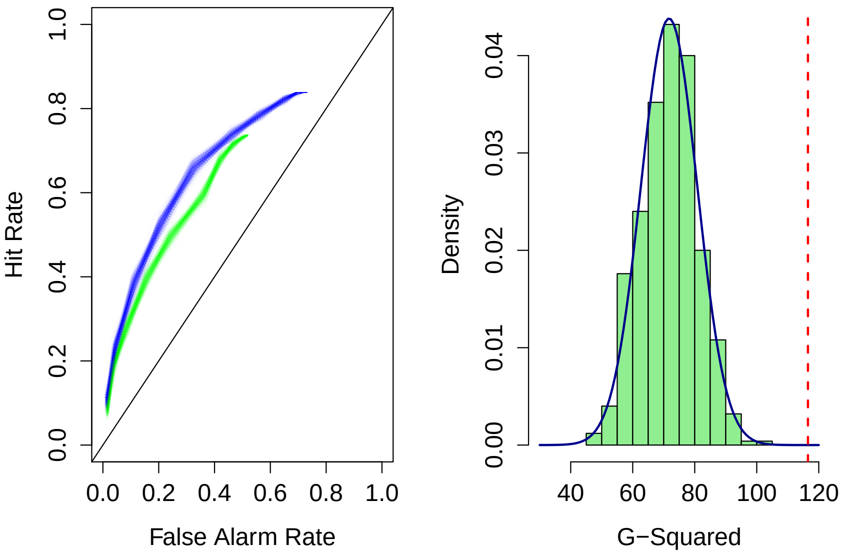


Figure 6: ‘Best’ case scenario simulation results. New ECL values are simulated for each observation - 10% of the time they are ten points higher, 10% of the time they are ten points lower, and 80% of the time they remain the same. We see the G^2 values for the corresponding log-linear model are all within an acceptable range (below the red dotted line), but the ROC curves begin to overlap near the tails, and thus the procedures become indistinguishable because we cannot choose a lineup procedure based on the reported confidence. This suggests that the log-linear model is more robust to variability in ECL.

-20	-10	0	+10	+20	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 5: Simulation results for different assumed ECL distributions. Columns 1-5 represent the probabilities that each simulated expressed confidence level differs from the observed by the given quantity. Column six represents a goodness of fit statistic, column seven represents the standard error of that test statistic, and the eighth column represents the number of times (out of 1000) the log linear model was rejected.

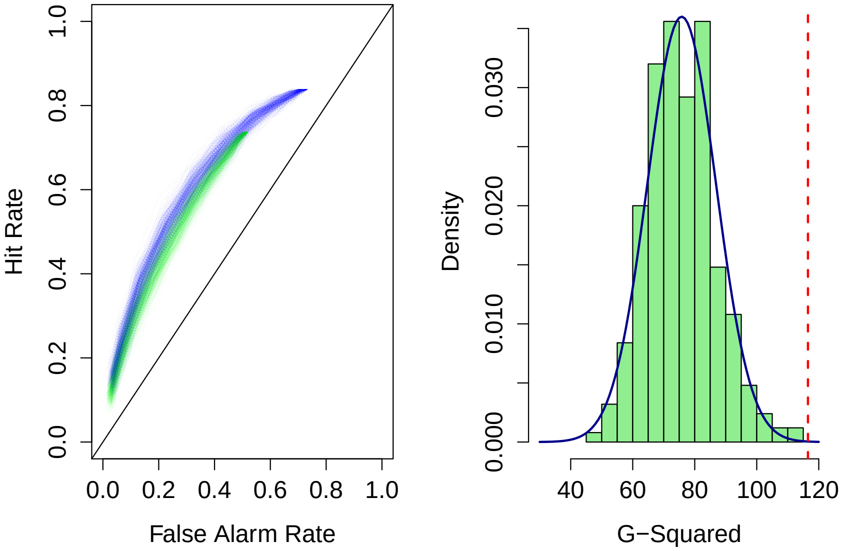


Figure 7: 'Worst' case scenario simulation results. New ECL values are simulated for each observation - 20% of the time they are twenty points higher, 20% of the time they are ten points higher, 20% of the time they remain the same, 20% of the time they are ten points lower, and 20% of the time they are twenty points higher. We see the G^2 values for the corresponding log-linear model continue to be within an acceptable range (below the red dotted line), but the ROC curves overlap over the entire range of ECL, and thus the procedures are indistinguishable because we cannot choose a lineup procedure based on the reported confidence.. This suggests that the log-linear model is more robust to variability in ECL.

5.4 Log-Linear Model is not dependent on 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 $\frac{1}{\text{Size of Lineup}} \cdot \text{Filler ID's}$. 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.

We see that in the first organization, there is a designated innocent suspect that is integrated into the experimental design. The number of times they are identified is recorded in the 'Innocent ID' cell of the table. Often, there is no designated innocent suspect, as in the second table, where there is an 'X' where the 'Innocent ID' should be. Later, when the experiment is analyzed, the second table is transformed into the third table, where $\frac{1}{6}$·Filler ID (TA) is substitued for 'Innocent ID'.

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

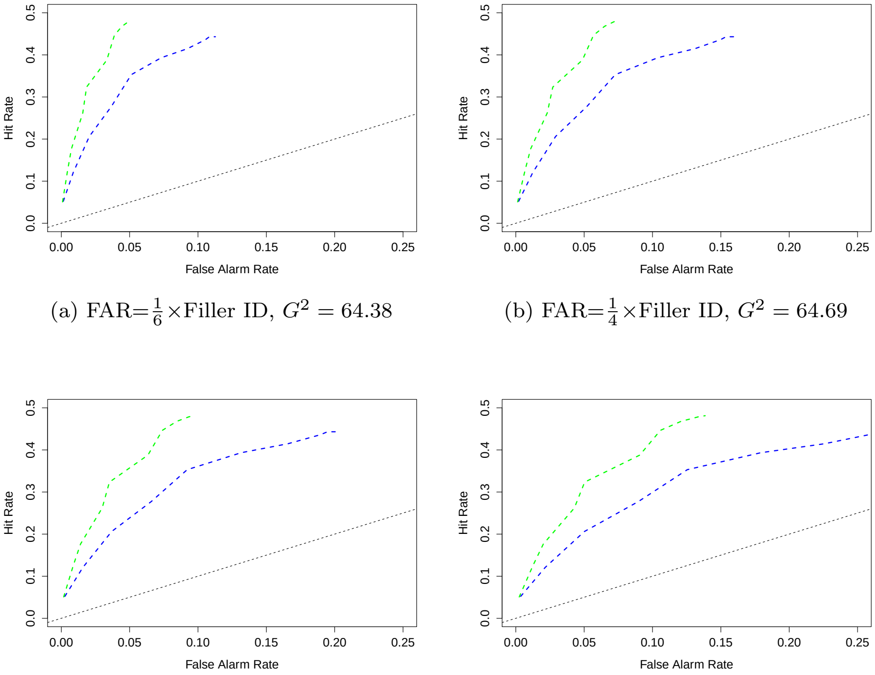
Since this set-up assumes that each filler is chosen uniformly at random, one question we aim to answer is whether the analysis methods are dependent on this assumption. To do so, we arbitrarily chose different fractions to calculate innocent suspect identifications. For instance, if we choose $\frac{1}{3}$, we are 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 8, the ROC curves stretch out over more of the FAR axis the larger the fraction is. Although this doesn't change which curve produces better classification results, it does have implications for calculating uncertainty in results. 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. Whether fillers are chosen uniformly at random, or one filler is chosen more often than another, the log linear model continues to fit the data well. Note that while the graphical model structure remains the same, the expected values for each cell may change.

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) 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 8: Comparison of four different assumptions of innocent suspect designation fractions. The G^2 values included correspond to the fit of the log-linear model produced by the transformed data. We see that although the shape of the curve doesn't change, the range of the ROC space does. This has important implications for standard error calculations.

a new method of analysis to retain the power to detect these differences. Also, while other quantities of interest are identified in the log-linear model approach, such as the positive and negative predictive value, these quantities are obscured when using ROC analysis alone.

As an alternative, we have proposed treating lineup outcomes as a contingency table and utilizing log-linear analysis, 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. Once this uncertainty is taken into account, we have illustrated both the robustness of the log-linear model approach and the variability in ROC curves through a simulation study. 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. More complex experimental design and data collection is necessary to fully understand the effect and interaction of 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. 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

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

7.2 IPF Results

	ID Suspect	ID Foil	Reject Lineup	
Target-Absent	0	329	272	601
Target-Present	367	132	100	599
	367	461	372	1200

Table 7: Observed data collapsed into two dimensions without an innocent suspect designation.

	Suspect	Foil	Reject
TA	183.81	230.88	186.31
TP	183.19	230.12	185.69

Table 8: Expected values from IPF algorithm for Table 7 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 9: Expected values from IPF algorithm under independence for Table 7 with correction for fixed-zero cells. Note that the fits are much closer to the observed values than without the correction for fixed-zero cells. $\chi^2 = 0.3143; G^2 = .3149; df = 1$

	Suspect	Foil	Reject
TA	55	274	272
TP	367	132	100

Table 10: Observed counts from Table 7 with correction for innocent suspect designation.

	Suspect	Foil	Reject
TA	211.35	203.34	186.31
TP	210.65	202.66	185.69

Table 11: Expected values from IPF algorithm under independence for Table 10. We see the fits are sufficiently poor and we cannot conclude that the model is correct. $\chi^2 = 359.86; G^2 = 391.73; df = 2$

References

[1] Alan Agresti. *Categorical Data Analysis.* John Wiley and Sons.

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

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

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

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

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

[7] 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.

[8] Nancy A. Obuchowski. New methodological tools for multiple-reader roc studies. *Radiology*, 2007.

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

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

[11] 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.

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

[13] 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.

[14] 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.

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

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

[17] 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.