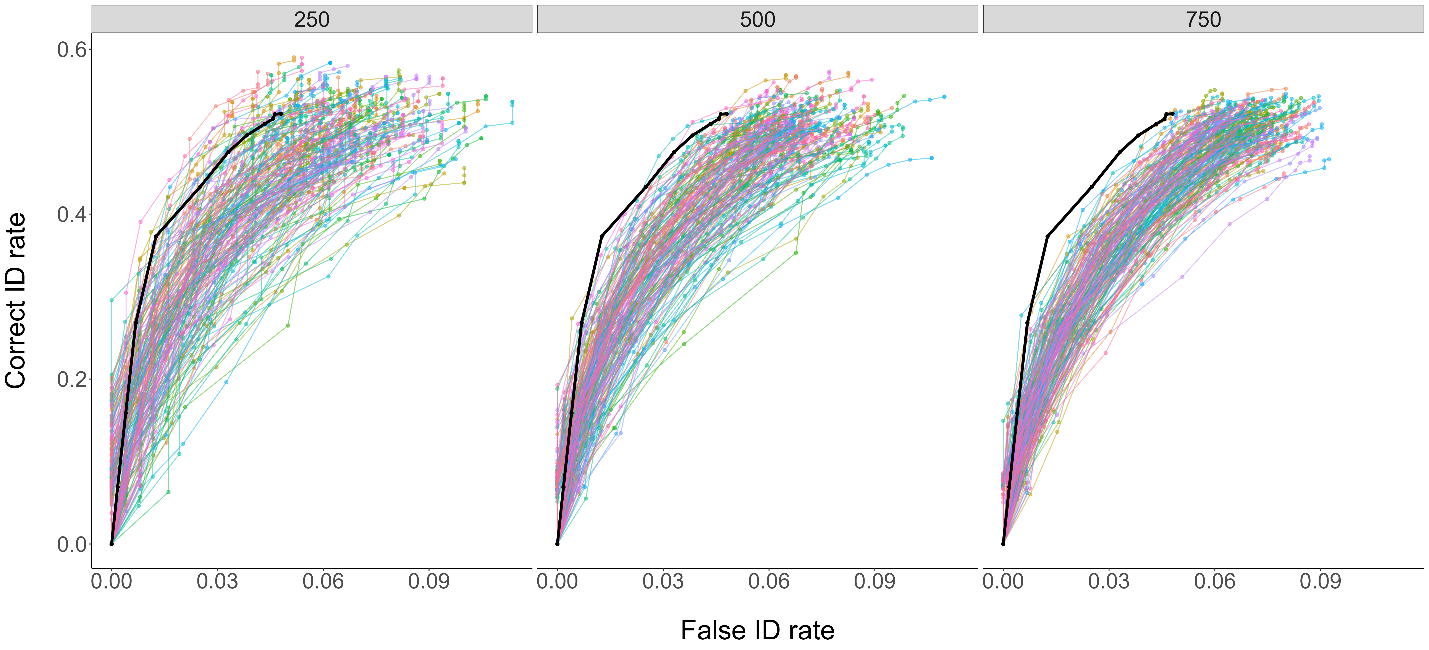
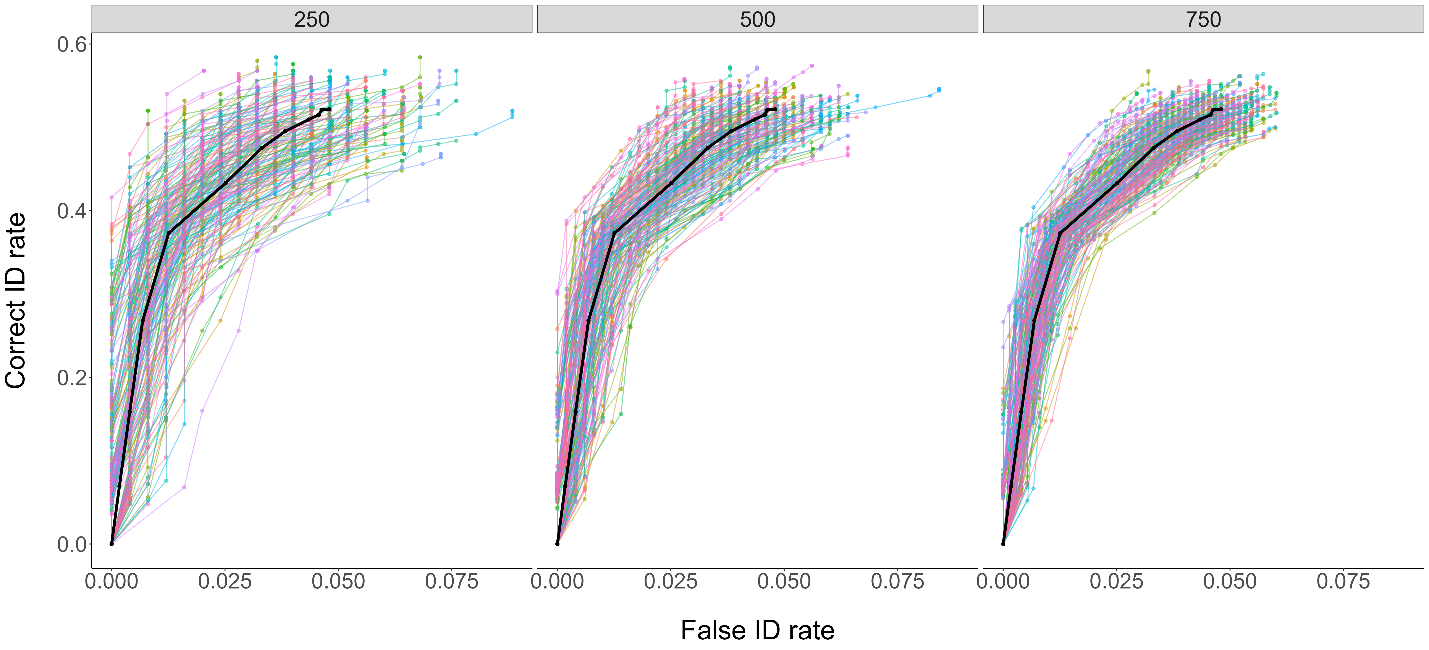
**powe(R)OC testing**

* Show that similar pAUC differences, *D* statistics can lead to different power estimates – hence need for simulation-based methods
* Possible addition: Analogue with a simpler analysis (e.g., *t-*test, same Cohen’s *d* but different power)
* Test robustness of effect size measure:
  + Same effect, same *N*, different base ROCs: DIFFERENT POWER, EFFECT SIZE MEASURE INADEQUATE
  + Model version, everything but discriminability held constant
* Test robustness of the sdtlu generation:



…vs. the data generation method:



* + *sdtlu* method slightly overestimates pAUC?
* Test how provided data and sample size of the provided data affects the results:
  + Generate multiple datasets from the same sdtlu model, with different sample sizes
    - Colloff et al., 2021a (Exp 2): High- vs. Low-similarity fillers
      * Lineup size = 6
      * Ns = 500, 1000, 2000, 3000
    - Kaesler et al., 2020: Simultaneous vs. Sequential
    - Palmer et al., 2013: Long- vs. short-delay
      * Lineup size = 8
      * Ns = 500, 1000, 2000, 3000
    - Seale-Carlisle & Mickes (2016): US vs. UK lineups
  + Testing discrepancies:
    - Look at variability in raw pAUC difference in the base data files – does this correlate with variability in the power estimates?
    - Simulate subsets of two data files (e.g., sample ¼ of base dataset for Colloff et al. and Seale-Carlisle & Mickes)
    - Up the number of model fits?
    - Up the number of pROC bootstrap iterations?
* Worked example: sequential vs. simultaneous
  + Kaesler et al. (2020) Simultaneous vs. sequential
  + Seale-Carlisle & Mickes (2016) US (simultaneous 6) vs. UK (sequential 9) lineups

**powe(R)OC revisions**

* Definitions of pAUC and *D* test statistic, more coverage of DPP
* Justify why the analytic methods are not appropriate (e.g., not designed for pAUC)
  + Existing methods that are implemented are only appropriate for full AUC, single or paired ROC curves (e.g., pROC)
  + No user-friendly tools that implement solutions that can be used for pAUC / unpaired ROC curves, and these methods may require information from the data anyway (e.g., variability among observers/participants), make assumptions that may not be valid for eyewitness data (unobserved test results are binormally distributed with equal variances), and effect size metrics may not be intuitive (e.g., overall accuracy, differences in FPR/TPR) (Obuchowski, 2000)
* Consider dropping the current effect size metric altogether
* Inclusion of methods to simulate from model parameters
  + Description of *sdtlu* models

**Full reviews**

Reviewer: 1  
  
Comments to the Author  
This paper describes an R package that conducts power analyses for Receiver Operating Characteristic (ROC) comparisons in eyewitness lineup studies. ROC methods are very popular in the current eyewitness literature, so the proposed package could be a valuable study-planning tool for many researchers. I see a lot of positives in the current approach; for example, the data base of past data sets and power analyses conducted by other researchers is a very nice idea. However, the current version of the package is somewhat limited in functionality and is not grounded in a proven quantitative approach. As such, I have some suggestions for changes.  
  
I will start with a point on exposition: It would be nice to have some worked-through examples described in the manuscript. The online materials have some of this, but I’d recommend going through at least one full, step-by-step application of the package in the manuscript itself.  
  
Maybe the biggest limitation of the current approach is that it relies on bootstrapping from available data sets without consideration for an appropriate statistical or cognitive model of performance. Of course, assuming that a given data set is an effective stand-in for the full population is precarious, and I imagine this can affect the power estimates. How much do the power estimates differ when different “base” data sets are used for the bootstrapping? How does this depend on the sample size of the original data set?  
  
The lack of a deeper model is apparent in the strategy for creating differences between conditions, which is to shift all of the hit rates on an ROC function by a multiplicative factor. The paper admits that this strategy is a bit out-of-step with actual data, and it is not justified by any model of lineup performance that I know of. Letting the user specify different shifts in hit rate across confidence levels is unlikely to produce desirable results. How would researchers know which confidence levels should have bigger shifts than which others? This seems like a place where theoretical guidance from a reasonable model would be critical.  
  
By the way, Figure 5 should be on the log scale. The raw ratios give the impression that deviations above 1 are larger than deviations below 1, but this is an artifact of the measure. (When Condition One is higher, ratios can go from 1 to infinity. When Condition 2 is higher, ratios can go from 1 to zero.)  
  
What should researchers do if they are testing a variable that they expect will affect false-alarm rates in addition to (or instead of) hit rates?  
  
It would be a big advantage to be able to simulate condition differences using a model that does a better job of matching lineup ROC data. The Max-SDT model is a good candidate, and perhaps the current package could call sdtlu as a subsidiary to introduce some of this functionality. For example, researchers could create differences between conditions by changing the distributions representing witness memory, the criterion match strength for making an identification, the confidence criteria, the lineup size, or any combination of these variables. That would provide more flexibility, but also help researchers make choices that are more likely to correspond to real data patterns (as opposed to just putting in a different shift in the hit rate for each confidence level, say).  
  
The proposed effect size measure is interesting, but it would be a good idea to explore its properties before codifying it. Many effect size measures are linked to known sampling distributions. For example, Cohen’s d determines the non-centrality parameter of a t distribution. That is a big ask in this case, but something to think about. Perhaps more realistically, effect size measures are useful when they are robust to extraneous factors. For example, Cohen’s d is prominent because a particular d value is tied to a particular distribution of expected t values from an experiment regardless of the population mean and standard deviation (and somewhat regardless of shape, although a normal distribution is technically required). So by working with d as a measure, researchers can plan for a certain level of power given a certain sample size without worrying about a lot of unknown factors for the planned experiment. Does the proposed effect size measure have this quality?    
  
To elaborate the last point, I recommend that the proposed effect size measure should be tested across situations that vary extraneous factors. For a given value of effect size, can researchers expect the same power across changes in other factors, like overall memory discriminability, criteria positions, lineup size, etc.? One way to do this is to start from “base” data sets that have a wide range of performance patterns, conduct the power analysis using the same effect size and planned sample size, and evaluate the variability in the resulting power values. Another way to do it would be to use a reasonable model to simulate different performance patterns. Ideally, both methods would be explored.  
  
Is it possible to vary lineup size in the current package, or would researchers need to find a “base” data set that used the exact lineup size that they would like to use in their study?  
  
In summary, developing this package was a great idea and it would definitely serve a need in the field. This potential for widespread use is a great reason to make sure that the assumptions are sound and that researchers have the flexibility they need to be creative in designing new studies.   
  
Jeff Starns  
  
  
  
Reviewer: 2  
  
Comments to the Author  
This manuscript proposes a simulation-based power analysis for ROC curves generated from eyewitness lineup experiments. An R shiny app was developed.   
  
The topic of this work is very timely and fills an important gap. The basic idea is excellent and will be of great benefit. The manuscript is fairly well written (but see below) and of appropriate length. The work is suitable for the journal.  
  
I hope a version of this work is published. There are, however, a few issues that I would like to see addressed before publication.  
  
It is unfortunate that in a paper developed to determine power for pAUC that neither pAUC nor the test statistic that is being analyzed are defined or even described. The closest we get is the reference in the caption of Figure 7.  Similar for DPP. Also, in general, DPP is not discussed nearly in depth as pAUC.   
  
The general idea of the next few comments is that the method needs to be tested more thoroughly.  
  
There is a good discussion of the advantages of a bootstrapping/simulation-based method. However, simulation-based methods are not without disadvantages. The lack/minimization of discussion and testing surrounding these disadvantages is problematic. Is the method consistent, asymptotically valid, does independence of samples hold, etc? How strongly does the provided data affect the results (e.g., test with subsets of the data and check variance)?   
  
I understand the reticence to rely on models for the method. However, a potentially informative test would be to compare power analysis using the simulation-based method with data derived from a model (perhaps a SDT model). For example, generate data from the model, use those data to seed the simulations. That way you can determine the accuracy of your method. You could fit a model to the data to get appropriate parameters for testing.    
  
The multiplier assumption for correct IDs is probably a good approximation, however, it would be good to know whether that approximation comes with any bias. Does it depend on the absolute pAUC? Is it possible for all performance levels (e.g., is it possible to assume too many trials)? Etc. Basically, test this assumption a bit more, especially as the provided analysis (Fig 5, which was useful) was equivocal. What happens if this assumption doesn’t hold?  
  
Another major missing piece is an actual application of the method. There is very little justification in the literature for the sample sizes used in published papers. This work would be an excellent opportunity, and it would help the field, to backtrack and determine whether the sample sizes used in a few landmark previous studies were appropriate. Similarly, pick a few very common paradigms (e.g., sim vs seq) and give us some recommendations (feel free to hedge, but even a vague recommendation would be better than the current state of affairs).  
  
Provide a better rationale for why the analytic methods (p6) are not appropriate. Is no part of those methods feasible?  
  
This is probably my lack of understanding, but how does the sample size of the provided data influence the results?  
  
Consider citing the Smith full ROC paper earlier, e.g., p3.  
  
The discussion surrounding Fig 3 is confusing. I was expecting to be shown something like P(SID & Conf=X | Guity) and P(SID & Conf=X | Innocent). Instead, what was shown was P(Conf=x & Guilty | SID) and P(Conf=x & Innocent | SID).   
  
The reference in FN 1 is circular.  
  
P11 L15 vector.  
  
P13 L3 missing word.  
  
P13 L13 this is an effect size, not test statistic.