Behavior Research Methods (2021) 53:278–300
https://doi.org/10.3758/s13428-020-01402-7

sdtlu: An R package for the signal detection analysis of eyewitness lineup data

Andrew L. Cohen[1] · Jeffrey J. Starns[1] · Caren M. Rotello[1]

Published online: 22 July 2020

Abstract

In a standard eyewitness lineup scenario, a witness observes a culprit commit a crime and is later asked to identify the culprit from a set of faces, the lineup. Signal detection theory (SDT), a powerful modeling framework for analyzing data, has recently become a common way to analyze lineup data. The goal of this paper is to introduce a new R package, `sdtlu` (Signal Detection Theory – LineUp), that streamlines and automates the SDT analysis of lineup data. `sdtlu` provides functions to process lineup data, determine the best-fitting SDT parameters, compute model-based performance measures such as area under the curve (AUC) and diagnosticity, use bootstrapping to determine uncertainty intervals around these parameters and measures, and compare parameters across two different data sets. The package incorporates closed-form solutions for both simultaneous and sequential lineups that allow for model-based analyses without Monte Carlo simulation. Show-ups are also supported. The package can estimate the base-rate of lineups that include a guilty suspect when the guilt or innocence of each suspect in the data set is unknown, as in "real-world" lineups. The package can also produce a full set of graphs, including data and model-based ROC curves and the underlying SDT model.

Keywords Eyewitness lineups · Computational modeling · Signal detection · R package

In a typical eyewitness lineup scenario, a witness observes a culprit commit a crime. After some delay, the witness is asked to identify the culprit from a set of faces, the lineup. One of the faces, the *suspect*, is the person the police believe committed the crime. The other *filler* faces, of people known to have not committed the crime, are typically selected to be similar to either the suspect (Wogalter, Malpass, & McQuiston, 2004) or a description of the culprit (Wells, Rydell, & Seelau, 1993; Tunnicliff & Clark, 2000). Lineups can vary in length. In the minimal lineup, a show-up, the suspect is shown without any fillers (e.g., Gonzalez, Ellsworth, & Pembroke, 1993). Lineups commonly contain six to eight faces, including the suspect (Police Executive Research Forum, 2013), although other lengths are possible (e.g., Levi, 2012). For lineups with at least one filler, the faces can be shown to the witness in different ways. The two most commonly employed methods are *simultaneous* presentations, in which the witness can view all of the faces at the same time, and *sequential* presentations, in which the witness views the faces one at a time (Police Executive Research Forum, 2013). Although variations exist (e.g., Horry, Brewer, Weber, & Palmer, 2015; Wells, 2014), when discussing sequential lineups, we assume here that the witness can only view each face once and the lineup ends after an identification is made. Regardless of the lineup specifics, the witness identification (ID) falls into one of three broad categories: *suspect*, in which the witness selects the suspect, *filler*, in which the witness selects one of the filler faces, if available, and *reject* or *no identification*, in which the witness does not identify any lineup member as the culprit.

To emphasize the obvious, the suspect and the culprit are not necessarily the same person (e.g., Wells & Olson, 2003). That is, the suspect may not have committed the crime. Thus, a suspect ID has the potential to endanger an innocent person, especially given the weight juries tend to place on eyewitness testimony (Brewer & Burke, 2002; Cutler, Penrod, & Dexter, 1990). It is therefore vital to make every effort to assess the accuracy of a suspect ID. One path forward is to collect confidence ratings. Although eyewitness confidence was initially disregarded as uninformative (e.g., Bothwell, Deffenbacher, & Brigham 1987; Deffenbacher, 1980), recent evidence

✉ Andrew L. Cohen
alc@umass.edu

[1] Department of Psychological and Brain Sciences, University of Massachusetts, 135 Hicks Way, Amherst, MA 01003-7710, USA

strongly suggests that there is a tight relationship between the confidence with which a witness makes an identification and the accuracy of that identification (e.g., Mickes, 2015; Wixted, Mickes, Dunn, Clark, & Wells, 2016; Wells, Yang, & Smalarz, 2015), in particular, a high confidence suspect ID is a good indicator of guilt. Designated fillers are particularly useful because, unlike suspects who can be guilty or innocent, fillers are known to be innocent and so a filler identification is known to be an incorrect response (Wells & Turtle, 1986).

It is also important to know the probability that a suspect is actually guilty before the witness ID is determined, that is, how often the culprit is put into the lineup (Wells et al., 2015). For example, Wells et al. (2015) thoroughly explored the impact of base rate and how it interacts with a number of other variables commonly investigated by eyewitness memory researchers, and they make a convincing argument that base rate is an important factor to consider when interpreting lineup identifications. At one extreme, in a police precinct in which the suspect is almost never guilty, even a highly confident suspect ID is not a strong indication of guilt. At the other extreme, a suspect ID in a precinct that almost always finds the culprit should be given considerable weight, regardless of confidence. In experimental settings, this *base rate*, i.e., the probability of a guilty suspect, is under experimenter control. Outside of the lab, however, the base rate is unknown, and so must be estimated (Cohen, Starns, Rotello, & Cataldo, 2020; Wixted et al., 2016).

All of these experimental factors and statistical concerns can be naturally addressed using signal detection theory (SDT), which has become a common modeling framework to analyze lineup data within the last decade (e.g., Mickes, Flowe, & Wixted, 2012; Wixted & Mickes, 2012; Dobolyi & Dodson, 2013; Carlson & Carlson, 2014; Wetmore, Neuschatz, Gronlund, Wooten, Goodsell, & Carlson, 2015; Wixted et al., 2016; Colloff, Wade, Wixted, & Maylor, 2017). Such analysis can be difficult, and often relies on numerical simulations. The goal of this paper is to introduce a new R package, `sdtlu`, that streamlines and automates the SDT analysis of lineup data. Furthermore, the package implements closed-form solutions for both simultaneous and sequential lineups that allow for model-based analyses.

In what follows, we first briefly describe signal detection theory and how it can be applied to lineup data. We then introduce `sdtlu` and illustrate its capabilities through a series of examples.

Signal detection theory for lineups

Within SDT, a lineup identification is conceived of as a memory task. The witness is asked to match a memory representation of a culprit to a currently viewed face in the lineup. Consider Fig. 1. The *x*-axis represents the strength of match between the culprit memory representation and an individual face in the lineup. The higher the value, the better the match. As represented by the two distributions, this match is generally higher for target faces, the culprit or guilty suspect, than to lure faces, either a filler or innocent suspect. The lure and target distributions are assumed to be normally distributed with means μ_l and μ_t, respectively, and standard deviations σ_l and σ_t, respectively. Differentiation between targets and lures increases with the distance between μ_l and μ_t, as measured in σ_l and σ_t units. The application of SDT to both simultaneous and sequential lineups relies on this basic framework, but differs in how a choice is determined.

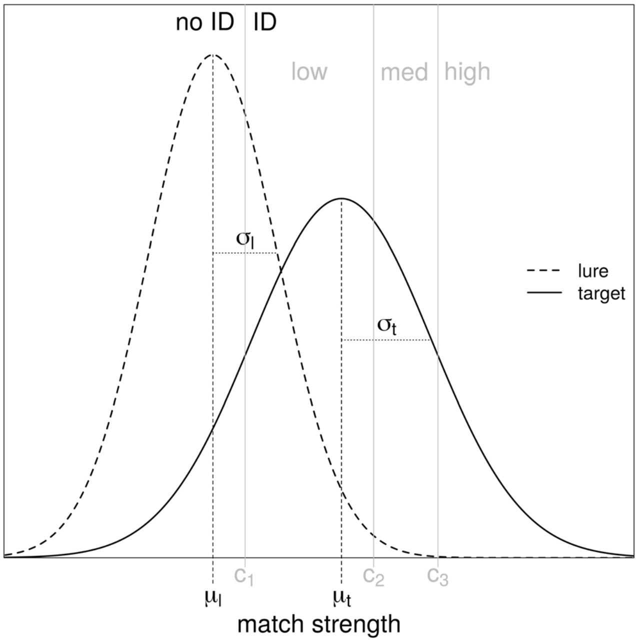


Fig. 1 Signal detection model of lineup responses. The *x*-axis represents the strength of association between the culprit and either a lure or target item. Lure and target strengths are assumed to be normally distributed with means μ_l and μ_t, respectively, and standard deviations σ_l and σ_t, respectively. With three confidence levels, low, medium, and high, there are three response criteria, c_1, c_2, and c_3

First, consider a simultaneous lineup with six faces. (This lineup size is not in any way critical to our conclusions, but is useful for expository purposes.) The basic idea is that a match strength is independently sampled for each of the six faces from the appropriate distribution in Fig. 1 and the highest sampled strength $s*$ from face $f*$ determines the witness's decision. In a *target absent* lineup, the suspect is not the culprit and the match strengths for all six faces are sampled from the lure distribution. In a *target present* lineup, the suspect is the culprit, and the match strengths of the five filler faces are sampled from the lure distribution and the match strength for the suspect face is sampled from the target distribution. The ID is determined by comparing $s*$ to a set of response criteria.

There is one response criterion for each confidence level. Consider the common situation with three confidence levels: Low, medium, and high. If $s*$ is lower than the first criterion or ID threshold, c_1, the witness rejects the lineup, that is, none of the faces was a strong enough match to the culprit to warrant an ID. Any $s*$ above c_1, however, results in an ID. If $s*$ falls between criteria c_1 and c_2, the witness identifies face $f*$ with low confidence. Likewise, an $s*$ between criteria c_2 and c_3 generates a medium confidence ID for $f*$ and an $s*$ above c_3 produces a high confidence ID for $f*$. If $f*$ is a suspect, the ID is a suspect ID (regardless of guilt), and if it is a filler, the ID is a filler ID. Because it is easier to make an ID with a lower criterion, lower criteria are considered more *liberal*. Similarly, because it is more difficult to make an ID with a higher criterion, higher criteria are considered more *conservative*.

Next, consider a sequential lineup with six faces. Recall that in a sequential lineup, the faces are seen exactly once, one at a time, and that the witness must stop after making an ID. Thus, the match strengths are likewise sampled one at a time. If the face is from the culprit, the match strength is sampled from the target distribution, otherwise, the match strength is sampled from the lure distribution. If the current match strength is less than c_1, that face is rejected, the next face is viewed, and the process continues. If the current match strength exceeds c_1, the current face is identified, confidence is determined in the same way as for the simultaneous lineup, i.e., by comparison to the response criteria, and the process ends. If none of the match strengths exceed c_1, the lineup is rejected. Note that, unlike the model of the simultaneous lineup, predictions from the sequential model depend on the order in which faces are presented.[1] For example, a suspect is more likely to be identified if viewed earlier in the lineup because there is less of a chance of a filler ID before the suspect is viewed. This detail is important because suspects are often not equally likely to appear in each position. For example, position is randomly selected in only about 60% of police lineups, and in many experimental designs and about 5% of surveyed agencies, the suspect is never in the first or last position (Police Executive Research Forum, 2013). Thus, to calculate the probability of an ID, this procedure must be run for each possible suspect position with the combined results weighted by the probability of each suspect position.

Numerical simulations are typically used to generate model predictions, i.e., the probability of a rejection or a suspect or filler ID at different confidence levels. As discussed below, however, performing analyses on these models can be computationally demanding. It is therefore useful to be able to rely on closed-form solutions. Such equations are provided in Appendix A for a simultaneous lineup and in Appendix B for a sequential lineup (also see Wixted, Vul, Mickes, & Wilson, 2018). The equations work for lineups of an arbitrary size and an arbitrary number of confidence levels.

As discussed previously, it is important to be able to estimate the base rate p, i.e., the probability that a given lineup includes a guilty suspect. The SDT modeling approach provides a method for estimating this value even when the guilt or innocence of each suspect is unknown (Cohen, Starns, Rotello, & Cataldo, 2020; Wixted, Mickes, Dunn, Clark, & Wells, 2015). In an eyewitness memory experiment, the experimenter knows whether each lineup was target present (i.e., included the guilty suspect) or target absent, what we will call *full* data. In a real-world lineup, what we will call *restricted* data, this classification is exactly the information of interest. In the latter case, the available data are how often each ID was made (suspect, filler, or reject) and the associated confidence of the ID, and these values are collapsed across both target-present and target-absent lineups. As shown in the appendices, calculating the predictions of the SDT model requires specifying the probability that a lineup will include a guilty suspect. For both simultaneous and sequential lineups, whether the lineup is target present or target absent determines whether the suspect match strength is drawn from the target or lure distribution, respectively. The critical insight, therefore, is that this base rate or target present probability becomes a parameter of the model which, in turn, can be estimated from data. Indeed, prior work has shown that such probabilities can be reliably estimated and applied to field data (Wixted et al., 2015). Cohen et al. (2020) extended these results and clarified how the model is able to estimate base rate by evaluating the relative probability and confidence distributions of suspect and filler IDs. Essentially, the model can produce a higher ratio of suspect to filler IDs by either assuming that a higher proportion of lineups have guilty suspects or assuming that witnesses have better memory (and thus are more likely to identify guilty suspects when they are present). These scenarios are distinguishable because higher memory increases confidence levels for suspect IDs to a greater degree than higher base rates.

In summary, to apply the SDT model, the following experimental design parameters are needed: l, the lineup size and n, the number of confidence levels. In addition, for sequential lineups, the model needs to know the probability that the suspect appears in each lineup position. In principle, the full SDT model has the following parameters: μ_l and σ_l, the mean and standard deviation of the lure distribution, respectively; μ_t and σ_t, the mean and standard deviation of the target distribution, respectively; c_1-c_n, the response criteria; and p, the probability that the suspect is guilty. Without loss of generality, it is

[1] We recognize that suspect position can influence identification in both simultaneous and sequential lineups (e.g., Palmer, Sauer, & Holt, 2017). Whereas the sequential model naturally incorporates position effects, the baseline simultaneous model does not, and doing so introduces a host of difficulties because the witness can consider the faces in any order and can return to previously considered faces. To avoid the additional model complexity involved in incorporating position effects into the simultaneous model, we leave that change for future work.

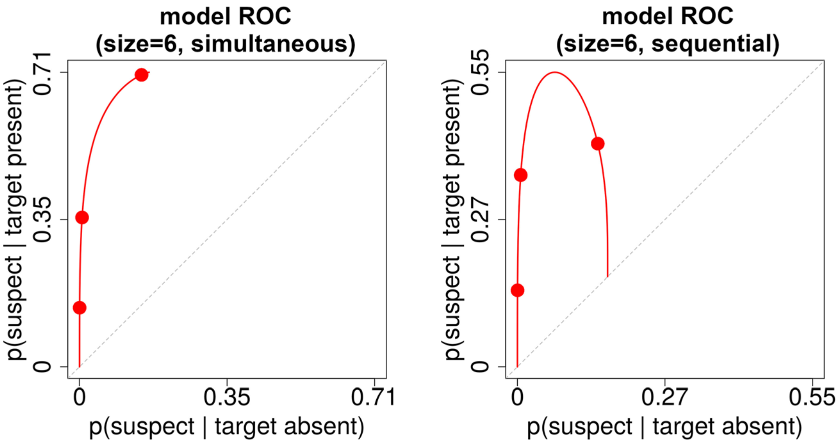


Fig. 2 Example receiver operating characteristic (ROC) curves for a simultaneous (left) and sequential (right) signal detection model

typically assumed that $\mu_l = 0$ and $\sigma_l = 1$. For full data, p can be directly set to the observed proportion of a target-present lineups in the data set. For restricted data, p must be estimated. Unfortunately, in this case, not all of the parameters are identifiable. In particular, it is not possible to estimate both p and σ_t. In this situation, we adopt the solution of estimating p and setting $\sigma_t = \sigma_l = 1$, i.e., assuming an equal-variance SDT model. In either the full or restricted case, the model has $2+n$ parameters. For example, with three confidence levels, the model has five parameters.

Receiver operating characteristic curves

Receiver operating characteristic (ROC) curves are a very convenient and common way to represent SDT data. For lineup data, an ROC curve plots the probability of a suspect ID when the target is present against the probability of a suspect ID when the target is absent at every possible confidence level. Example ROC curves are provided in Fig. 2. The left and right panels display ROC curves for a simultaneous and sequential lineup of size 6 using the SDT parameters from Fig. 1. For the sequential lineup, all suspect positions were assumed equally likely. The *x*- and *y*-axes represent the probability of a suspect ID in a target absent and target present lineup, respectively. The curves were determined from the SDT models described previously by sweeping the ID criteria across a wide range. Being able to easily produce such curves is one advantage of the closed-form solution of the SDT model. The dots represent the probabilities at the three response criteria from Fig. 1. ROC curves that lie on the diagonal represent chance

```
id_type,   conf_level,   culprit_present,   lineup_size
 filler,        2,             absent,            8
 filler,        2,             absent,            8
suspect,        2,            present,            8
 reject,        3,             absent,            8
suspect,        2,             absent,            8
 filler,        2,             absent,            8
 reject,        1,            present,            8
 filler,        3,             absent,            8
 reject,        3,             absent,            8
 filler,        3,             absent,            8
```

Fig. 3 Ten randomly selected rows of the Palmer et al. (2013) data set formatted for `sdtlu`. The whitespace is arbitrary

```
> sdtlu_process_data(data_name='palmer_data.csv')

$resp_data_restr
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]
[1,]  100   93   55   16   13   30   73   88   42    28   370

$resp_data_full
      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10] [,11]
[1,]    98    84    49    12    10    13    25    31    16    13   105
     [,12] [,13] [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22]
[1,]     2     9     6     4     3    17    48    57    26    15   265

$overall_resp_data_restr_prop
 [1] 0.110 0.102 0.061 0.018 0.014 0.033 0.080 0.097 0.046 0.031 0.407

$overall_resp_data_full_prop_joint
 [1] 0.108 0.093 0.054 0.013 0.011 0.014 0.028 0.034 0.018 0.014 0.116 0.002 0.010
0.007 0.004 0.003 0.019 0.053 0.063 0.029 0.017 0.292

$overall_resp_data_full_prop_cond
 [1] 0.215 0.184 0.107 0.026 0.022 0.029 0.055 0.068 0.035 0.029 0.230 0.004 0.020
0.013 0.009 0.007 0.038 0.106 0.126 0.058 0.033 0.586

$lu_sizes
[1] 8

$pos_prop
NULL

$ptp
[1] 0.5022026

$n_trials
[1] 908

$acc_data
[1] 0.980 0.903 0.891 0.750 0.769 0.716

$sus_g_id_data
[1] 0.769 0.560 0.385 0.276 0.317

$n_resp_cats
[1] 11
```

Fig. 4 Output of **sdtlu_process_data** when applied to the Palmer et al. (2013) simultaneous lineup data set. Values were rounded to 3 decimals

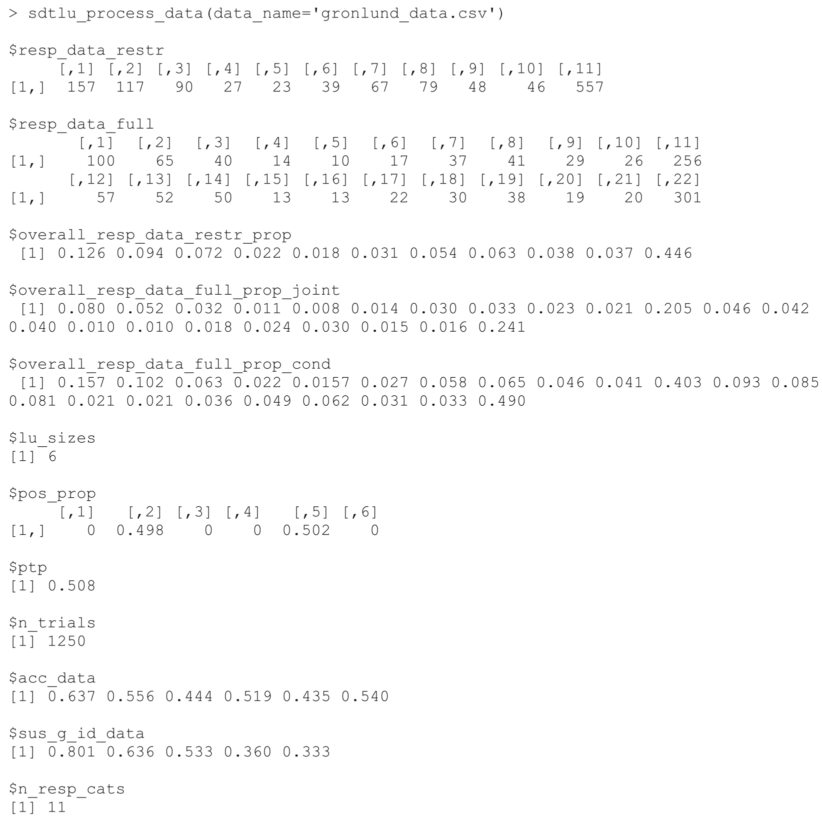
performance at selecting the culprit. Performance improves as the ROC curve moves into the upper left corner, where correct IDs are more likely than errors. As the response criteria become more liberal or conservative, the dots will move up or down the curve, respectively.

Standard ROC curves typically end at the point (1, 1). Because suspect selection when the target is absent is assumed to be at chance, i.e., selected from the lure distribution, ROC curves for lineup data usually do not reach (1, 1), and so are partial ROCs. The simultaneous ROC ends at the inverse of the lineup size, in this case 1/6, which is the probability of selecting the suspect by chance. The ROC for the sequential lineup is somewhat more complex. If the response criterion is extremely liberal, i.e., any face is highly likely to be selected, only a suspect in the first position will realistically be selected. Thus, the sequential ROC curve will end with both the hit and false alarm rates equal to the probability that the suspect appears in the first position (Rotello & Chen, 2016). In this example, that is 1/6. However, if the suspect appears in the first position with probability less than 1/6, which can easily be the case in sequential lineups, the ROC can reverse and can even end at the origin (see Fig. 12).

The ROC curves in Fig. 2 are model-based. Similar, data-based ROC curves are also common. In such plots, the points represent the probabilities of a suspect ID present in the data and are usually simply connected by a straight line. An example is shown in Fig. 7.

Measuring performance

Within the SDT framework, there are a number of ways to measure subject performance. Perhaps the most common performance measure within the SDT framework is d',



```
> sdtlu_process_data(data_name='gronlund_data.csv')

$resp_data_restr
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]
[1,]  157  117   90   27   23   39   67   79   48    46   557

$resp_data_full
       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10] [,11]
[1,]    100    65    40    14    10    17    37    41    29    26   256
      [,12] [,13] [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22]
[1,]     57    52    50    13    13    22    30    38    19    20   301

$overall_resp_data_restr_prop
 [1] 0.126 0.094 0.072 0.022 0.018 0.031 0.054 0.063 0.038 0.037 0.446

$overall_resp_data_full_prop_joint
 [1] 0.080 0.052 0.032 0.011 0.008 0.014 0.030 0.033 0.023 0.021 0.205 0.046 0.042
0.040 0.010 0.010 0.018 0.024 0.030 0.015 0.016 0.241

$overall_resp_data_full_prop_cond
 [1] 0.157 0.102 0.063 0.022 0.0157 0.027 0.058 0.065 0.046 0.041 0.403 0.093 0.085
0.081 0.021 0.021 0.036 0.049 0.062 0.031 0.033 0.490

$lu_sizes
[1] 6

$pos_prop
     [,1]   [,2] [,3] [,4]   [,5] [,6]
[1,]    0  0.498    0    0  0.502    0

$ptp
[1] 0.508

$n_trials
[1] 1250

$acc_data
[1] 0.637 0.556 0.444 0.519 0.435 0.540

$sus_g_id_data
[1] 0.801 0.636 0.533 0.360 0.333

$n_resp_cats
[1] 11
```

Fig. 5 Output of **sdtlu_process_data** when applied to the Gronlund et al. (2009) sequential lineup data set. Values were rounded to 3 decimals. Confidence levels 5-7 were combined

which is the distance between the distribution means in standard deviation units (Macmillan & Creelman, 2005). However, d' is only defined for the equal-variance model, i.e., when the standard deviations of the two distributions are identical. As discussed previously, the equal-variance model in the `sdtlu` package assumes that $\mu_l = 0$ and $\sigma_l = \sigma_t = 1$. Thus, under the equal-variance assumption, d' is identical to the `sdtlu` estimate of μ_t, the mean of the target distribution.

Area under the curve (AUC) is commonly used in the lineup context (Mickes et al., 2012; Rotello & Chen, 2016). The AUC measure is derived from the ROC curve. Recall that the ROC curve lies on the diagonal when performance is at chance and moves into the upper-left corner when performance is good. Thus, a natural way to measure performance is to determine the area under the ROC curve. Higher values mean the curve is farther into the upper-left corner and therefore indicate better performance. Because lineup ROC curves do not reach (1, 1), the AUC measure is technically a partial AUC (McClish, 1989). For simplicity, however, we refer to it as AUC throughout. It is important to note, however, that the AUC is affected by lineup length for a simultaneous lineup (Rotello & Chen, 2016) and the probability of a first-position suspect in a sequential lineup. Thus, AUC is a relative, not absolute, measure and care is required when comparing AUC across experiments with different designs. Analytic solutions do not exist, so AUC is calculated by numerical approximation. In the current work, iterative quadrature is used to compute AUC from model-defined ROC functions. Because the numerical approximation becomes unstable when part of the ROC curve is vertical, AUC should be treated with caution in situations like the right panel of Fig. 2. AUC is not defined when the ROC curve is not monotonic on p(suspect | target absent), as is the case for some sequential lineups.

Diagnosticity is another common measure of lineup performance (Wells & Lindsay, 1980), although it has been shown to be confounded with response bias, leading some to conclude that AUC is a better measure of ID accuracy (Mickes et al., 2012; Rotello & Chen, 2016). Nevertheless, diagnosticity is a critical consideration in evaluating how lineup identifications provide evidence about the guilt or innocence of the suspect (e.g., Wells et al., 2015). Diagnosticity is the ratio of the probability of a suspect ID given target present and target absent lineups. A value of 1 means that guilty and innocent suspects are equally likely to be selected, i.e., chance performance. The higher the value, the more likely a selected suspect is guilty. Diagnosticity can be computed directly from data. Here we take a different approach and use the SDT model to estimate diagnosticity. Diagnosticity can be calculated from the SDT model in two different ways. The first method collapses over all confidence levels. That is, the relevant probabilities are calculated without regard to confidence. The second method restricts calculation within a confidence region. That is, the relevant probabilities are calculated based on confidence levels between two response criteria. All other things being equal, diagnosticity increases with confidence level. Two other, related measures of performance are discussed below in relation to the `sdtlu` package figures.

Equations for AUC and diagnosticity are provided in Appendix C.

The sdtlu package

Overview

The SDT model discussed in the previous section is implemented in the `sdtlu` (Signal Detection Theory - LineUp) R package, which is publicly available at https://osf.io/mfk4e. The package instantiates functions for the preparation and processing of lineup data, fitting SDT models to data, comparing models across two data sets, generating accuracy measures, graphing data and model results, and simulating experiments. In this section, we introduce how to use `sdtlu` for each of these tasks. To illustrate the functionality and use of this package, we walk through a simultaneous lineup example and then a sequential lineup example. All references to the `sdtlu` package, functions, variables, and output files are provided in `Courier font`. For clarity, the package functions are also shown in bold.

To preview, the main functions are **sdtlu_process_data**, which processes lineup data, **sdtlu_fit**, which fits the SDT model to lineup data, and **sdtlu_compare_2**, which uses bootstrapping to compare parameters from the SDT model across two data sets. The functions **sdtlu_sim_sim** and **sdtlu_seq_sim** can be used to simulate simultaneous and sequential SDT lineup data, respectively.

Data

As an example of a simultaneous lineup study, we will use Palmer, Brewer, Weber, and Nagesh (2013), which was also the data set used by Wixted et al. (2016). This is a field study in which 908 participants were asked to identify a culprit from a simultaneous, eight-person lineup in public. Approximately 50% of the lineups were target present and 50% were target absent. Participants provided a confidence rating on an 11-point scale, which the authors collapsed onto a five-point scale by combining some ratings categories. For simplicity, we ignore the other manipulated factors including exposure time and delay, although we revisit delay below. For the target absent trials, we randomly selected one of the eight faces to act as the innocent suspect.

For a sequential data set, we use Gronlund, Carlson, Dailey, and Goodsell (2009). This was a joint lab and online study in which subjects watched a video of a crime and then were asked to ID the culprit from a 6-person lineup and rate their confidence on a 1–7 scale. Because suspect IDs were rarely, if ever, made at the lowest confidence levels, some measures were undefined and the model fits were unstable. To demonstrate the full functionality of the package, we collapsed confidence levels 5, 6, and 7. Thus, there are five confidence levels in the analyzed data. The authors designated innocent suspects. Suspects (both guilty and innocent) were only tested in positions 2 and 5 of the lineups. We ignore all other factors including view quality, lineup bias, and suspect similarity to culprit. This data set includes 1250 trials.

Data format

For use in `sdtlu`, the data must be a comma-separated (csv) file with the following columns: `id_type` is the participant's ID (`suspect/filler/reject`); `conf_level` the participant's confidence level (e.g., `1,2,3,4,5`), where `1` is the highest confidence level; `culprit_present` is whether the lineup was target present or absent (`present/absent`); and `lineup_size` is how many individuals were in the lineup. Ten randomly selected rows of the Palmer et al. (2013) data set are shown in Fig. 3. Sequential data would have an additional column, `suspect_position`, which provides the position of the suspect (an integer from `1` to `lineup_size`).

Processing data

Raw lineup data, like that shown in Fig. 3, can be processed using the function **sdtlu_process_data**. You can pass the data either as a file name, with the file formatted as described previously, as is done in this example, or as a data frame with the same variables present in a data file. A sample call and output are shown in Fig. 4. We describe this output next.

`resp_data_restr` provides the response counts when collapsed across target present and target absent conditions. This output is useful for real-world studies when it is unknown whether the suspect was guilty, and so it is unknown whether a datum is from a target present or target absent condition. The output is in the following order: suspect ID highest confidence, …, suspect ID lowest confidence, filler ID highest confidence,

```
> sim_data <- sdtlu_process_data('palmer_data.csv')
> options <- list(model_type='sim',
+                 fit_fcn='G2',
+                 fix_p='data', fix_sigma_t='free',
+                 use_restr_data=FALSE,
+                 n_bootstrap_samps=50)
> sdtlu_fit(data=sim_data, options=options)

quantiles = (0.01, 0.025, 0.05, 0.25, 0.5, 0.75, 0.95, 0.975, 0.99)

----------

G2 = 8.601 (8.172, 8.539, 10.043, 17.821, 22.352, 26.765, 35.724, 44.255, 46.86)
p (fixed) = 0.502 (0.502, 0.502, 0.502, 0.502, 0.502, 0.502, 0.502, 0.502, 0.502)
mu_t = 1.812 (1.643, 1.688, 1.704, 1.75, 1.815, 1.853, 1.9, 1.93, 1.952)
sigma_t = 0.989 (0.8, 0.854, 0.864, 0.916, 0.994, 1.063, 1.12, 1.163, 1.233)
c1 = 1.511 (1.439, 1.443, 1.45, 1.484, 1.503, 1.525, 1.557, 1.56, 1.567)
c2 = 1.593 (1.517, 1.524, 1.539, 1.567, 1.583, 1.607, 1.634, 1.638, 1.642)
c3 = 1.712 (1.636, 1.645, 1.652, 1.683, 1.706, 1.731, 1.754, 1.758, 1.763)
c4 = 2.035 (1.945, 1.954, 1.959, 2.006, 2.02, 2.053, 2.088, 2.095, 2.11)
c5 = 2.579 (2.442, 2.446, 2.46, 2.511, 2.566, 2.626, 2.692, 2.716, 2.724)

mu_f (fixed) = 0
sigma_f (fixed) = 1

----------

area under the curve (lineup size)
auc (8) = 0.066 (0.059, 0.061, 0.062, 0.064, 0.066, 0.068, 0.071, 0.072, 0.074)

diagnosticity collapsed across confidence level (lineup size)
diag (8) = 10.548 (8.875, 9.334, 9.357, 9.853, 10.329, 10.908, 11.838, 11.952, 12.109)

diagnosticity by confidence level (lineup size, confidence level)
diag (8,5) = 3.257 (2.574, 2.693, 2.78, 2.95, 3.175, 3.476, 3.914, 3.975, 4.218)
diag (8,4) = 3.908 (3.062, 3.179, 3.283, 3.564, 3.797, 4.143, 4.718, 4.746, 5.1)
diag (8,3) = 5.789 (4.648, 4.784, 4.929, 5.277, 5.711, 6.036, 6.807, 6.996, 7.471)
diag (8,2) = 12.113 (9.539, 9.666, 10.089, 11.216, 11.822, 12.63, 14.07, 14.32, 14.55)
diag (8,1) = 44.49 (29.247, 29.584, 30.379, 37.382, 43.93, 49.93, 63.77, 64.50, 70.25)

----------

Data and best-fitting model predictions (lineup size)
(Conditioned on target present/absent)

Data  (8)  0.215 0.184 0.107 0.026 0.022 0.029 0.055 0.068 0.035 0.029 0.230 0.004
0.020 0.013 0.009 0.007 0.038 0.106 0.126 0.058 0.033 0.586
Model (8)  0.217 0.177 0.104 0.033 0.021 0.029 0.069 0.067 0.027 0.019 0.237 0.005
0.015 0.018 0.008 0.006 0.034 0.102 0.126 0.059 0.045 0.582

----------
```

Fig. 6 Partial output from **sdtlu_fit** with the Palmer et al. (2013) simultaneous lineup data

…, filler ID lowest confidence, reject. In all inputs and outputs, confidence responses range from the highest confidence level to the lowest confidence level, left to right. Because this data set has five confidence levels, there are 11 counts (five suspect, five filler, and one reject[2]). For example, 100 participants identified the suspect with the highest confidence level, regardless of whether the suspect was the culprit or not.

`resp_data_full` provides the response counts categorized by target absent and target present condition if those data are available, as they are in this data set and most experimental work. The counts are in the same order as for `resp_data_restr`, but are shown twice, first for target present trials and then for target absent trials. For example, 98 participants identified the suspect at the highest confidence level, when the suspect was the culprit, and two participants identified the suspect at the highest confidence level, when the suspect was not the culprit.

[2] Confidence ratings for rejections are not used because it is unclear how one should make these confidence ratings and because confidence is often fairly unsystematic for rejections in data sets.

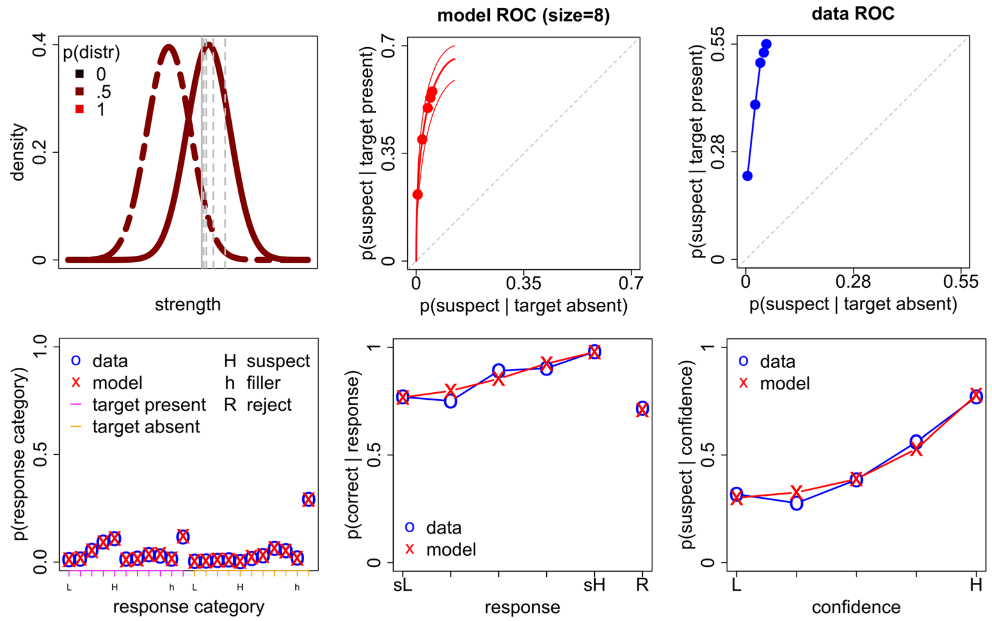


Fig. 7 Figure generated from **sdtlu_fit** with the Palmer et al. (2013) simultaneous lineup data. The figure layout has been modified to remove whitespace

Although we do not show an example here, lineups of different lengths can be processed together. This functionality is useful for meta-analyses in which data from multiple experiments are combined. When multiple lineup sizes are present, `resp_data_restr` and `resp_data_full` provide one row of counts per lineup size.

The next three outputs provide the proportion of response categories, rather than counts. If multiple lineup sizes exist, they collapse across the lineup sizes, i.e., lineup size is ignored. If there is only a single lineup size, the proportions are for that lineup size. `overall_resp_data_restr_prop` shows `resp_data_restr`, but as a proportion. For example, 11% of participants, i.e., 100/908, provided a high confidence suspect ID. `overall_resp_data_full_prop_joint` and `overall_resp_data_full_prop_cond` provide the count data from resp_data_full as proportions. The former shows them as joint probabilities, e.g., $P(resp = sus \cap conf = i \cap tar = pres)$, and the latter as conditional probabilities, e.g., $P(resp = sus \cap conf = i | \ tar = pres)$. For example, from `overall_resp_data_full_prop_joint` we learn that 11% of trials were high confidence, target present, suspect IDs and from `overall_resp_data_full_prop_cond` we learn that 22% of target present trials were high confidence suspect IDs.

`lu_sizes` is a vector of the lineup sizes in the data set. In the Palmer et al. (2013) example data, lineups were only of size 8.

`pos_prop` provides the proportion of times the suspect was present in each lineup position. The distributions of suspects over positions was assumed to be the same for both guilty and innocent suspects. This output is only relevant to sequential lineups and will be discussed below.

`ptp` is the base rate, i.e., the probability of a target present lineup, if known. In this case, 50.2% of lineups were target present.

`ntrials` is the number of data points, trials, or participants. This data set had 908 participants.

`acc_data` provides a measure of accuracy, i.e., the proportion of correct IDs for suspect picks and rejects (e.g., Wixted et al., 2016). This measure is similar to diagnosticity. There is one proportion for each confidence level (five in this example, starting with the highest confidence level) and one for rejections. For suspect IDs, this measure is $P(tar = pres|resp = sus \cap conf = i)$. For rejects, this measure is $P(tar = abs|resp = rej)$. For example, out of all high confidence suspect IDs, 98% were actually the culprit, and out of all rejections, the culprit was not present 72% of the time. This measure is only available for full data sets.

`sus_g_id_data` provides the probability of a suspect ID at a given confidence level, excluding rejections (e.g., Wixted et al., 2016). Each value is taken at a single confidence level (starting with the highest). For example, out of all high confidence IDs, 77% were suspect IDs.

`n_resp_cats` shows the number of possible response categories. With five confidence levels, this data set had 11 response categories (five suspect + five filler + one reject).

The output for the sequential Gronlund et al. (2009) study is shown in Fig. 5. The output format is identical to Fig. 4. Because suspect position was specified, we

```
> sim_data <- sdtlu_process_data(data_name='palmer_data.csv')
> options <- list(model_type='sim',
+                 fit_fcn='G2',
+                 fix_p='free', fix_sigma_t='sigma_f',
+                 use_restr_data=TRUE,
+                 n_bootstrap_samps=50)
> sdtlu_fit(data=sim_data, options=options)

quantiles = (0.01, 0.025, 0.05, 0.25, 0.5, 0.75, 0.95, 0.975, 0.99)

----------

G2 = 1.485 (0.406, 0.466, 0.595, 2.046, 3.512, 6.659, 9.76, 13.28, 16.965)

p = 0.509 (0.406, 0.432, 0.435, 0.46, 0.486, 0.537, 0.663, 0.67, 0.727)
mu_t = 1.81 (1.376, 1.439, 1.556, 1.743, 1.864, 1.919, 1.985, 2.005, 2.056)
sigma_t (fixed) = 1 (1, 1, 1, 1, 1, 1, 1, 1, 1)
c1 = 1.511 (1.451, 1.457, 1.465, 1.489, 1.508, 1.528, 1.555, 1.564, 1.57)
c2 = 1.594 (1.536, 1.539, 1.548, 1.566, 1.593, 1.612, 1.641, 1.641, 1.652)
c3 = 1.712 (1.659, 1.661, 1.664, 1.689, 1.714, 1.727, 1.763, 1.767, 1.78)
c4 = 2.039 (1.967, 1.977, 1.982, 2.021, 2.043, 2.059, 2.083, 2.098, 2.107)
c5 = 2.588 (2.487, 2.494, 2.512, 2.567, 2.591, 2.616, 2.668, 2.702, 2.727)

mu_f (fixed) = 0
sigma_f (fixed) = 1

----------

area under the curve (lineup size)
auc (8) = 0.066 (0.047, 0.05, 0.055, 0.063, 0.069, 0.071, 0.074, 0.075, 0.077)

diagnosticity collapsed across confidence level (lineup size)
diag (8) = 10.527 (7.272, 7.583, 8.65, 9.939, 10.723, 11.451, 12.16, 12.286, 12.317)

diagnosticity by confidence level (lineup size, confidence level)
diag (8,5) = 3.229 (2.793, 2.878, 2.919, 3.038, 3.185, 3.311, 3.522, 3.539, 3.547)
diag (8,4) = 3.873 (3.428, 3.531, 3.543, 3.657, 3.828, 3.944, 4.193, 4.244, 4.292)
diag (8,3) = 5.751 (4.837, 5.024, 5.232, 5.476, 5.762, 5.893, 6.186, 6.293, 6.488)
diag (8,2) = 12.134 (8.158, 8.621, 10.164, 11.281, 12.36, 13.202, 14.20, 14.54, 14.65)
diag (8,1) = 45.516 (20.927, 23.655, 30.704, 40.07, 49.05, 54.69, 65.38, 68.02, 74.89)

----------

Data and best-fitting model predictions (lineup size)

Data  (8)  0.110 0.102 0.061 0.018 0.014 0.033 0.080 0.097 0.046 0.031 0.407
Model (8)  0.112 0.097 0.062 0.021 0.014 0.031 0.085 0.097 0.043 0.032 0.407

----------
```

Fig. 8 Partial output from **sdtlu_fit** with the Palmer et al. (2013) restricted simultaneous lineup data

learn that approximately 50% of the suspects were shown in positions 2 and 50% in postion 5.

Fitting SDT data

Perhaps the most fundamental aspect of sdtlu is its ability to fit the SDT model to experimental data. Fitting is done using the **sdtlu_fit** function. This function takes as arguments experimental data and a set of options.

The data can be in one of three forms. First, the data can be the output from **sdtlu_process_data**. This method is used in the following example. Second, you can provide a file name and **sdtlu_fit** will then internally analyze the data using **sdtlu_process_data**. Third, the data can be a matrix of count data of each response at each confidence level, of the same form as `resp_data_full` or `resp_data_restr`. When count data are provided, the function requires the lineup size, i.e., the number of people in the lineup. An example of this third method is also provided below.

There are numerous options available, which are grouped in a list. `model_type` specifices a simultaneous (‘`sim`’, default) or sequential (‘`seq`’) SDT model. `fit_fcn` allow you to select a method for calculating the fit of the model to the data. The default is to use G^2 (‘`G2`’), but χ^2 (‘`chi-square`’) is also available[3]. If use_restr_data is TRUE, the data are assumed to be collapsed over target present and absent, as in resp_data_restr. If FALSE (the default), target present and absent are separated, as in resp_data_full. n_fits tells the function how many times to restart the parameter search with a different starting parameter set, which can be useful for avoiding local minima. save_file_name (sdtlu_save.RData, default) and fig_file_name (sdtlu_figs.png, default) allow you to specify where the numeric and graphic outputs are stored.

The arguments fix_p and fix_sigma_t control the behavior of the parameters for the proportion of target present trials p and the standard deviation σ_t of the target distribution. If set to 'free', the associated parameters are free to vary. These arguments can also be set to a value, which fixes the associated parameter to that value. If fix_p is 'data', and the base rates are known, p (also referred to as *ptp*, probability of target present trials, depending on context) is set to the proportion of target present trials in the data (for restricted data, p is allowed to vary). If fix_sigma_t is set to 'sigma_f' [4], then $\sigma_t = \sigma_l = 1$. Recall that, for restricted data, it is not possible to simultaneously estimate the proportion of target present trials p and the standard deviation σ_t of the target distribution. Thus, when restricted data are used, fix_p and fix_sigma_t should not both be set to 'free'. For restricted data, the defaults are fix_p = 'free' and fix_sigma_t='sigma_f'. For full data, the defaults are fix_p='data' and fix_sigma_t='free'.

To provide bounds on the fit, parameter, and accuracy values, **sdtlu_fit** can also bootstrap the data. That is, the data are resampled with replacement (within target present/absent conditions, if available), the fit is repeated on this new sample, and all measures are recalculated. Bounds are then determined from quantiles on these samples. The bootstrap is run if run_bootstrap is TRUE (the default). If you only want point estimates, it is faster to set run_bootstrap to FALSE. The number of bootstrap samples is determined by n_bootstrap_samps. Bootstrapping can be slow, so we suggest trying it with a small n_bootstrap_samps first before running a longer version. The default is 1000 samples. The function returns the .01, .025, .05, .025, .5, .75, .95, .975, and .99 quantiles from the bootstrapped samples. If all samples are required, not just the quantiles, set output_bootstrap_samps to TRUE (default is FALSE). Only one fit is run per bootstrap, with starting parameters determined by the best-fitting parameters to the non-bootstrapped data. If the bootstrap creates an error in the fit function, e.g., the sample is impossible under the model, a warning is provided, and a new sample is drawn.

A sample run for the Palmer et al. (2013) simultaneous data is provided in Fig. 6 along with partial output. First, the data are processed, as in Fig. 4. The options are then set. For this example, we set many of the options to their default value, but for convenience, we only use 50 bootstrap samples. In general, we suggest at least 1000 bootstrap samples.

The output provides means and bootstrapped quantiles for the fit measure, all model parameters, AUC, and diagnosticity. The output also provides the data (in the same order as resp_data_full) and the best-fitting model prediction. Left out of the figure are the fit-by-fit fit measures, the full set of bootstrapped results produced, and the function options. These results are stored, by default, in sdtlu_save.RData in the current directory.

The function also produces a set of figures as shown in Fig. 7, also saved in the current directory. The upper-left panel shows the SDT model as in Fig. 1 with the best fitting model parameters. The color/brightness of the distribution lines represent the probability of a target present or absent sample. Because the base rate is near .5, the two colors are indistinguishable for this example. The upper-center panel shows the model-based ROC as in Fig. 2. The thinner lines are generated from the parameter sets that produced AUC values at the .05 and .95 quantiles of the distribution across bootstrap samples, and thus provide a form of equal-tailed interval on the ROC curve. The upper-right panel shows the data ROC. The bottom-left panel shows how well the model fits the data and provides both the data and model predicted proportions for each response category for each condition. The bottom-center panel provides the data and predicted model accuracy, i.e., the probability of a correct response given either a suspect ID at the different confidence levels or a rejection. The bottom-right panel shows the data and model predicted probability of a suspect ID at different confidence levels. To be consistent with previous work, in the latter three graphs, confidence increases from left to right. The model does an overall good job of accounting for the experimental data, both qualitatively and quantitatively.

[3] The G^2 method provides maximum likelihood estimates of parameter values, as G^2 is determined by the likelihood of the data under a given parameter set in the model of interest (the "research" model) and the likelihood of the data in a "full" model in which the probability of each response category matches its proportion in the data. The latter value does not change for different parameter sets, so G^2 differences across parameter sets are determined only by the likelihood of the "research" model.

[4] Although this should technically be sigma_l, we use sigma_f to avoid confusion with sigma_1.

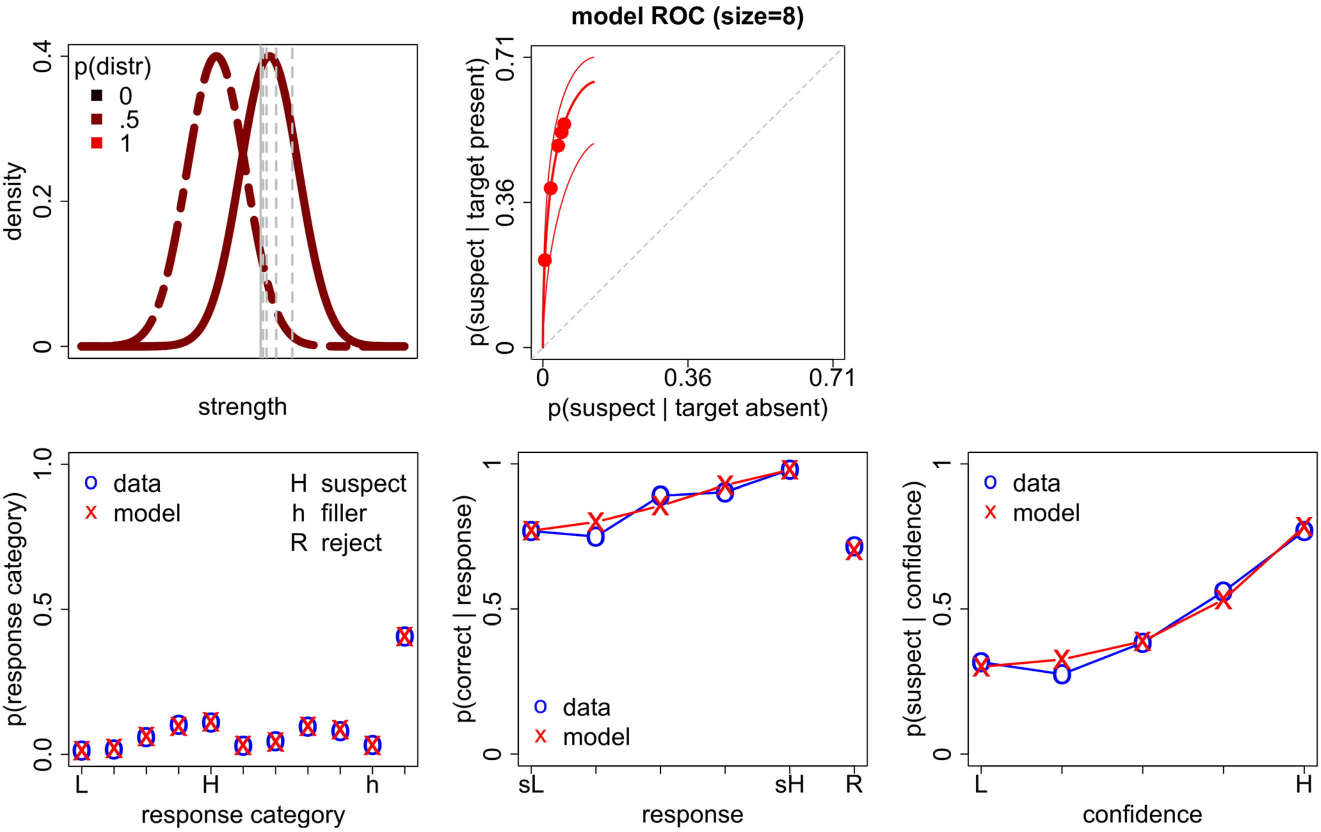


Fig. 9 Figure generated from **sdtlu_fit** with the Palmer et al. (2013) restricted simultaneous lineup data. The figure layout has been modified to remove whitespace

We can use the same data to simulate restricted data, i.e., real-world data in which the proportion of target present trials is not known. To do this, we simply collapse the counts at a confidence level across target present and absent conditions. In the **`sdtlu_fit`** function, you can do that by setting the `use_restr_data` option argument to `TRUE`. Now p is free to vary, and we set $\sigma_t = \sigma_l = 1$. Example output is provided in Fig. 8, and the associated figures are provided in Fig. 9. Note that the estimated value of *p,* .509, is very close to the actual value from Fig. 4 of .502. Because it is impossible to separate the data into target present and absent categories for restricted data, a data ROC curve cannot be drawn.

As mentioned previously, **`sdtlu_fit`** can also be applied directly to counts. This application of the function is useful when fitting to previously analyzed data, data from tables, or simulated data. We provide a sample call to the Palmer et al. (2013) restricted in Fig. 10. Because these are restricted data, the counts are in the same order as `resp_data_restr`, full data would be in the same order as `resp_data_full`. Because these are count data, the lineup size now needs to be specified as an argument. The output is identical to Figs. 8 and 9.

```
> sim_count_data <- c(100, 93, 55, 16, 13, 30, 73, 88, 42, 28, 370)
> options <- list(model_type='sim',
+                 fit_fcn='G2',
+                 fix_p='free', fix_sigma_t='sigma_f',
+                 use_restr_data=TRUE,
+                 n_bootstrap_samps=50)
> sdtlu_fit(data=sim_count_data, lineup_sizes=c(8), options=options)
```

Fig. 10 Call to **sdtlu_fit** for counts from Palmer et al. (2013) restricted simultaneous lineup data

We can also apply **sdtlu_fit** to sequential data. This function was applied to the Gronlund et al. (2009) data. The call and output are shown in Fig. 11 and the associated figures are in Fig. 12. The suspect position proportions, which are needed for a sequential model, are included in the processed data, as was shown in Fig. 5[5]. In our experience, the parameter search for the sequential model is more difficult. To avoid local minima, we increased the number of model fits to 10. Because of the unusual shape of the sequential ROC curve, it is also more difficult to compute a stable AUC. AUC is computed by default, but should be checked carefully for numerical stability and non-monotonicity. For similar reasons, it may be necessary to lower the maximum number of recursion steps (max_steps) in the computation of AUC in the **sdtlu_auc** function. Although the data do show increasing performance with increasing confidence, the model does not fit these data nearly as well as the simultaneous Palmer et al. (2013) data. Indeed, the model fits very poorly. Note the oddly shaped, non-monotonic ROC curve, which is a clear indication that AUC should not be used here. We should note that the sequential model can accurately recover parameters generated from the SDT sequential model, thus, these results strongly suggest that these data were not generated from the sequential SDT model we have implemented. Different assumptions about sequential lineups might result in a better fit.

It is likely that this poor fit is due, in large part, to the Gronlund et al. (2009) experimental paradigm in which the memory strength of the innocent and guilty suspects was manipulated (see Cohen, et al. 2020, for an in-depth discussion). Specifically, manipulating innocent suspect strength likely violated the signal detection model's assumption that innocent suspects and fillers come from the same memory distribution[6]. These data, however, serve as an illustration that the model is not overly flexible and cannot fit any data set, which is a desirable quality because it means that finding a good fit is a better indication that the processes generating the data conform to the model's assumptions. When fit to a different sequential data set from Horry, Palmer, & Brewer (2012), the fit was greatly improved with a G^2 of approximately 21 (see OSF for results). This is perhaps not surprising, given that these researchers did not have the goal of manipulating how much the innocent suspect matched the culprit, and thus took careful measures to ensure lineup fairness as assumed by the model, i.e., innocent suspects are no more likely to strongly match the culprit than fillers.

As mentioned previously, **sdtlu_fit** can be used on data with multiple lineup sizes. When there are multiple lineup sizes, a single set of parameters are estimated across lineup size, however, by necessity, the accuracy measures and predictions are provided separately for each lineup size. In the figures, there will also be one model ROC curve per lineup size. When fitting the model, the model predictions are computed for each lineup size and combined across lineup sizes weighted by the number of trials at that lineup size. These combined predictions are used to compute the fit measure.

Comparing data

The sdtlu package also contains a function, **sdtlu_compare_2**, to compare parameter values from an SDT model across two different data sets. This function is useful, for example, to researchers who want to know how an experimental manipulation affects the SDT parameters. This function takes many of the same fitting and model parameters as **sdtlu_fit**. The major difference is that there are now two sets of data (data_1 and data_2) and lineup sizes (lineup_sizes_1 and lineup_sizes_2). You can also name each of the data sets (data_name_1 and data_name_2), which changes labels in the figure. The options are also similar to **sdtlu_fit**, however, you can also change the output and figure files names including the names of the two individual data sets (save_file_name_1, save_file_name_2 and fig_file_name_1, fig_file_name_2) and the comparison (save_file_name_compare and fig_file_name_compare).

Recall that Palmer et al. (2013) varied delay, i.e., the time between viewing the culprit and participating in the lineup. There were two levels of delay, immediate testing and a week's delay, what we will refer to as short and long delays, respectively. We compare the recovered SDT model parameter values for the Palmer et al. (2013) short and long delay data. Because the results relied on very small differences, we increased the number of bootstrap samples to 500.

The example calls and output are provided in Fig. 13 and a subset of the figure is provided in Fig. 14. This example sends in the data as counts and lineup sizes. The **sdtlu_compare_2** function starts by calling **sdtlu_fit** on both data sets and provides all of the associated output and figures, as described previously. To avoid repetition, this output is not provided here. The new output is shown in Figs. 13 and 14. To provide a measure of statistical differences, the 2.5, 50, and 97.5% quantiles for the differences in bootstrapped fit measures, parameters, and

[5] Trial-by-trial data are not used. The data and predictions are collapsed over suspect positions.

[6] Note that Gronlund et al. (2009) did not rely on SDT modeling to interpret their data, so noting that their paradigm violated the model's assumptions is not a criticism of these authors and does not undermine the original purpose of their study.

```
> seq_data <- sdtlu_process_data(data_name='gronlund_data.csv')
> options <- list(model_type='seq',
+                 fit_fcn='G2',
+                 fix_p='data',
+                 fix_sigma_t='free',
+                 n_bootstrap_samps=50,
+                 n_fits=10)
> sdtlu_fit(data=seq_data, options=options)

quantiles = (0.01, 0.025, 0.05, 0.25, 0.5, 0.75, 0.95, 0.975, 0.99)

----------

G2 = 354.191 (287.042, 293.86, 315.63, 343.61, 364.30, 402.42, 434.23, 449.72, 466.19)

p (fixed) = 0.508 (0.508, 0.508, 0.508, 0.508, 0.508, 0.508, 0.508, 0.508, 0.508)
mu_t = 1.139 (0.944, 0.954, 0.995, 1.087, 1.14, 1.183, 1.224, 1.241, 1.276)
sigma_t = 1.079 (0.858, 0.93, 0.937, 0.992, 1.086, 1.171, 1.288, 1.326, 1.37)
c1 = 1.327 (1.287, 1.287, 1.289, 1.306, 1.324, 1.343, 1.364, 1.368, 1.378)
c2 = 1.397 (1.344, 1.349, 1.354, 1.374, 1.394, 1.41, 1.436, 1.446, 1.449)
c3 = 1.48 (1.431, 1.434, 1.435, 1.461, 1.479, 1.494, 1.518, 1.52, 1.526)
c4 = 1.709 (1.648, 1.653, 1.657, 1.69, 1.709, 1.727, 1.759, 1.77, 1.779)
c5 = 2.084 (2.012, 2.019, 2.025, 2.061, 2.086, 2.12, 2.165, 2.173, 2.195)

mu_f (fixed) = 0
sigma_f (fixed) = 1

----------

area under the curve (lineup size)
auc (6) = 0.058 (0.052, 0.053, 0.053, 0.056, 0.058, 0.059, 0.061, 0.061, 0.063)

diagnosticity collapsed across confidence level (lineup size)
diag (6) = 4.667 (3.947, 3.985, 4.036, 4.48, 4.685, 4.83, 5.046, 5.058, 5.271)

diagnosticity by confidence level (lineup size, confidence level)
diag (6,5) = 2.291 (1.74, 1.789, 1.9, 2.081, 2.29, 2.542, 2.636, 2.715, 2.938)
diag (6,4) = 2.507 (1.893, 1.96, 2.068, 2.297, 2.512, 2.767, 2.887, 2.963, 3.187)
diag (6,3) = 3.007 (2.314, 2.369, 2.475, 2.785, 3.01, 3.29, 3.428, 3.537, 3.749)
diag (6,2) = 4.296 (3.501, 3.536, 3.631, 4.025, 4.38, 4.583, 4.909, 4.931, 4.969)
diag (6,1) = 10.251 (7, 7.119, 7.678, 9.264, 10.25, 11.511, 13.016, 13.468, 14.516)

----------

Data and best-fitting model predictions (lineup size)
(Conditioned on target present/absent)

Data  (6)  0.157 0.102 0.063 0.022 0.016 0.027 0.058 0.065 0.046 0.041 0.403 0.093
0.085 0.081 0.021 0.021 0.036 0.049 0.062 0.031 0.033 0.489
Model (6)  0.151 0.086 0.061 0.024 0.020 0.062 0.084 0.086 0.039 0.037 0.351 0.015
0.020 0.020 0.009 0.009 0.074 0.100 0.102 0.047 0.044 0.559

----------
```

Fig. 11 Partial output from **sdtlu_fit** with the Gronlund et al. (2009) sequential lineup data

performance measures are provided along with the proportion of differences that are less than and greater than 0. The differences are for the long condition (data set 1) minus the short condition (data set 2). We consider there to be a significant difference if zero is not included in the 2.5 to 97.5% interval. The range of this interval is determined by the alpha parameter (the default alpha is .05).

For parameters, only c_4 definitively differs across data sets, although there is a strong trend for overall more conservative responding in the short delay data condition. There are also marginal differences in and μ_t and σ_t, with μ_t tending to be larger and σ_t tending to be smaller in the short delay condition. These results go along with overall better performance with a short delay. Histograms of the differences for μ_t and σ_t are shown in Fig. 14 (the full figure is shown in Fig. 16 in Appendix D). For performance measures, AUC is larger in the short delay condition, a sensible outcome. Diagnosticity is also larger in the

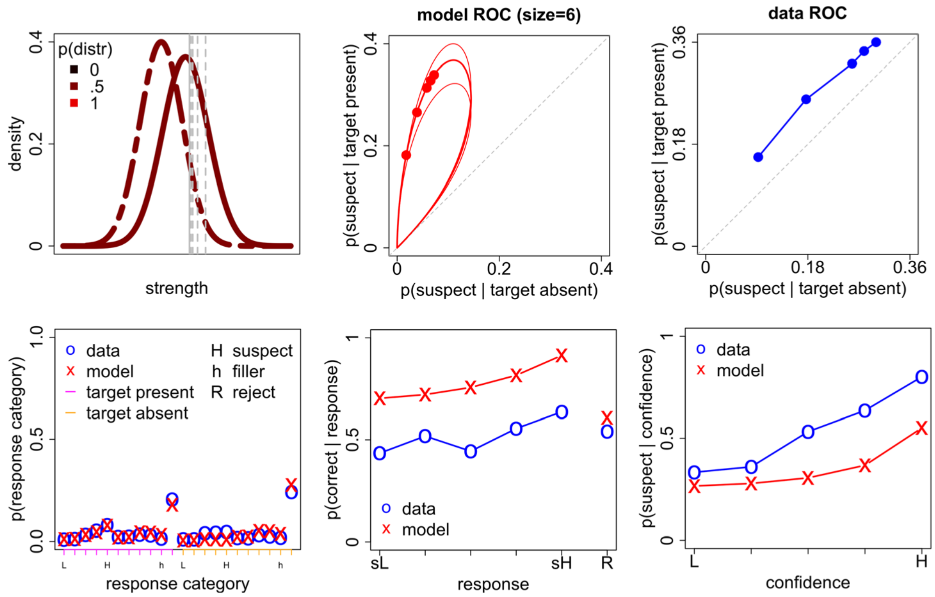


Fig. 12 Figure generated from **sdtlu_fit** with the Gronlund et al. (2009) sequential lineup data. The figure layout has been modified to remove whitespace

short delay condition, but only definitively at confidence levels 2–4 and marginally at confidence level 5.

Simulating SDT models

The sdtlu package provides the ability to simulate the SDT models described previously and provided in the Appendix A and Appendix B. The **sdtlu_sim_sim** and **sdtlu_seq_sim** functions simulate data from the simultaneous and sequential SDT models, respectively.

Both functions were run using the parameters from Fig. 1. Recall that the full set of parameters are the proportion of target present trials p, the mean μ_t and standard deviation σ_t of the target distribution, and the response criteria c_1-c_n. These functions also need to know the lineup size (lineup_sizes) and how many trials are being simulated (n_trials). If use_restr_data is TRUE, restricted data is simulated. In addition, the sequential model is provided with a distribution of suspect positions (pos_prop). The function calls and output are provided in Fig. 15. In this case we simulate two experiments (specified by n_sims).

Help and other functions

There are other useful sdtlu functions. To get additional help, after installing the package, typing ??sdtlu will list all of the available functions and a "how to" file that includes another set of examples. Each function also has its own help file that includes examples.

Discussion

Signal detection theory is a powerful framework for analyzing data. This power has been implemented in several existing R packages that are available for the analysis of ROC data. The sdtlu package leverages the power of the signal detection framework specifically for the analysis of lineup data, or other similarly structured data such as the identification of the location of a tumor within a radiological image (Starr, Metz, Lusted, & Goodenough, 1975; Swets & Pickett, 1982). sdtlu provides functions to process lineup data, determine the best-fitting SDT parameters, compute model-based performance measures such as AUC and diagnosticity, use bootstrapping to determine intervals around these parameters and measures, and compare parameters across two different data sets. Both simultaneous and sequential lineups are supported, as well as show-ups. Closed-form solutions are used. The package can also produce a full set of graphs, including data and model-based ROC curves and the underlying SDT model.

To our knowledge, the sdtlu package represents the first R package implementation of equations that can be solved as integrals to define the predictions of the lineup SDT model. This form offers a computationally more efficient way to determine precise model predictions than the simulation methods often used in past studies (although see Wixted et al., 2018). That said, this package can also simulate data by randomly sampling observations from the model, a function that makes it easy to perform parameter recovery simulations, and it provides an easy way to

```
> long_delay <- sdtlu_process_data(data_name='palmer_delay_long_data.csv')
> long_delay_counts <- long_delay$resp_data_full
> long_delay_sizes <- long_delay$lu_sizes

> short_delay <- sdtlu_process_data(data_name='palmer_delay_short_data.csv')
> short_delay_counts <- short_delay$resp_data_full
> short_delay_sizes <- short_delay$lu_sizes

> options <- list(model_type='sim',
+                 fit_fcn='G2',
+                 fix_p='data', fix_sigma_t='free',
+                 n_bootstrap_samps=500)

> compare_2_return <-
+   sdtlu_compare_2(data_1=long_delay_counts, lineup_sizes_1=long_delay_sizes,
+                   data_2=short_delay_counts, lineup_sizes_2=short_delay_sizes,
+                   options=options)

==========

Key: parameter difference = mean (0.025, 0.5, 0.975 quantiles) [p(<0) p(>0)]

G2 diff = -1.991 (-23.34, -1.584, 19.418) [0.572, 0.428]

p diff = -0.017 (-0.017, -0.017, -0.017) [1, 0]
mu_t diff = -0.242 (-0.529, -0.239, 0.021) [0.964, 0.036]
sigma_t diff = 0.317 (-0.049, 0.309, 0.772) [0.054, 0.946]
c_1 diff = -0.054 (-0.174, -0.056, 0.067) [0.808, 0.192]
c_2 diff = -0.094 (-0.228, -0.096, 0.031) [0.926, 0.074]
c_3 diff = -0.097 (-0.226, -0.097, 0.036) [0.938, 0.062]
c_4 diff = -0.141 (-0.287, -0.137, -0.006) [0.978, 0.022]
c_5 diff = -0.087 (-0.31, -0.083, 0.143) [0.784, 0.216]

area under the curve (lineup size)
auc(8) diff = -0.011 (-0.022, -0.011, 0) [0.972, 0.028]

diagnosticity collapsed across confidence level (lineup size)
diag(8) diff = -2.57 (-5.351, -2.531, 0.409) [0.966, 0.034]

diagnosticity by confidence level (lineup size, confidence level)
diag(8,5) diff = -1.12 (-2.455, -1.119, 0.1) [0.962, 0.038]
diag(8,4) diff = -1.628 (-3.306, -1.599, -0.149) [0.984, 0.016]
diag(8,3) diff = -2.954 (-5.29, -2.966, -0.675) [0.996, 0.004]
diag(8,2) diff = -6.565 (-11.394, -6.466, -2.109) [1, 0]
diag(8,1) diff = -6.309 (-41.739, -6.359, 35.998) [0.66, 0.34]
```

Fig. 13 Partial output from an example use of the **sdtlu_compare_2** function. Comparing parameter values across the short and long delay conditions of Palmer et al. (2013)

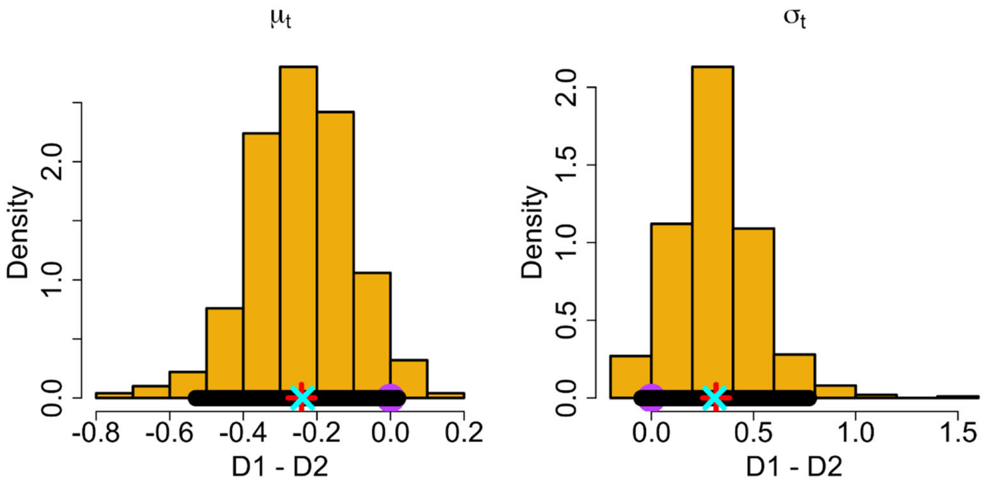


Fig. 14 Figure generated from1 **sdtlu_compare_2** with the Palmer et al. (2013) short and long delay simultaneous lineup data. Only the results of μ_t and σ_t are shown. The full figure is shown in Fig. 16 in Appendix D

```
> p <- .5
> mu_t <- 2
> sigma_t <- 1.4
> cs <- c(.5, 2.5, 3.5)

> params <- c(p, mu_t, sigma_t, cs[1], cs[2], cs[3])

> lineup_sizes <- 6
> n_trials <- 1000
> n_sims <- 2

> sdtlu_sim_sim(params, lineup_sizes, n_trials, n_sims, use_restr_data=FALSE)

     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
[1,]   69  100  202    1    8  128    9    0    2    69     2    19   337    54
[2,]   66   94  164    0   14  128    9    0    6    80     0    13   372    54

> pos_prop <- c(0, 1/2, 0, 0, 1/2, 0)

> sdtlu_seq_sim(params, lineup_sizes, n_trials, n_sims, pos_prop=pos_prop,
use_restr_data=FALSE)

     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
[1,]   35   53  127    1    1  282   13    0    1    51     0    10   375    51
[2,]   40   50  106    0    2  277   14    0    1    75     0     6   381    48

> sdtlu_sim_sim(params, lineup_sizes, n_trials, n_sims, use_restr_data=TRUE)
     [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,]   65  104  255    2   27  479   68
[2,]   64  100  237    1   29  493   76
```

Fig. 15 Examples of the use of **sdtlu_sim_sim** and **sdtlu_seq_sim** to simulated experimental data

explore position effects in sequential lineups. sdtlu returns a variety of performance measures used by lineup researchers, and it is, to our knowledge, the first package to calculate theoretical AUC measures from the lineup SDT model.

Thus, the sdtlu package offers eyewitness researchers a number of specialized functions that are not available in packages designed for more general applications of ROC analysis. The most downloaded packages are ROCR (Sing, Sander, Beerenwinkel, & Lengauer, 2005, downloaded ~63 k times in the month prior to 8/10/19) and pROC (Robin et al., 2011, downloaded ~46 k times), with the more functionally limited sROC in third place (Wang, 2012, downloaded ~5 k times). ROCR and pROC both offer AUC estimation, including for partial AUCs, as well as statistical comparison of two ROC curves, smoothing of data, and a range of plotting options. They also both provide tools for analysis of ROC data based on continuously valued measures; in psychological research, the reaction-time based ROC is one example (Thomas & Myers, 1972). pROC easily implements bootstrapping of samples and plotting of confidence intervals; ROCR has the advantage of generating predictions for how a new sample will be classified (i.e., as target or lure). One key difference between these packages and sdtlu is that the latter does not just estimate AUC using the empirical ROC points, but defines the theoretical ROC function based on an SDT model designed for lineup tasks and uses iterative quadrature to find the area under this continuous function. Thus, sdtlu provides an option for eyewitness memory researchers who wish to use more theoretically motivated performance measures. Another important consideration is that, among these packages, only sdtlu can estimate the base-rate of lineups that include a guilty suspect. sdtlu also includes functions for generating simulated data sets from an SDT model with either a simultaneous and sequential lineup design, which makes it easy for researchers to conduct parameter recovery simulations or explore model predictions when planning a new study.

In conclusion, the sdtlu package offers a number of unique tools for lineup researchers, and we hope that it will contribute to the growing sophistication in the analysis and interpretation of both empirical and real-world eyewitness identification data.

Acknowledgements The authors would like to thank Andrea Cataldo for her help collating the example data sets, Tanya Leise for help verifying AUC, and to the authors of the data sets for making them available.

Open Practices Statement The R package and all data are available at https://osf.io/mfk4e.

Appendix A Equations for simultaneous lineups

Let $\phi(s, \mu, \sigma)$ be the density of a normal distribution with mean μ and standard deviation σ and let $\Phi(s, \mu, \sigma) = \int_{-\infty}^{s} \phi(x, \mu, \sigma)dx$ be the cumulative normal.

Let *resp*=response, with values *sus*=suspect, *fil*=filler, and *rej*=reject. Let *conf*=the response confidence level, with values 1…max confidence level. Note that, for notational convenience, 1 is the lowest confidence level here. Let *tar*=target, with values *pres*=present and *abs*=absent.

The following model parameters are used: p=P(target present), μ_t=target distribution mean, σ_t=target distribution standard deviation, μ_l =lure distribution mean, σ_l =lure distribution standard deviation, c_i=the values of the i^{th} response criterion, where c_1 is the lowest response criterion.

The lineup size is given by l.

Also see Fig. 1.

Suspect response

The probability of a suspect response at confidence level i is given by[7]

$$\begin{aligned} &P(resp = sus \cap conf = i) \\ &\quad = pP(resp = sus \cap conf = i | tar = pres) \\ &\qquad + (1-p)P(resp = sus \cap conf = i | tar = abs), \end{aligned} \tag{A1a}$$

where

$$\begin{aligned} &P(resp = sus \cap conf = i | tar = pres) \\ &\quad = \int_{c_i}^{c_{i+1}} \phi(s, \mu_t, \sigma_t)\Phi(s, \mu_l, \sigma_l)^{l-1} ds \end{aligned} \tag{A1b}$$

and

$$\begin{aligned} &P(resp = sus \cap conf = i | tar = abs) \\ &\quad = \int_{c_i}^{c_{i+1}} \phi(s, \mu_l, \sigma_l)\Phi(s, \mu_l, \sigma_l)^{l-1} ds. \end{aligned} \tag{A1c}$$

In Equation A1a finds the overall probability of observing a suspect ID at a given confidence level by taking the weighted average of the corresponding probability for target present and target absent lineups, where the weight is the proportion of target-present lineups (p). This equation is needed to get predictions for restricted data, such as data from real lineups, where the guilt status of each suspect is unknown.

Equations A1b and A1c provide the probabilities of choosing the suspect at a given confidence level for target-present and target-absent lineups individually. These equations are used for full data in which each suspect can be classified as guilty or innocent.

In Equation A1b, the first term in the integral is the probability density at a value of s for a guilty suspect (i.e., a draw from the target distribution). The second term in the integral is the probability that a given filler (i.e., a draw from the lure distribution) has a strength value below s, raised to the power of the number of fillers (l – 1) to give the joint probability that all of the fillers have a strength value below s (the model assumes that filler strengths are independent of one another and independent of the suspect strength). Multiplying the two terms gives the joint probability density that the suspect has a strength value of s and has a higher strength than all of the fillers, i.e., the suspect is selected as the lineup member whose strength value is used to make the identification decision. Integrating this equation between c_i and c_{i+1} gives the joint probability that the suspect both has a strength value in this range and has the highest strength value in the lineup, i.e., the probability of selecting the suspect with confidence level i.

In Equation A1c has the same structure as Equation A1b, but the probability density for the suspect is based on the lure distribution to represent a target absent lineup.

Filler response

The probability of a filler response at confidence level i is given by

$$\begin{aligned} &P(resp = fil \cap conf = i) \\ &\quad = pP(resp = fil \cap conf = i | tar = pres) \\ &\qquad + (1-p)P(resp = fil \cap conf = i | tar = abs), \end{aligned} \tag{A2a}$$

where

$$\begin{aligned} &P(resp = fil \cap conf = i | tar = pres) \\ &\quad = (l-1) \int_{c_i}^{c_{i+1}} \phi(s, \mu_l, \sigma_l)\Phi(s, \mu_l, \sigma_l)^{l-2}\Phi(s, \mu_t, \sigma_t) ds \end{aligned} \tag{A2b}$$

and

$$\begin{aligned} &P(resp = fil \cap conf = i | tar = abs) \\ &\quad = (l-1) \int_{c_i}^{c_{i+1}} \phi(s, \mu_l, \sigma_l)\Phi(s, \mu_l, \sigma_l)^{l-2}\Phi(s, \mu_l, \sigma_l) ds. \end{aligned} \tag{A2c}$$

Equation A2a finds that the overall probability of selecting a filler at confidence level i by taking the weighted average of the

[7] Note that the package code implements the joint, rather than conditional probabilities, for example, $P(resp = sus \cap conf = i) = P(resp = sus \cap conf = i \cap tar = pres) + P(resp = sus \cap conf = i \cap tar = abs)$, $P(resp = sus \cap conf = i \cap tar = pres) = pP(resp = sus \cap conf = i| \; tar = pres)$, and $P(resp = sus \cap conf = i \cap tar = abs) = (1 - p)P(resp = sus \cap conf = i| \; tar = abs)$. For ease of exposition, the conditionals are presented here, but the results are equivalent. The same transformation holds for all equations below.

corresponding probability for target-present and target-absent lineups, which would be needed for restricted data.

Equations A2b and A2c give the probability of selecting a filler at confidence level i for target-present and target-absent lineups individually, so these equations would be used for full data.

For Equation A2b, the first term in the integral is the probability density at strength value s for a filler F_1 (i.e., a draw from the lure distribution). The second term in the integral is the probability that one of the other $l - 2$ fillers has a strength value below s. Exponentiating provides the joint probability that all of these other fillers all have a strength value below s. The third term in the integral is the probability that the suspect (i.e., a random draw from the target distribution) has a strength value below s. Multiplying these three terms gives the probability density that F_1 has a strength value of s and this strength value is higher than the strength values for all the other fillers and the suspect. That is, F_1 is selected as the lineup member whose strength value will inform the identification decision and has a strength value of s. Integrating this equation between c_i and ci+1 gives the joint probability that filler F_1 both has a strength value in this range and has the highest strength value in the lineup, that is, the probability of selecting filler F_1 with confidence i. Finally, multiplying this value by the number of fillers ($l - 1$) gives the probability of selecting any of the fillers at confidence level i.

Equation A2c has the same structure as Equation A2b, except that the suspect becomes a draw from the lure distribution to represent a target-absent lineup.

No identification

The probability of rejecting the lineup, that is, not identifying any lineup member as the culprit, is given by

$$P(resp = rej) = pP(resp = rej|tar = pres) + (1{-}p)P(resp = rej|tar = abs), \qquad \text{(A3a)}$$

where

$$P(resp = rej|tar = pres) = \int_{-\infty}^{c_1} \phi(s, \mu_t, \sigma_t)\Phi(s, \mu_l, \sigma_l)^{l-1} ds + (l{-}1)\int_{-\infty}^{c_1} \phi(s, \mu_l, \sigma_l)\Phi(s, \mu_l, \sigma_l)^{l-2}\Phi(s, \mu_t, \sigma_t) ds \qquad \text{(A3b)}$$

and

$$P(resp = rej|tar = abs) = \int_{-\infty}^{c_1} \phi(s, \mu_l, \sigma_l)\Phi(s, \mu_l, \sigma_l)^{l-1} ds + (l{-}1)\int_{-\infty}^{c_1} \phi(s, \mu_l, \sigma_l)\Phi(s, \mu_l, \sigma_l)^{l-2}\Phi(s, \mu_l, \sigma_l) ds. \qquad \text{(A3c)}$$

Equation A3a indicates that the overall probability of rejecting a lineup is the weighted average of the probability of rejection for target-present and target-absent lineups. This value is needed to fit restricted data.

Equations A3b and A3c give the probability of rejection for target-present and target-absent lineups, and so are used for full data.

In Equation A3b, the first integral has the same structure as Equation A1b and gives the probability that the suspect has the highest strength value and a strength value below c_1. The second integral has the same structure as Equation A2b and gives the probability that a given filler has the highest strength value and a strength value below c_1, which is multiplied by the number of fillers that could potentially have the highest strength value ($l - 1$). Adding these two terms gives the total probability that the lineup member with the highest strength value (whether suspect or filler) has a strength below c_1, that is, the probability that the lineup will be rejected.

Equation A3c has the same structure as Equation A3b, except that the suspect is now a draw from the lure distribution to represent a target-absent lineup.

Appendix B Equations for sequential lineups

Let $\phi(s, \mu, \sigma)$ be the density of a normal distribution with mean μ and standard deviation σ and let $\Phi(s, \mu, \sigma) = \int_{-\infty}^{s} \phi(x, \mu, \sigma)dx$ be the cumulative normal.

Let *resp*=response, with values *sus*=suspect, *fil*=filler, and *rej*=reject. Let *conf*=the response confidence level, with values 1…max confidence level. Note that, for notational convenience, 1 is the lowest confidence level here. Let *tar*=target, with values *pres*=present and *abs*=absent. Let *spos*=the subject position, with values=1…lineup size.

The following model parameters are used: p=P(target present), μ_t=target distribution mean, σ_t=target distribution standard deviation, μ_l =lure distribution mean, σ_l =lure distribution standard deviation, c_i=the values of the i^{th} response criterion, where c_1 is the lowest response criterion.

The lineup size is given by l.

For the sequential lineups, suspect and filler responses depend on the suspect position.

Also see Fig. 1.

Suspect response

The probability of a suspect response at confidence level i is given by[8]

[8] Note that the package code implements the joint, rather than conditional probabilities. See FN 7.

$$P(resp = sus \cap conf = i)$$

$$= pP(resp = sus \cap conf = i | tar = pres)$$

$$+ (1{-}p)P(resp = sus \cap conf = i | tar = abs), \quad \text{(B1a)}$$

where

$$P(resp = sus \cap conf = i | tar = pres)$$

$$= \sum_{j=1}^{l} P(spos = j)P(resp = sus \cap conf = i | spos = j \cap tar = pres), \quad \text{(B1b)}$$

where $P(spos = j)$is the probability that the suspect is in position j, as determined by the lineup designer and

$$P(resp = sus \cap conf = i | spos = j \cap tar = pres)$$

$$= \Phi(c_1, \mu_l, \sigma_l)^{j-1} \int_{c_i}^{c_{i+1}} \phi(s, \mu_t, \sigma_t)ds \quad \text{(B1c)}$$

and

$$P(resp = sus \cap conf = i | tar = abs)$$

$$= \sum_{j=1}^{l} P(spos = j)P(resp = sus \cap conf = i | spos = j \cap tar = abs), \quad \text{(B1d)}$$

where

$$P(resp = sus \cap conf = i | spos = j \cap tar = abs)$$

$$= \Phi(c_1, \mu_l, \sigma_l)^{j-1} \int_{c_i}^{c_{i+1}} \phi(s, \mu_l, \sigma_l)ds. \quad \text{(B1e)}$$

Equation B1a provides the overall probability of a suspect ID at confidence level i, used for restricted data, by taking the weighted average of the probability of a suspect ID at confidence level i for target-present and target-absent lineups. The weight is determined by the proportion of target-present lineups (p).

Equations B1b and B1c give the probability of a suspect ID at confidence level i for target-present lineups. Equation B1c assumes a given suspect position and Equation B1b calculates this value across the full distribution of suspect positions. In Equation B1c, the first term is the probability that the witness would reach the suspect position (j) in the lineup; that is, that all $j - 1$ preceding fillers would have strength values below the identification criterion c_1. The second term is the probability that the witness would identify the suspect with confidence level i, found by integrating the probability density of the target distribution between the criteria defining the bounds of confidence region i. Multiplying the two terms gives the probability that the witness would reach the suspect in the lineup sequence and would identify them with confidence level i once they do so. Equation B1b takes the weighted average of the values returned by Equation B1c for each suspect position, where the weights are taken from the probability distribution of suspect positions.

Equations B1d and B1e have the same structure as Equations B1b and B1c, except that the suspect strength comes from the lure distribution instead of the target distribution to represent a target-absent lineup.

Filler response

The probability of a filler response at confidence level i is given by

$$P(resp = fil \cap conf = i)$$

$$= pP(resp = fil \cap conf = i | tar = pres)$$

$$+ (1{-}p)P(resp = fil \cap conf = i | tar = abs), \quad \text{(B2a)}$$

where

$$P(resp = fil \cap conf = i | tar = pres)$$

$$= \sum_{j=1}^{l} P(spos = j)P(resp = fil \cap conf = i | spos = j \cap tar = pres), \quad \text{(B2b)}$$

where $P(spos = j)$is the probability that the suspect is in position j, as determined by the lineup designer, and

$$P(resp = fil \cap conf = i | spos = j \cap tar = pres)$$

$$= \sum_{k=1}^{l} \begin{cases} 0 & \mathit{if}\ k = j \\ \Phi(c_1, \mu_l, \sigma_l)^{k-1} \int_{c_i}^{c_{i+1}} \phi(s, \mu_l, \sigma_l)ds & \mathit{if}\ k < j \\ \Phi(c_1, \mu_l, \sigma_l)^{k-2} \Phi(c_1, \mu_t, \sigma_t) \int_{c_i}^{c_{i+1}} \phi(s, \mu_l, \sigma_l)ds & \mathit{if}\ k > j \end{cases} \quad \text{(B2c)}$$

where k denotes each lineup position, and

$$P(resp = fil \cap conf = i | tar = abs)$$

$$= \sum_{j=1}^{l} P(spos = j)P(resp = fil \cap conf = i | spos = j \cap tar = abs), \quad \text{(B2d)}$$

where

$$P(resp = fil \cap conf = i | spos = j \cap tar = abs)$$

$$= \sum_{k=1}^{l} \begin{cases} 0 & \mathit{if}\ k = j \\ \Phi(c_1, \mu_l, \sigma_l)^{k-1} \int_{c_i}^{c_{i+1}} \phi(s, \mu_l, \sigma_l)ds & \mathit{if}\ k < j. \\ \Phi(c_1, \mu_l, \sigma_l)^{k-2} \Phi(c_1, \mu_l, \sigma_l) \int_{c_i}^{c_{i+1}} \phi(s, \mu_l, \sigma_l)ds & \mathit{if}\ k > j \end{cases} \quad \text{(B2e)}$$

Equations B2a, B2b, and B2d are directly analogous to Equations B1a, B1b, and B1d. See the explanation of those equations. Equations B2c and B2e are similar to Equations B1c and B1e, but they give the probability of selecting any of the fillers (as opposed to the single suspect) with a given confidence level.

First consider Equation B2c. The sum goes over all the positions (k) in the lineup (1 through l). For each position, the bracketed equations give the probability of selecting a filler at confidence level i for that position. This value is 0 if the suspect (and not a filler) is in that position (the first "if" statement). Recall that the suspect is in position j. The second "if" statement applies to filler positions that come before the suspect position in the lineup. For these positions, the equation is the probability of rejecting all of the $k - 1$ fillers that came before the filler in position k – that is, the probability that $k - 1$ random draws from the lure distribution would all fall below the identification criterion c_1 – multiplied by the probability of selecting confidence level i for the filler in position k, which is found by integrating the probability density of the lure distribution between the confidence criteria defining this confidence region. The third "if" statement applies to filler positions that come after the suspect position. The only change from the equation just discussed is that now the probability of getting "past" the faces before position k is found by multiplying the probability that each of the $k - 2$ preceding fillers fell below the identification criterion and the probability that the one preceding guilty suspect fell below the identification criterion.

Appendix D Full sdtlu_compare_2 figure

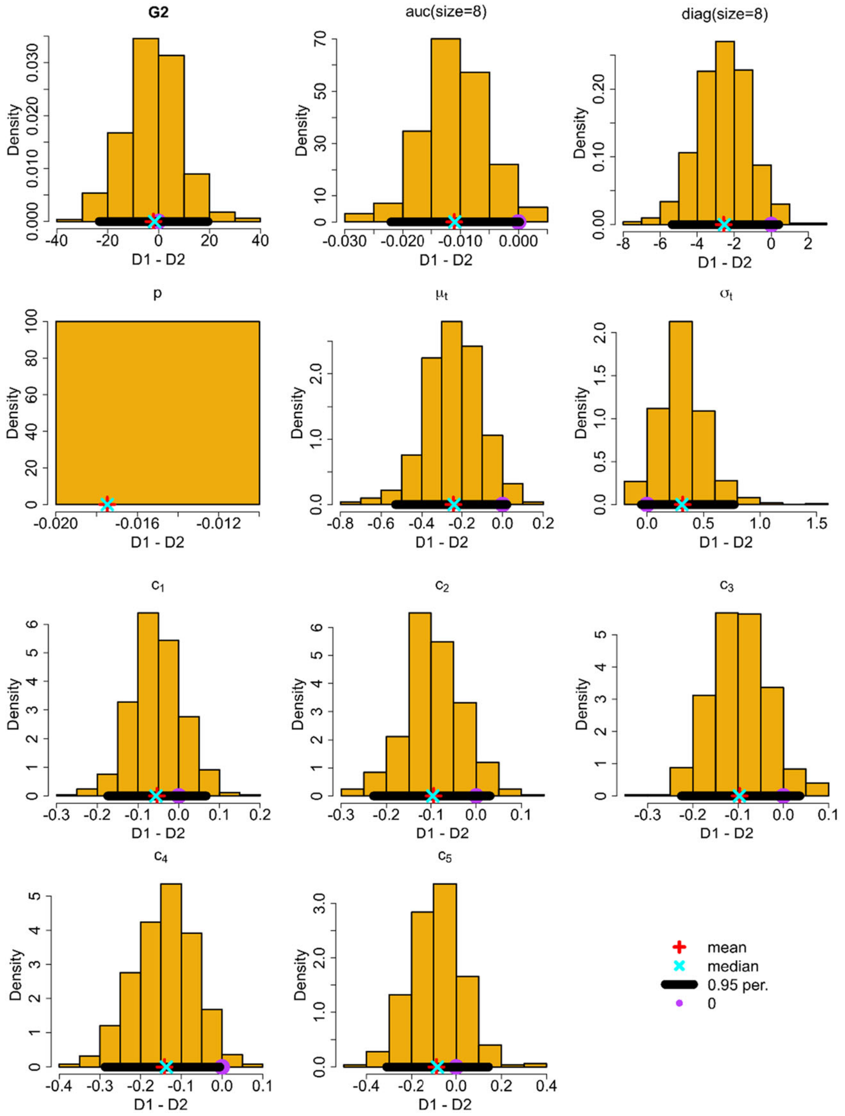


Fig. 16 Figure generated from **sdtlu_compare_2** with the Palmer et al. (2013) short and long delay simultaneous lineup data. To remove whitespace, the figure has been reformatted

Equation B2e is like Equation B2c, except that, because this is for a target-absent lineup, the suspect term in the third “if” statement is defined by the lure distribution.

No identification

Lineup rejections do not depend on suspect position. The probability of rejecting the lineup, that is, failing to identify any individual, is given by

$$P(resp = rej) = pP(resp = rej|tar = pres) + (1{-}p)P(resp = rej|tar = abs), \qquad \text{(B3a)}$$

where

$$P(resp = rej|tar = pres) = \Phi(c_1, \mu_t, \sigma_t)\Phi(c_1, \mu_l, \sigma_l)^{l-1} \qquad \text{(B3b)}$$

and

$$P(resp = rej|tar = abs) = \Phi(c_1, \mu_l, \sigma_l)\Phi(c_1, \mu_l, \sigma_l)^{l-1}. \qquad \text{(B3c)}$$

Equation B3a provides the overall probability of rejection across all lineups, used for fitting restricted data, by taking the weighted average of the probability of rejection for target-present and target-absent lineups. Equation B3b gives the probability that all members of a target-present lineup would fall below the identification criterion c_1, obtained by multiplying the probability that a draw from the target distribution falls below this criterion by the probability that all of the $l - 1$ fillers also fall below this criterion. Equation B3c is the same as equation B3b, except now all lineup members are assumed to be draws from the lure distribution.

Appendix C Equations for AUC and diagnosticity

Area under the curve

Let T_a = *P*(*suspect pick* | *target absent*) for a given set of model parameters θ. Let T_a^* be the highest T_a on the ROC curve, which can be less than 1 for lineups. Let $T_p(x)$ = *P*(*suspect pick* | *target present*) when $T_a = x$. Then,

$$AUC = \int_0^{T_a^*} T_p(x)dx. \qquad \text{(C1)}$$

Analytic solutions do not typically exist for AUC. Here, AUC is computed numerically using iterative quadrature.

Diagnosticity

Let $\phi(s, \mu, \sigma)$ be the density of a normal distribution with mean μ and standard deviation σ and let $\Phi(s, \mu, \sigma) = \int_{-\infty}^{s} \phi(x, \mu, \sigma) dx$ be the cumulative normal. Let μ_t=target distribution mean, σ_t=target distribution standard deviation, μ_l =lure distribution mean, σ_l =lure distribution standard deviation, c_i=the values of the i^{th} response criterion, where c_1 is the lowest response criterion. Let the lineup size be given by *l*. Also see Fig. 1.

Simultaneous lineups

First, consider diagnosticity for simultaneous lineups. Overall diagnosticity, collapsing over response threshold, is given by

$$diagnosticity = \frac{\int_{c_1}^{\infty} \phi(s, \mu_t, \sigma_t)\Phi(s, \mu_l, \sigma_l)^{l-1} ds}{\int_{c_1}^{\infty} \phi(s, \mu_l, \sigma_l)\Phi(s, \mu_l, \sigma_l)^{l-1} ds} \qquad \text{(C2)}$$

and diagnosticity at confidence level *i* is given by

$$diagnosticity = \frac{\int_{c_i}^{c_{i+1}} \phi(s, \mu_t, \sigma_t)\Phi(s, \mu_l, \sigma_l)^{l-1} ds}{\int_{c_i}^{c_{i+1}} \phi(s, \mu_l, \sigma_l)\Phi(s, \mu_l, \sigma_l)^{l-1} ds}. \qquad \text{(C3)}$$

Sequential lineups

Next, consider diagnosticity for sequential lineups. Overall diagnosticity, collapsing over confidence, is given by

$$diagnosticity = \frac{\sum_{j=1}^{l} P(spos = j)\Phi(c_1, \mu_l, \sigma_l)^{j-1} \int_{c_1}^{\infty} \phi(s, \mu_t, \sigma_t) ds}{\sum_{j=1}^{l} P(spos = j)\Phi(c_1, \mu_l, \sigma_l)^{j-1} \int_{c_1}^{\infty} \phi(s, \mu_l, \sigma_l) ds}, \qquad \text{(C4)}$$

where *P(spos=j)* is the probability of a suspect at position *j*. The diagnosticity at confidence level *i* is given by

$$diagnosticity = \frac{\sum_{j=1}^{l} P(spos = j)\Phi(c_1, \mu_l, \sigma_l)^{j-1} \int_{c_i}^{c_{i+1}} \phi(s, \mu_t, \sigma_t) ds}{\sum_{j=1}^{l} P(spos = j)\Phi(c_1, \mu_l, \sigma_l)^{j-1} \int_{c_i}^{c_{i+1}} \phi(s, \mu_l, \sigma_l) ds}. \qquad \text{(C5)}$$

References

Bothwell, R. K., Deffenbacher, K. A., Brigham, J. C. (1987). Correlation of eyewitness accuracy and confidence: Optimality hypothesis revisited. *Journal of Applied Psychology, 72*, 691–695.

Brewer, N., & Burke, A. (2002). Effects of testimonial inconsistencies and eyewitness confidence on mock-juror judgments. *Law and Human Behavior*, *26*, 353–364.

Carlson, C. A., & 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*, 45–53.

Cohen, A.L., Starns, J.J., Rotello, C.M., & Cataldo, A.M. (2020). Estimating the proportion of guilty suspects and posterior probability of guilt in lineups using signal-detection models. *Cognitive Research: Principles and Implications, 5*, 21. https://doi.org/10.1186/s41235-020-00219-4

Colloff, M. F., Wade, K. A., Wixted, J. T., & Maylor, E. A. (2017). A signal-detection analysis of eyewitness identification across the adult lifespan. *Psychology and Aging, 32*, 243–258. DOI: https://doi.org/10.1037/pag0000168

Cutler, B. L., Penrod, S. D., & Dexter, H. R. (1990). Juror sensitivity to eyewitness identification evidence. *Law and Human Behavior, 14*, 185–191.

Deffenbacher, K. A. (1980). Eyewitness accuracy and confidence: Can we infer anything about their relationship? Law and Human Behavior, 4, 243–260. https://doi.org/10.1007/BF01040617

Dobolyi, D. G., & Dodson, C. S. (2013). Eyewitness confidence in simultaneous and sequential lineups: A criterion shift account for sequential mistaken identification overconfidence. *Journal of Experimental Psychology: Applied, 19*, 345–357.

Gonzalez, R., Ellsworth, P. C., & Pembroke, M. (1993). Response biases in lineups and showups. *Journal of Personality and Social Psychology, 64*, 525–537

Gronlund, S. D., Carlson, C. A., Dailey, S. B., Goodsell, C. A. (2009). Robustness of the sequential lineup advantage. *Journal of Experimental Psychology: Applied, 15*, 140–152. DOI: https://doi.org/10.1037/a0015082

Horry, R., Brewer, N., Weber, N., & Palmer, M. A. (2015). The effects of allowing a second sequential lineup lap on choosing and probative value. *Psychology, Public Policy, and Law, 21*, 121–133. doi: https://doi.org/10.1037/law0000041

Horry, R., Palmer, M., & Brewer, N. (2012). Backloading in the sequential lineup prevents within-lineup criterion shifts that undermine eyewitness identification performance. *Journal of Experimental Psychology: Applied, 18*, 346–360.

Levi, A. M. (2012). Much better than the sequential lineup: a 120-person lineup. *Psychology, Crime & Law, 18*, 631–640. DOI: https://doi.org/10.1080/1068316X.2010.526120

Macmillan, N. A., & Creelman, D. C. (2005). *Detection Theory: A User's Guide.* New Jersey: Lawrence Erlbaum Associates.

McClish, D. K. (1989). Analyzing a portion of the ROC curve. *Medical Decision Making, 9*, 190–195.

Mickes, L. (2015). Receiver operating characteristic analysis and confidence–accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory and Cognition, 4,* 93–102. DOI: https://doi.org/10.1016/j.jarmac.2015.01.003

Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous versus sequential lineups. *Journal of Experimental Psychology: Applied, 18*, 361–376.

Palmer, M. A., Brewer, N., Weber, N., & Nagesh, A. (2013). The confidence–accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. *Journal of Experimental Psychology: Applied, 19*, 55–71. DOI: https://doi.org/10.1037/a0031602

Palmer, M. A., Sauer, J. D., & Holt, G. A. (2017). Undermining position effects in choices from arrays, with implications for police lineups. *Journal of experimental psychology: applied*, 23(1), 71.

Police Executive Research Forum. (2013). A national survey of eyewitness identification procedures in law enforcement agencies. Retrieved from https://www.ncjrs.gov/pdffiles1/nij/grants/242617.pdf

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011) "pROC: an open-source package for R and S+ to analyze and compare ROC curves". *BMC Bioinformatics, 7,* 77. DOI: https://doi.org/10.1186/1471-2105-12-77

Rotello, C. M., & Chen, T. (2016). ROC curve analyses of eyewitness identification decisions: An analysis of the recent debate. *Cognitive Research: Principles and Implications*, 1:10. DOI https://doi.org/10.1186/s41235-016-0006-7

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics, 21*, 3940–3941.

Starr, S. J., Metz, C. E., Lusted, L. B., & Goodenough, D. J. (1975). Visual detection and localization of radiographic images. *Radiology, 116*, 553–538.

Swets, J. A., & Pickett, R. M. (1982). *Evaluation of diagnostic systems: Methods from signal detection theory*. New York: Academic Press.

Thomas, E. A. C., & Myers, J. L. (1972). Implications of latency data for threshold and nonthreshold models of signal detection. *Journal of Mathematical Psychology, 9*, 253–285. DOI: https://doi.org/10.1016/0022-2496(72)90018-1

Tunnicliff, J. L., & Clark, S. E. (2000). Selecting foils for identification lineups: Matching suspects or descriptions? *Law and Human Behavior, 24*, 231–258.

Wang, X.-F. (2012). Nonparametric smooth ROC curves for continuous data. R package version 0.1–2.

Wells, G. L. (2014). Eyewitness identification: Probative value, criterion shifts, and policy regarding the sequential lineup. *Current Directions in Psychological Science, 23*, 11–16. doi:https://doi.org/10.1177/0963721413504781

Wells, G. L., & Lindsay, R. C. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin, 88*, 776–784. doi:https://doi.org/10.1037/0033-2909.88.3.776

Wells, G. L., & Olson, E. A. (2003). Eyewitness testimony. *Annual Review of Psychology, 54,* 277–295.

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

Wells, G. L., & Turtle, J. W., (1986). Eyewitness identification: The importance of lineup models. *Psychological Bulletin, 99,* 320–329.

Wells, G. L., Yang, Y., & Smalarz, L. (2015). Eyewitness identification: Bayesian information gain, base-rate effect equivalency curves, and reasonable suspicion. *Law and Human Behavior, 39(2)*, 99.

Wetmore, S. A., Neuschatz, J. S., Gronlund, S. D., Wooten, A., Goodsell, C. A., Carlson C. A. (2015). Effect of retention interval on showup and lineup performance. *Journal of Applied Research in Memory and Cognition, 4*, 8–14. DOI: https://doi.org/10.1016/j.jarmac.2014.07.003

Wixted, J. T., Mickes, L. (2012). The field of eyewitness memory should abandon "probative value" and embrace receiver operating characteristic analysis. *Perspectives on Psychological Science, 7*, 275–278. doi:https://doi.org/10.1177/1745691612442906

Wixted, J. T., Mickes, L., Dunn, J. C., Clark, S. E., & Wells, W. (2016). Estimating the reliability of eyewitness identifications from police lineups. *Proceedings of the National Academy of Sciences, 113*, 304-309. https://doi.org/10.1073/pnas.1516814112

Wixted, J. T., Vul, E., Mickes, L., & Wilson, B. M. (2018). Models of lineup memory. *Cognitive Psychology, 105*, 81–114.

Wogalter, M. S., Malpass, R. S., & McQuiston, D. E. (2004). A national survey of U.S. police on preparation and conduct of identification lineups. *Psychology, Crime & Law, 10*, 69 –82. DOI: https://doi.org/10.1080/10683160410001641873