Law, Probability and Risk (2017) **16**, 241–257 doi:10.1093/lpr/mgx014
Advance Access publication on November 2, 2017

Strengthening analyses of line-up procedures: a log-linear model framework

Amanda S. Luby[†]
Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, USA

[Received on 28 July 2017; revised on 11 September 2017; accepted on 20 September 2017]

Over the past three decades, there has been considerable interest among both researchers and the criminal justice system to reform line-up procedures to ensure eyewitness identifications are as accurate as possible. Recently, the Receiver Operating Characteristic (ROC) methodology has been adopted to analyse line-up procedures, but it has not been universally accepted in the field. This article examines the application of ROC methodology to line-up data and proposes an approach based on log-linear models as an alternative method to analyse line-up procedures. We find that log-linear models allow for non-binary classification schemes that are necessary for analysing line-up outcomes. Conditional independence relationships between variables can also be identified through a log-linear model. Log-linear models also provide a natural framework to account for multiple sources of uncertainty present in eyewitness identification data. While the ROC analysis provides valuable insight into the changes in outcomes across different decision-making thresholds, incorporating a log-linear analysis allows us to examine these outcomes in finer detail.

Keywords: eyewitness identification; line-ups; ROC; log-linear models.

1. Introduction

The criminal justice system, with support from eyewitness memory researchers, has been actively seeking to improve line-up outcomes for over 30 years. Line-ups are presented to witnesses to see if they can identify a suspect, and often include reports of confidence, although specific procedures vary across departments. Many studies have examined the effects of different procedures on line-up outcomes including the presentation method, instructions and post-identification feedback (Loftus and Palmer, 1996; Steblay *et al.*, 2011, 2001, 2014; Steblay, 1997; Gronlund *et al.*, 2009). However, there continues to be considerable uncertainty over how to best analyse these different procedures, and different analysis methods can lead to vastly different conclusions.

For example, when experiments first addressed the question of sequential versus simultaneous line-up presentation, researchers concluded that sequential line-ups decreased the false identification rate with a negligible decrease in the true identification rate (Lindsay and Wells, 1985; Steblay *et al.*, 2001; Clark, 2012). The Receiver Operating Characteristic (ROC) curves instead demonstrated that simultaneous procedures produced better line-up outcomes (Mickes *et al.*, 2012; Clark *et al.*, 2014; Gronlund *et al.*, 2015). Given that different analysis methods have led to different conclusions using the same set of data, a model-based approach should be adopted in order to make assumptions in the analysis explicitly and to be able to justify conclusions.

Researchers have recently adopted ROC curves to analyse how true and false identifications in line-ups change as confidence ratings vary (Mickes *et al.*, 2012; Gronlund *et al.*, 2014; Wixted and Mickes,

[†]Corresponding author. Email: aluby@stat.cmu.edu

2014; Carlson and Carlson, 2014). However, there are fundamental differences between the eyewitness identification problem and a typical classification problem for which ROC analysis is traditionally used. First, a line-up task has more than one type of false positive identification, due to the presence of known innocent people in the line-up, and some information is lost when there is no distinction between the two identifications. Secondly, the self-reported confidence level of the witness is used in place of the classification threshold. Use of the confidence scale is not constant across witnesses, or even within witnesses, in contrast to the classification threshold in a conventional ROC analysis (National Research Council, 2014). Addressing these differences is necessary to better understand the effects of line-up procedures and determine a statistically sound method for analysing this type of data.

We propose analysing line-up procedures by cross-classifying outcomes of line-up procedures and applying a log-linear model to the resulting contingency table. Not only does a log-linear analysis allow for non-binary classification schemes, it also provides a natural framework to incorporate variability in reported confidence into the analysis.

In eyewitness identification data, variables may appear to interact with each other, but this apparent interaction may actually be due to a common interaction with a third variable. One way this relationship occurs is known technically as conditional independence and, as shown in Section 5 below, log-linear analysis can uncover conditional independences between variables that are not evident from ROC analysis. By using a log-linear analysis in conjunction with an ROC approach to eyewitness identification, we can not only visualize the trade-off between true positives and false positives, but also identify which variables are interacting with one another to explain that trade-off.

To illustrate these concepts, we use data collected in an eyewitness identification experiment performed by Wells and Brewer (2006). This experiment tested the difference between 'unbiased' instructions, in which participants were given an explicit option that the perpetrator may not be in the line-up, versus 'biased' instructions, in which participants were not given the explicit option that the perpetrator may not be present. Although using unbiased instructions is not as controversial as other changes to line-up procedures, these experimental data are still useful for comparing ROC and log-linear analysis, especially when other variables are taken into account.

In Section 2, we introduce a contingency table for line-up data that we use for both ROC and log-linear analysis. In Section 3, we provide an overview of ROC analysis in general, and its application to eyewitness identification in particular. In Section 4, we introduce the log-linear model approach. In Section 5, we use the aforementioned data to construct and interpret an ROC analysis as well as perform a log-linear analysis, followed by a simulation study that addresses the within-witness variation of the Expressed Confidence Level (ECL).

2. A contingency table for line-up data

Contingency tables represent the observed frequency for each combination of categorical variables in a given data set. In the line-up setting, there are three distinct outcomes: a 'line-up rejection', where the witness has not identified anyone in the line-up as the perpetrator; a 'filler identification', where the witness has selected one of the known innocent 'fillers'; and a 'suspect identification', where the witness has selected the suspect. This suspect can be either the true perpetrator or an innocent suspect. These three outcomes, along with whether or not the true perpetrator was present in the line-up, fully define the potential outcomes for a line-up. If the true perpetrator is in the line-up, we call the line-up 'target-present'; and if the true perpetrator is not in the line-up, we call the line-up 'target-absent'. This 2×3 classification of outcomes is illustrated as a contingency table in Table 1.

TABLE 1. *2 × 3 classification structure of line-up outcomes*

	ID suspect	ID filler	Reject line-up
Target-present	True Positive	Filler ID	False Negative
Target-absent	False Positive	Filler ID	True Negative

The hit rate (HR) is an important quantity for constructing ROC curves and is also used in other forms of analysis. The HR is defined as the fraction of the target-present line-ups in which the guilty suspect is correctly identified. That is,

$$\text{HR} = \frac{\sharp \text{ of Guilty Suspect IDs}}{\sharp \text{ of Target-present Line-ups}}$$

Another quantity of interest is the false alarm rate (FAR), or how often an innocent person is incorrectly identified. The calculation of this quantity is more ambiguous for line-ups. Any formulation of the FAR should include the proportion of target-absent line-ups in which an innocent suspect is chosen (labelled False Positives in Table 1). It is unclear, however, where to put filler identifications. These observations are often not included in the FAR calculation at all, with the justification that in an actual line-up situation, these fillers are known to be innocent and thus would not be prosecuted (Mickes *et al.*, 2012; Gronlund *et al.*, 2014). The FAR is then calculated using the following:

$$\text{FAR}_1 = \frac{\sharp \text{ of Innocent Suspect IDs}}{\sharp \text{ of TA Line-ups}}$$

In the real world, however, a filler identification in a target-present line-up means that the actual culprit is not identified and could then possibly go free. By leaving these observations out of the analysis, we are excluding a consequential outcome.

An alternative method is to include filler identifications in the calculation of the FAR, and adjust the denominator accordingly. We would then use the following definition of the FAR instead:

$$\text{FAR}_2 = \frac{\sharp \text{ of Innocent Suspect IDs} + \sharp \text{ of Filler IDs}}{\sharp \text{ of TA Line-ups} + \sharp \text{ of Filler IDs in TP Line-ups}}$$

However, as we show in Section 5.1, using these two different definitions of the FAR may lead to contradictory results.

3. ROC curves

ROC curves are often used in classification tasks that may be completely categorized by two outcomes, often called 'positives' and 'negatives'. In this 2 × 2 classification, each outcome is either a True Positive, True Negative, False Positive or False Negative. Counts of the four possible outcomes in this 2 × 2 classification can be assembled into a contingency table, illustrated in Table 2, which is also called a confusion matrix. The rows correspond to the true state, or whether or not the observations are actual positives ('Actual +') or actual negatives ('Actual −'). The columns correspond to what the classifier predicted each observation to be, either positive ('Predict +') or negative ('Predict −').

TABLE 2. *Standard 2 × 2 classification task, or confusion matrix*

	Predict +	Predict −
Actual +	True Positive	False Negative
Actual −	False Positive	True Negative

For example, suppose we only consider positive eyewitness identifications in which the witness was over 70% confident as evidence of guilt. The number of true identifications with a reported confidence above 70% will be in the 'True Positive' cell of the table, while the number of true identifications with a reported confidence below 70% will be in the 'False Negative' cell of the table. Similarly, the number of false identifications with a reported confidence above 70% will be in the 'False Positive' cell, and the number of false identifications with a reported confidence below 70% will be in the 'True Negative' cell.

Each point along a ROC curve represents the HR and FAR at a certain threshold, where $\text{HR} = \frac{\text{True Positives}}{\text{Actual Positives}}$ and $\text{FAR} = \frac{\text{False Positives}}{\text{Actual Negatives}}$. The threshold determines the cut-off for predicting an observation as positive, such as using 70% confidence as the cut-off in the example above. If we lower the confidence needed for evidence of guilt, we will correctly classify more of the true perpetrators (True Positives) but will incorrectly classify more of the innocent suspects as guilty (False Positives). If we calculate the HR and the FAR at each of these thresholds and plot each pair of values, connected by a line, with the FAR on the *x*-axis and HR on the *y*-axis, we obtain an ROC curve. An ROC curve is thus a visualization of the trade-off between True Positives and False Positives across different thresholds. An ideal classifier will have a high HR and low FAR across all thresholds, which will lead to an ROC curve close to the upper left corner of the graph (Pepe, 2000; Fawcett, 2004).

Although ROC curves are a powerful visual comparison tool for a binary classification problem, there are complications when they are applied to eyewitness identification.

The first complication stems from the ambiguity in calculating the FAR introduced in Section 2. In both possible definitions of the FAR, we are no longer able to make conclusions about the filler identifications through the use of a 2 × 2 structure and, by extension, ROC analysis. This is important because filler identifications can be diagnostic of innocence of the suspect (Wells *et al.*, 2015b), and collapsing this category of outcomes obscures the filler siphoning effect (Wells *et al.*, 2015a): when the actual perpetrator is not in the line-up and if the fillers are chosen to closely resemble the perpetrator, these fillers will siphon away some of the false identifications from the innocent suspect.

A second complication we must consider when adapting ROC analysis to eyewitness identification is the addition of uncertainty in characterizing classification thresholds. In a conventional classifier, only a single classifier is considered and the classification threshold is known with certainty, since it is an adjustable parameter of the classification algorithm. When constructing an ROC curve for line-up procedures, however, we are measuring the HR and FAR for many different witnesses, who can be thought of as an ensemble of classifiers with different internal classification thresholds and who cannot report those thresholds without error. The outcomes that are predicted will thus depend on which witnesses were included in the sample, and there will be some uncertainty in both the HR and the FAR at each confidence threshold due to this sampling variation.

Additionally, the threshold value is replaced with the ECL taken from the eyewitness at the time of identification (Mickes *et al.*, 2012). After an identification (or a line-up rejection), the eyewitness is

asked how confident they are in their decision, which is then recorded as the ECL. Even if the scale and data collection procedures for ECL were standardized, there is no guarantee that every witness would use the scale in the same way to report their internal confidence. This between-witness variability in ECL reporting must be accounted for in the ROC curve in order to make valid conclusions.

We would also expect a single witness to report different ECL's across many different trials, even if their underlying 'True Confidence' was constant, which leads to additional variability (National Research Council, 2014). This within-subject variability must also be accounted for in any complete analysis, which we address in Section 5.3.

4. Log-linear model

In Section 2, a 2×3 contingency table was identified as a natural formulation of line-up outcomes. Log-linear modelling provides a natural framework to analyse this tabular formulation directly. In addition, log-linear models provide a natural way to treat uncertainty in ECL reporting and sampling variability. Log-linear models also provide additional information about the data by identifying conditional independence relationships between variables.

Cross-classifying line-up outcomes in a contingency table allows us to implement a log-linear analysis of the data. That is, the logarithm of the expected counts in each of the cross-classified cells (Y) can be modelled as a linear function of predictors ($\mathbf{x}$, e.g. witness choice, target status, presentation type, line-up instructions), such that

$$\ln E(Y|\mathbf{x}) = \beta'\mathbf{x},$$

where β is the model coefficient vector to be estimated. We can then compute the maximum likelihood estimates for the expected values of each cell directly (Bishop *et al.*, 2007; Agresti, 2002; Wasserman, 2003; Fienberg, 2007). This allows us to model each of the distinct outcomes of a line-up, as in Table 1, for the different line-up procedures, and treat each of the variables as a predictor in the model.

A subtle difference between an ROC approach and the log-linear model approach is the incorporation of ECL. In an ROC curve, ECL is included as the classification threshold, which is a constant value for each witness. In the log-linear model approach, ECL is a predictor included in the model that is associated with a parameter estimate and variability. As we shall see in Section 5, log-linear analysis also provides a method to account for sampling variability in ECL in ROC analysis.

As a further motivation for log-linear modelling, recall the possibility of variables appearing to interact with each other when they are actually conditionally independent. Consider line-ups given by two different officers: one of whom gave instructions emphasizing the need to identify a perpetrator and one of whom gave instructions emphasizing the need to reject the line-up if the perpetrator is not present. Without recording the type of instructions given, we would likely conclude that there is a relationship between line-up outcome and the officer that gave the line-up, while controlling for instructions would likely show that officer and line-up outcome are conditionally independent, given the instruction type.

If there seems to be a relationship between two variables that are actually conditionally independent, and researchers make recommendations to law enforcement based on this conclusion, they may implement an expensive and time-intensive change on the wrong variable. In the example above, we might see a department decide the officers who should conduct line-ups without addressing the type of instructions given, which is the actual variable affecting the outcome. In eyewitness identification studies, finding these conditional independence relationships is important to make valid conclusions

and optimal recommendations. Fortunately, as we will see shortly, a log-linear model naturally identifies these relationships.

5. Application

We now illustrate both ROC analysis and log-linear analysis using data collected in an eyewitness identification experiment performed by Wells and Brewer (2006). This experiment compared biased and unbiased instruction conditions in line-up outcomes.

In groups of 2–4, participants watched a video in which a thief enters a restaurant and waits in the background while a customer leaves his credit card on a counter for the waiter to process. When the customer leaves, the thief asks the waiter a question which causes him to turn around, at which point the thief takes the credit card from the counter. After watching the video, participants were given puzzles to work on for 15 min as a distraction task. Each participant was then given either a target-present or target-absent line-up for the thief, followed by the other option (target-present or target-absent) for the waiter. After the participant made a selection, he or she was asked to report his or her confidence level.

To avoid the issue of correlation between observations taken from the same subject, we only consider outcomes from one line-up, the waiter identification task. We then have 600 observations of each of the four following variables:

- Target status, a categorical variable representing whether the perpetrator was in the line-up that takes two values: (1) target-present or (2) target-absent
- Witness choice, a categorical variable with three factors: (1) suspect identification, (2) filler identification and (3) line-up rejection
- Line-up type, a categorical variable with two factors, one for each line-up condition. In this case, the experiment was testing instruction type, so this variable can either be (1) biased instructions or (2) unbiased instructions.
- ECL, a categorical variable with 11 factors that takes values 0, 10, 20, . . ., 90, 100.

5.1 *ROC analysis*

To perform an ROC analysis of the data, we must calculate the HR and FAR at each of the possible values for ECL for both biased and unbiased instructions. For example, to calculate the HR for biased instructions at the first ROC point, we calculate how many true suspect identifications were made with an ECL equal to 100 out of the biased line-ups, and divide by the number of target-present biased line-ups. To calculate the FAR at the same point, we use the first definition (FAR_1) provided in Section 2. To compute the HR and FAR at the second point, we perform the same calculations, except that we count true and innocent suspect identifications made with an ECL equal to 90 or 100. We continue with this process until we get to the 11th point on the ROC curve, which corresponds to identifications made with any ECL value. We then plot the points with the FAR on the x-axis and HR on the y-axis, and connect the points with a line. We complete this process for each of the experimental line-up conditions, resulting in an ROC curve for each instruction type. This produces Fig. 1A below.

Since ROC curves for line-up are measuring many different witnesses who make up some sample of the true population of all eyewitnesses, the addition of sampling variability is necessary. In Fig. 1B, we have added approximate 90% confidence intervals to both the HR and FAR. To do so, we assumed the

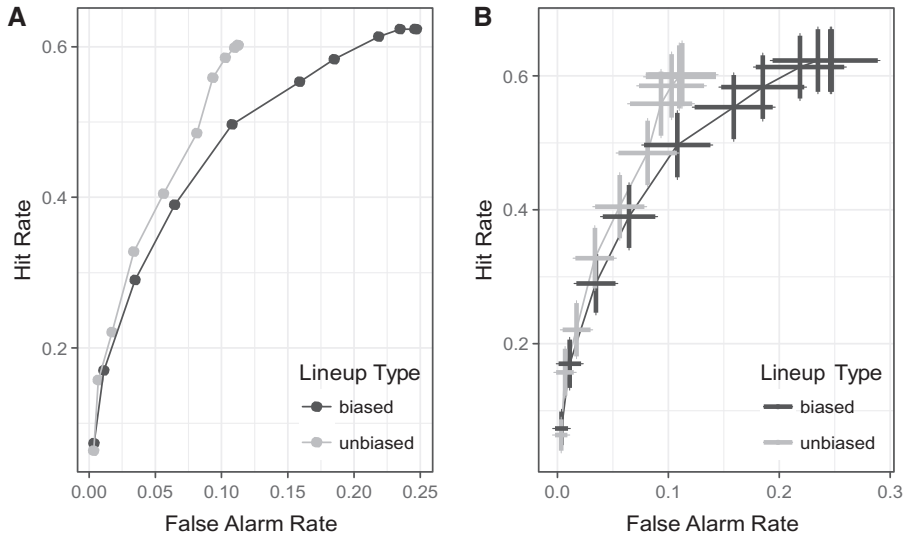

FIG. 1. Plot (A) shows the calculated (FAR, HR) pairs for each line-up type with no uncertainty included. Plot (B) shows approximate confidence intervals in the *X* and *Y* directions added to the points.

HR and FAR correspond to estimates for independent binomial distributions. Since the HR was calculated with only target-present line-ups, and the FAR was calculated with only target-absent line-ups, the confidence set for the (FAR, HR) point is simply a rectangle in the ROC space.

These witnesses may also be reporting different decision thresholds as the same ECL, which is an additional source of uncertainty present in the data that is not accounted for, so these approximate confidence intervals are likely too narrow. We will address the addition of this within-witness variability in Section 5.3.

If the curves cross, or cannot be distinguished from one another, the ROC analysis does not provide evidence that one procedure is better than the other. When we incorporate the confidence intervals, we are defining a rectangle in the ROC space where we think the true (FAR, HR) pair could be, based on the data in the sample. Even if one line-up procedure appears to have an ROC curve that lies above the ROC curve for the other procedure, if there is no separation between the confidence sets, we cannot conclude that this difference is due to anything other than sampling variation.

From the resulting pair of ROC curves in Fig. 1A, we would conclude that although the unbiased instructions initially appear to produce a more optimal trade-off between the HR and FAR, once we account for sampling error in Fig. 1B, the two instruction types become indistinguishable. We thus cannot conclude from the ROC analysis that the different instruction types produce different results.

If we instead use the alternative definition of FAR described in Section 2 (FAR_2), in which we include filler identifications as a False Positive outcome, we would conclude that biased instructions lead to better outcomes (Fig. 2B). Although these ROC curves do not include confidence sets, and would likely be indistinguishable after accounting for sampling variability, this may still help explain why researchers are able to come to conflicting conclusions based on the same data. Without specifying which formulation of the FAR was used, there is no way to tell from the curve itself how filler identifications are included, which could lead to a misunderstanding in interpretation.

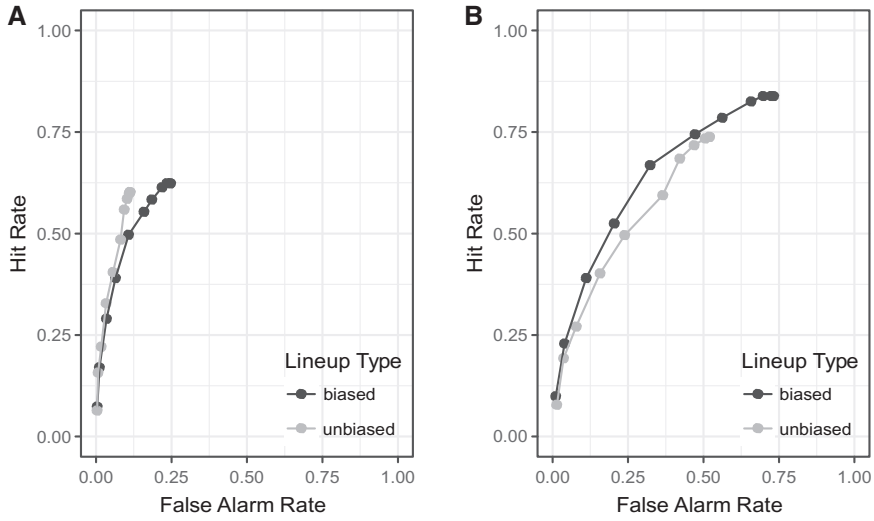

FIG. 2. Plot (A) shows the ROC curve that only includes innocent suspect identifications in the FAR_1, while plot (B) shows the alternative ROC curve that also includes filler identifications in the FAR_2. We see that the conclusions from the two plots are contradictory.

5.2 *Log-linear analysis*

The first step in performing a log-linear analysis is cross-classifying the data into a contingency table. Following the form of Table 1, the data will then consist of a 2×3 table for each combination of line-up type and ECL. That is, we will have a 4D array of size $2 \times 3 \times 2 \times 11$, where the number of layers in each dimension is determined by the number of possible values taken by each categorical variable.

As in ordinary least squares regression, we must determine which predictors to include in the log-linear model introduced in Section 4. Each of the four variables above will be a predictor in the model, but predictors for interactions between two or more variables can also be included, which would imply that the effect of one variable depends on the value of the other variable(s).

Our final model includes a predictor for each of the variables (witness choice, target status, instruction type and ECL), as well as interaction terms for each of target status, instruction type and ECL with witness choice. Details of the model selection process may be found in the Appendix.

We can also estimate expected values for the constructed $2 \times 3 \times 2 \times 11$ contingency table, along with confidence intervals for the estimates. Due to the small number of observations at each individual ECL value, we have collapsed the estimated table over all ECL values in order to observe the overall trends. The expected values for correct identifications, correct rejections, false rejections, filler identifications (TA and TP) and innocent suspect identifications for both biased and unbiased instructions, along with 90% confidence intervals computed from the model, are shown in Fig. 3.

From the two-way interactions identified above, along with the expected cell counts, we are able to make three key conclusions.

First, the interaction between witness choice and instruction type suggests that the type of instructions a witness receives is related to the choice they make. Figure 3 shows that the unbiased instruction group is more likely than the biased instruction group to reject the line-up for both correct and false

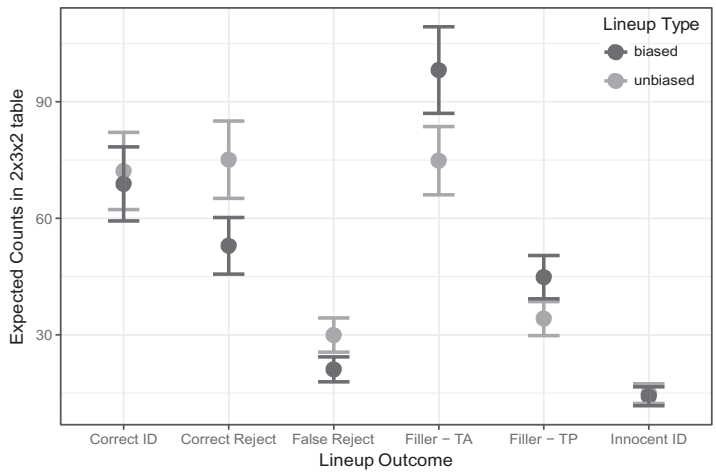

FIG. 3. Estimated expected values and 90% confidence intervals from the log-linear model for the entries in the $2 \times 3 \times 2$ contingency table corresponding to each possible combination of target status, witness choice and instruction type after collapsing over all possible ECL values.

rejections. This conclusion makes intuitive sense, as witnesses who receive unbiased instructions are given the explicit option that the perpetrator is not in the line-up. The biased instructions group, on the other hand, is more likely to identify fillers regardless of whether the line-up was target-present or target-absent.

Secondly, we consider the interaction between witness choice and target status. In Fig. 3, we observe a larger estimate of filler identifications in target-absent line-ups ('Filler–TA') than in target-present line-ups ('Filler–TP'), whether the witness received biased or unbiased instructions. This suggests that when the actual perpetrator is not in the line-up, witnesses are more likely to pick fillers than when the perpetrator is included in the line-up. This result is consistent with the filler siphoning effect mentioned in Section 3.

Finally, we note that although the fitted model has interactions between witness choice and instruction type, and between witness choice and ECL, instruction type and ECL are conditionally independent given witness choice. Some researchers have thought that the type of instructions given should have an effect on confidence level of the witness, but this model suggests that once we account for witness choice, instruction type has no effect on ECL.

In addition to the conclusions made from the fitted model, we can also use the estimated expected values from the fitted model to construct an ROC curve. We are able to obtain smoother ROC curves with the model-based estimates than from those that were obtained from the raw data, as in Section 5.1. The smooth ROC curves are shown in Fig. 4A.

Rather than using a binomial approximation, we can apply standard theory for calculating the variance of linear functions of cell means to estimate confidence intervals (Fienberg, 2007). In Fig. 4B, 90% confidence sets computed from the model are shown. A close comparison of Fig. 4B with Fig. 1B shows that the model-based confidence intervals are narrower than the approximate binomial confidence intervals for the raw data.

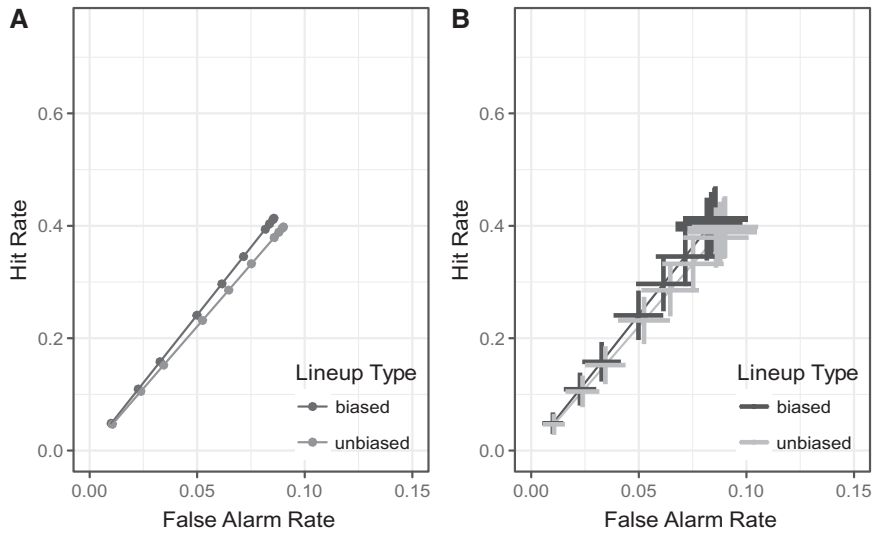

FIG. 4. ROC curves constructed using estimates from the log-linear model.

After constructing smooth ROC curves using the model estimates, the curves may be further away from the observed data than the ROC curves constructed directly from the data, but the estimates are associated with smaller variance. The bias-variance trade-off illustrated here is a well-known statistical phenomenon when selecting a model; generally, smoother models exhibit lower variance but higher bias (Wasserman, 2003). Since our purpose is explanatory rather than predictive, we will not dwell on optimizing the trade-off here.

Figure 4, by itself, or for that matter, Fig. 1, would lead us to believe that the two instruction types are indistinguishable, even though the more refined log-linear analysis leading to Fig. 3 shows there are differences in the estimates of both filler identifications and line-up rejections. The overlap of the confidence intervals in the ROC approach is due to the estimates for correct identifications and innocent suspect identifications being very close for both biased and unbiased instructions, even though the estimates for filler identifications and rejections for each instruction type are quite different.

While the ROC analysis provides valuable insight into the changes in outcomes across different decision-making thresholds, the log-linear analysis allows us to examine these outcomes in finer detail. We can then ascertain conclusions, such as those regarding filler identifications and line-up rejections that are not clear from ROC analysis alone.

5.3 *Accounting for within-witness variability*

As discussed in Section 3, each point on the ROC curve represents the HR and FAR for a given ECL. We would expect some within-witness variability in the reporting of the ECL, meaning we would expect a single witness to report different ECL's across many different trials, even if their 'true' confidence did not change. In this section, we explore this within-witness variability and compare its effect on log-linear models and ROC curves with a simulation study.

TABLE 3. *ECL reporting distribution used in each simulation. C is the original ECL value and C* is the simulated ECL value*

			C^*				
			$C-20$	$C-10$	C	$C+10$	$C+20$
Simulation number	1	$P(C^*)$	0.0	0.1	0.8	0.1	0.0
	2	$P(C^*)$	0.0	0.25	0.5	0.25	0.0
	3	$P(C^*)$	0.1	0.2	0.4	0.2	0.1
	4	$P(C^*)$	0.2	0.2	0.2	0.2	0.2
	5	$P(C^*)$	0.0	0.0	0.7	0.2	0.1
	6	$P(C^*)$	0.1	0.2	0.7	0.0	0.0

In our simulation, we assume each witness in the original study has the same line-up conditions and outcome, but their ECL varies randomly according to some assumed distribution. Thus, each witness is simulated as a deterministic classifier that reports a noisy version of its classification threshold.

Since we are only varying one of the four variables in this simulation, we expect only small changes in the log-linear model fit, although we will check for possible overfitting of the data. We will assess the fit using the G^2 statistic at the $\alpha = 0.05$ significance level. In the ROC analysis, however, we expect to see considerable variation in the curves when varying the classification threshold.

We simulated ECL values (C^*) according to the distributions displayed in Table 3, where C is the observed ECL value from the original data.

Simulations 1–4 were chosen to reflect probability distributions with varying degrees of flatness, with Simulation 4 being the flattest distribution and Simulation 1 being the most centrally concentrated. Simulations 5 and 6 were chosen to test asymmetric distributions in which C^* overestimates C (Simulation 5) and C^* underestimates C (Simulation 6).

In all cases, the ROC curves overlapped on some range of the ECL values, making the procedure unable to discriminate between the two instruction types. However, even when C^* followed the most variable distribution in Simulation 4, the log-linear model fit the data over 99% of the time. This is not necessarily surprising as target status, witness choice and instruction type were held to their original values.

The lack of overfitting in virtually all simulations is a more surprising result. In all cases, removing the interaction between ECL and witness choice led to a significantly worse fit than the original model, even though ECL may be nearly independent of witness choice in some simulations. The log-linear modelling results from all simulations are shown in Table 4.

Due to the lack of differences between the simulations, we choose to focus on the simulation with the most centrally concentrated ECL distribution (Simulation 1) and the simulation with the flattest distribution (Simulation 4) for further analysis. In Figs 5 and 6, we show the ROC curves for both the raw simulated data and the fits from the associated model. We see that, in both cases, the raw ROC curves are spread out over the graph while the model ROC curves fall on a tight line. We also see that the raw ROC curves with 90% confidence sets included for Simulation 4 are noticeably wider than those for Simulation 1, while the model ROC curves do not produce noticeably different confidence sets between simulations. This is, again, an illustration of the bias-variance trade-off mentioned in Section 5.2. Nevertheless, even the smaller-variance model-based estimates do not provide any evidence of separation between these two line-up conditions.

TABLE 4. *Simulation results for different assumed ECL distributions. The $\bar{G}^2$ column contains the mean of the goodness of fit statistics and column 3 contains the standard error of that mean. The fourth column is the number of simulations the log linear model failed to fit the simulated data well. Column five is the number of simulations where removing the interaction term between ECL and witness choice led to an acceptable fit. The G^2 value for the original model was 69.37, $df = 93$*

Sim.#	$\bar{G}^2$	se($\bar{G}^2$)	Poor fit	Overfit
1	75.56	8.92	0	0
2	79.45	10.93	2	0
3	81.48	11.59	3	0
4	81.38	11.84	4	0
5	74.27	9.43	0	0
6	78.84	10.13	0	0

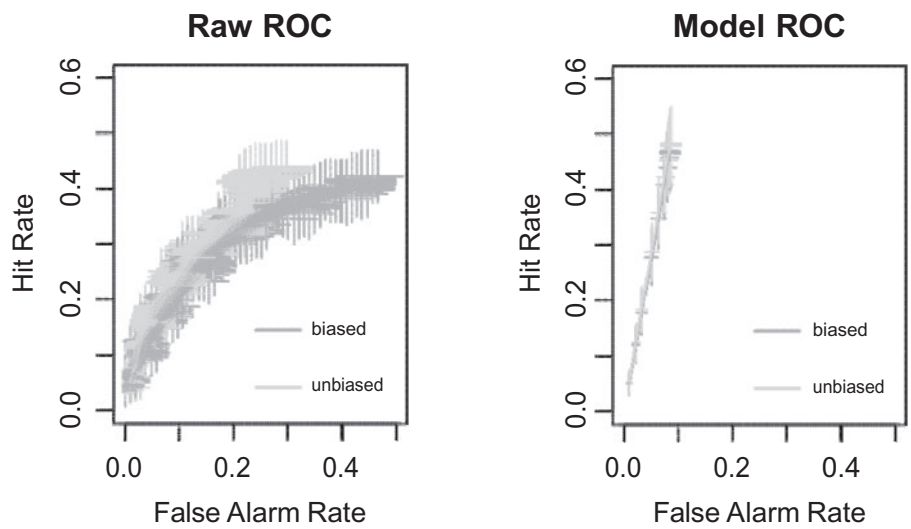

FIG. 5. ROC curves, with 90% confidence intervals, for both the raw simulated data and the log-linear model fits for simulation 1.

6. Discussion

When comparing line-up procedures, different analysis methods have led to different conclusions using the same set of data, producing considerable uncertainty over how to best analyse this type of data. We have proposed the use of log-linear models to analyse line-up outcomes and draw conclusions about line-up procedures.

Using experimental data, we have performed both an ROC analysis and a log-linear analysis, giving examples of conclusions that may be drawn from such analyses. The log-linear analysis led to the adoption of a model that is consistent with previous conclusions on eyewitness memory and line-up

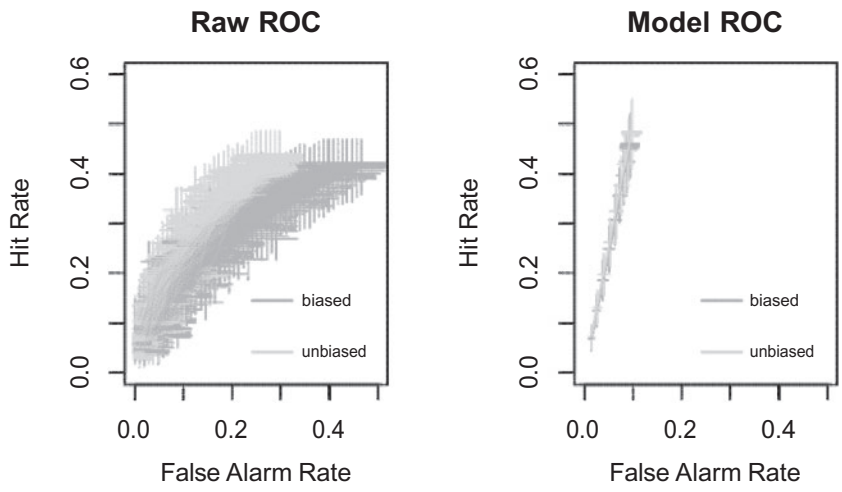

FIG. 6. ROC curves, with 90% confidence intervals, for both the raw simulated data and the log-linear model fits for simulation 4.

performance. We have also illustrated how using a log-linear model in conjunction with an ROC curve can lead to a deeper understanding of the relationship between line-up variables and line-up outcomes, while addressing the multiple types of uncertainty present in the data.

Throughout this work, we have mentioned three distinct sources of uncertainty in eyewitness identification data that must be accounted for: (1) Uncertainty in the model, whether it is an ROC analysis or a log-linear model, due to fitting the model to a finite amount of data that may not be perfectly representative of the true population we are trying to model. (2) Sampling uncertainty, or between-witness variability, due to using a sample of eyewitnesses (who may use different scales to report ECL) to construct either the ROC curve or the log-linear model. (3) Eyewitness uncertainty, or within-witness variability, in reporting their own ECL.

In an ROC curve, the uncertainty in the model is accounted for by constructing standard errors for the (HR, FAR) pair, but the sampling uncertainty and eyewitness uncertainty present in the ECL measurement is not. We have accounted for both model uncertainty and sampling uncertainty in the log-linear analysis through incorporating the standard errors for the parameter estimates into our confidence intervals for the expected values. We have examined eyewitness uncertainty, or within-witness variability, through a simulation study which showed the relative stability of the log-linear model after small changes in ECL.

Simulation studies, while a reasonable first step, are not the full solution for addressing within and between-witness variability. There are many factors, both random and systematic, that may influence both within-witness and between-witness variability in ECL reporting and could be included in a formal measurement model for ECL. The resulting model could then be incorporated into a log-linear analysis. There is substantial work to be done in developing such a measurement model, including more complex designs for measuring ECL and collection of experimental data.

We have restricted comparisons in both the log-linear analysis and ROC analysis to one line-up condition, although combinations of many line-up conditions are often considered. In an ROC

analysis, interactions between multiple line-up conditions can lead to ROC curves that are complicated to interpret (see Carlson and Carlson (2014) for one such example). The log-linear model, however, will naturally extend to the analysis of more line-up conditions. Since we can analyse each interaction term directly in a log-linear model, conclusions in higher dimensions are straightforward, although more data is generally needed for significant results.

There is existing work advocating for adoption of graphical models in legal and forensic decision-making, and graphical log-linear models could be incorporated into these frameworks. Using a graphical log-linear model approach could lead to further benefits, as visualizing the relationship between many different variables is straightforward and conditional independence relationships can be easily identified.

Log-linear analysis is not the only statistical procedure that can be used to analyse eyewitness identification data. Logistic regression, a tool used to model binary or multinomial responses based on explanatory variables, and its extensions have also been proposed (Andersen *et al.*, 2014; Wetmore *et al.*, 2015). Although logistic regression allows for multi-dimensional analysis, we prefer the different, but related, approach of log-linear modelling. Since we have seen that line-up outcomes are naturally described through two variables–target status and witness choice–log linear analysis allows us to study the relationships between these variables and explanatory variables in a way that logistic or multinomial regression does not.

We recommend incorporating a log-linear analysis when making conclusions about eyewitness identification experiments. A thorough model selection process should be used in order to identify any conditional independence relationships that may be present, and uncertainty should be accounted for when making conclusions.

Ultimately, the analysis method chosen is not as important as a careful consideration of the data structure, the uncertainty that may be present in the data and the assumptions that are made in the implementation of the method. We have found that a log-linear analysis provides the flexibility to account for both the structure and multiple types of uncertainty that are present in the data while making minimal assumptions, which leads to statistically sound conclusions.

Acknowledgement

This research was completed under the supervision of Professor Stephen Fienberg, and it would not have been possible without his guidance and support. I am grateful to Professors Brian Junker and Anjali Mazumder for their dedicated help in editing this manuscript.

References

Agresti, A. (2002), *Categorical Data Analysis*, New Jersey: John Wiley and Sons.

Andersen, S. M., Carlson, C. A., Carlson, M. A., and Gronlund, S. D. (2014), 'Individual differences predict eyewitness identification performance,' *Personality and Individual Differences*, **60**, 36–40.

Bishop, Y. M., Fienberg, S. E., and Holland, P. W. (2007), *Discrete Multivariate Analysis*, New York: Springer.

Carlson, C. A., and Carlson, M. A. (2014), 'An evaluation of lineup presentation, weapon presence, and a distinctive feature using ROC analysis,' *Journal of Applied Research in Memory and Cognition*, **3**(2), 45–53.

Clark, S. E. (2012), 'Costs and benefits of eyewitness identification reform: psychological science and public policy,' *Perspectives on Psychological Science*, **7**(3), 238–259.

Clark, S. E., Moreland, M. B., and Gronlund, S. D. (2014), 'Evolution of the empirical and theoretical foundations of eyewitness identification reform,' *Psychonomic Bulletin and Review*, **21**, 251–267.

Fawcett, T. (2004), 'ROC graphs: notes and practical considerations for researchers,' *Machine Learning*, **31**(1), 1–38.
Fienberg, S. E. (2007), *The Analysis of Cross-Classified Categorical Data*, New York: Springer.
Gronlund, S. D., Carlson, C. A., Dailey, S. B., and Goodsell, C. A. (2009), 'Robustness of the sequential lineup advantage,' *Journal of Experimental Psychology: Applied*, **15**(2), 140.
Gronlund, S. D., Wixted, J. T., and Mickes, L. (2014), 'Evaluating eyewitness identification procedures using receiver operating characteristic analysis,' *Current Directions in Psychological Science*, **23**(1), 3–10.
Gronlund, S., Mickes, L., Wixted, J., and Clark, S. (2015), 'Conducting an eyewitness lineup: how the research got it wrong,'' *Psychology of Learning and Motivation*, **63**, 1–43.
Lindsay, R., and Wells, G. L. (1985), 'Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation,' *Journal of Applied Psychology*, **70**(3), 556.
Loftus, E. F., and Palmer, J. (1996), *Eyewitness Testimony*, New York: Springer.
Mickes, L., Flowe, H. D., and Wixted, J. T. (2012), 'Receiver operating characteristic analysis of eyewit- ness memory: comparing the diagnostic accuracy of simultaneous versus sequential lineups,' *Journal of Experimental Psychology: Applied*, **18**(4), 361–376.
National Research Council (2014), *Identifying the Culprit: Assessing Eyewitness Identification*, Washington, DC: The National Academies Press.
Pepe, M. S. (2000), 'Receiver operating characteristic methodology,' *Journal of the American Statistical Association*, **95**(449), 308–311.
Steblay, N. (1997), 'Social influence in eyewitness recall: a meta-analytic review of lineup instruction effects,' *Law and Human Behavior*, **21**(3), 283–297.
Steblay, N., Dysart, J., Fulero, S., and Lindsay, R. C. (2001), 'Eyewitness accuracy rates in sequential and simultaneous lineup presentations: a meta-analytic comparison,' *Law and Human Behavior*, **25**(5), 459–473.
Steblay, N. K., Dysart, J. E., and Wells, G. L. (2011), 'Seventy-two tests of the sequential lineup superiority effect: A meta-analysis and policy discussion,' *Psychology, Public Policy, and Law*, **17**(1), 99–139.
Steblay, N. K., Wells, G. L., and Douglass, A. B. (2014), 'The eyewitness post identification feedback effect 15 years later: theoretical and policy implications,' *Psychology, Public Policy, and Law*, **20**(1), 1–18.
Wasserman, L. (2003), *All of Statistics: A Concise Course in Statistical Inference*, New York: Springer Science.
Wells, G. L., and Brewer, N. (2006), '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.
Wells, G. L., Smalarz, L., and Smith, A. M. (2015a), 'ROC analysis of lineups does not measure underlying discriminability and has limited value,' *Journal of Applied Research in Memory and Cognition*, **4**(4), 313–317.
Wells, G. L., Smalarz, L., and Smith, A. M. (2015b), 'ROC analysis of lineups obscures information that is critical for both theoretical understanding and applied purposes,' *Journal of Applied Research in Memory and Cognition*, **4**(4), 324–328.
Wetmore, S. A., Neuschatz, J. S., Gronlund, S. D., Wooten, A., Goodsell, C. A., and Carlson, C. A. (2015), 'Effect of retention interval on showup and lineup performance,' *Journal of Applied Research in Memory and Cognition*, **4**, 8–14.
Whittaker, J. (2009), *Graphical Models in Applied Multivariate Statistics*, New Jersey: Wiley Publishing.
Wixted, J. T., and Mickes, L. (2014), 'A signal-detection-based diagnostic-feature-detection model of eyewitness identification,' *Psychological Review*, **121**(2), 262–276.

Appendix: Model selection

We choose to represent the log-linear model using a graph of four nodes, one for each variable. If two nodes are connected by an edge, we have included an interaction term for those two variables in the model.

We implement a graphical approach to model selection. We begin by including all possible edges in the model, removing one edge at a time to evaluate the edges that are necessary to be included. We then add back the removed edges to see if any added parameters will significantly improve the model (See Agresti (2002); Bishop *et al.* (2007); Whittaker (2009); Fienberg (2007) for additional explanation of log-linear model selection).

We restrict possible models to hierarchical models of two-term interactions and lower. We first implement the unconditional edge exclusion test. In this step, we begin with the model including all two-factor interactions. We decide if the model provides a suitable fit using the likelihood ratio statistic, $G^2 = 2^{'},\ x\log\left(\frac{x}{n}\right)$, where x is the observed value and $m\hat{}$ is the fitted value. If the G^2 statistic falls within an acceptable range, with $p > 0.05$, it serves as the reference point for the remaining steps in the model selection process. If the G^2 statistic is not within the acceptable range, we must examine the three-factor interaction terms. Provided the G^2 statistic is acceptable, we then 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 this 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.

We then perform the conditional edge inclusion test, which tells us which additional interactions will lead to a significant increase in goodness of fit. The G^2 statistic from the model resulting from the unconditional edge exclusion test is computed and serves as the reference statistic for the 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 significantly better model fit, that edge is included in the final model. This process is repeated until we have found all conditionally significant edges.

We use the following notation to represent each variable: (1) To represent witness choice (possible values: suspect ID, filler ID and reject line-up); (2) to represent target status (target-absent or target-present); (3) to represent line-up type (biased or unbiased instruction); and (4) to represent ECL (0,10,20,. . .,90,100).

The results of the unconditional edge exclusion test are shown in Table A1. The entries in the table with significant p-values correspond to (1, 2), (1, 4) edges. This means that removing any one of the

TABLE A1. *Unconditional exclusion test results*

Excluded	G^2	p-value	df
(1,2)	186.38	0.00	74.00
(1,3)	62.18	0.83	74.00
(1,4)	118.01	0.04	92.00
(2,3)	53.36	0.96	73.00
(2,4)	62.17	0.95	82.00
(3,4)	60.46	0.96	82.00

TABLE A2. *Conditional edge inclusion test*

Edge	ΔG	Δdf	*p*-value
(1,3)	10.09	2.00	0.01
(2,3)	0.01	1.00	0.91
(3,4)	8.32	10.00	0.60
(2,4)	8.86	10.00	0.55

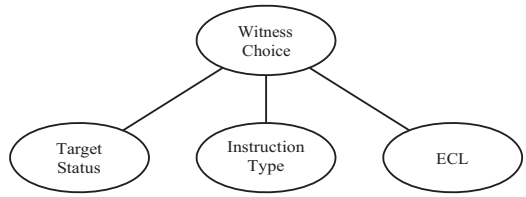

FIG. A1. Final graphical model results, indicating that witness choice interacts with three other variables, but no additional interactions are needed.

interactions between witness choice (1) and either target status (2) or ECL (4) will lead to a model with an insufficient fit.

As seen in Table A2, the conditional edge inclusion test concludes adding the (1, 3) edge back into the model (between witness choice and instruction type) significantly improves the fit. A second conditional edge inclusion test found no further edges to be significant. The resulting graphical model is shown in Fig. A1 and yielded the following goodness of fit results: $G^2 = 69.374$, $df = 93$, $p = 0.96$.