Heriot-Watt University
Research Gateway

Interaction Effects on Common Measures of Sensitivity: Choice of Measure, Type I Error and Power

Citation for published version:
Rhodes, S, Cowan, N, Parra Rodriguez, M & Logie, RH 2019, 'Interaction Effects on Common Measures of Sensitivity: Choice of Measure, Type I Error and Power', *Behavior Research Methods* , vol. 51, no. 5, pp. 2209–2227. https://doi.org/10.3758/s13428-018-1081-0

Digital Object Identifier (DOI):
10.3758/s13428-018-1081-0

Link:
Link to publication record in Heriot-Watt Research Portal

Document Version:
Peer reviewed version

Published In:
Behavior Research Methods

Publisher Rights Statement:
This is a post-peer-review, pre-copyedit version of an article published in Behavior Research Methods. The final authenticated version is available online at: http://dx.doi.org/10.3758/s13428-018-1081-0

General rights
Copyright for the publications made accessible via Heriot-Watt Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
Heriot-Watt University has made every reasonable effort to ensure that the content in Heriot-Watt Research Portal complies with UK legislation. If you believe that the public display of this file breaches copyright please contact open.access@hw.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 09. Sep. 2024

Interaction Effects on Common Measures of Sensitivity: Choice of Measure, Type I Error and Power

Stephen Rhodes[1], Nelson Cowan[1], Mario A. Parra[2,3], and Robert H. Logie[2]

[1]Department of Psychological Sciences, University of Missouri

[2]Human Cognitive Neuroscience, Centre for Cognitive Ageing and Cognitive Epidemiology, Department of Psychology, University of Edinburgh

[3]Department of Psychology, Heriot-Watt University

Accepted for publication at *Behavior Research Methods*. June 26, 2018

Author Note

Part of this work was completed while the first author was at the University of Edinburgh and was supported by a PhD studentship from the Centre for Cognitive Ageing and Cognitive Epidemiology, part of the UK cross council Lifelong Health and Well-being Initiative (MR/L501530/1). S.R., N.C., and R.H.L are currently supported by the ESRC, Grant ES/N010728/1. Part of this work was presented at the 57th annual meeting of the Psychonomic Society in Boston. The code written to perform these simulations can be found here: `https://github.com/stephenrho/MeasuresAndErrors`.

Stephen Rhodes, 9J McAlester Hall, Columbia MO 65211-2500. Email: rhodessp@missouri.edu.

Abstract

Here we use simulation to assess previously unaddressed problems in the assessment of statistical interactions in detection and recognition tasks. The proportion of hits and false-alarms made by an observer on such tasks is affected by both their sensitivity and bias, and numerous measures have been developed to separate out these two factors. Each of these measures makes different assumptions regarding the underlying process and different predictions as to how false-alarm and hit rates should covary. Previous simulations have shown that choice of an inappropriate measure can lead to inflated type I error rates, or reduced power, for main effects, provided there are differences in response bias between the conditions being compared. Interaction effects pose a particular problem in this context. We show that spurious interaction effects in analysis of variance can be produced, or true interactions missed, even in the absence of variation in bias. Additional simulations show that variation in bias complicates patterns of type I error and power further. This under-appreciated fact has the potential to greatly distort the assessment of interactions in detection and recognition experiments. We discuss steps researchers can take to mitigate their chances of making an error. (194 words)

Keywords: Recognition, Detection, Interactions, Type I Error, Type II Error, Power, Sensitivity, Bias

Interaction Effects on Common Measures of Sensitivity: Choice of Measure, Type I Error and Power

In binary choice recognition and detection paradigms, observers are presented with a sequence of stimuli and are required to distinguish targets (for example, previously studied items) from non-targets (for example, new items). Performance on such tasks is captured by the proportion of *target* responses conditional on whether the probe was actually a target or not. Participants make a *hit* if they correctly identify a target, whereas they make a *false-alarm* if they incorrectly characterize a non-target. The proportion of hits and false-alarms will not only be influenced by an observer's ability to distinguish targets and non-targets (which is usually of primary interest to researchers), but also by their preference for one response option over another. Thus, researchers regularly adopt measures that aim to separate out the sensitivity of observers from their response bias. In order to do this, these commonly used measures make particular assumptions about the process underlying discrimination. If these assumptions are unfounded then researchers are at risk of concluding that two conditions differ in sensitivity when truly they differ in bias, or they may miss true differences between conditions (Rotello, Masson, & Verde, 2008; Schooler & Shiffrin, 2005).

Here we consider the assessment of *interactions* with commonly used single point estimates of sensitivity (d', P_r, A') and find that this presents a particular problem not recognized before in the literature. Typically researchers interested in differences between groups will compare two or more conditions of theoretical interest and conduct an analysis of variance on the resulting sensitivity estimates. Interaction effects, such that certain groups do better or worse under certain conditions, are often interpreted as indicating that a certain process functions more or less well in certain populations. Unfortunately, if the chosen outcome measure is not appropriate for the data at hand the detected interactions may be spurious and real interactions may be missed entirely. We also show that this can happen in the *absence* of variation in response bias; that is, even if all participants adopt the same decision strategy in all conditions. Some basic difficulties in interpreting interactions such as these were well explained by Loftus

(1978), but remain under-appreciated (Wagenmakers, Krypotos, Criss, & Iverson, 2012). Here we illustrate a previously unarticulated problem in interpreting interactions from binary choice detection or recognition experiments that build on these earlier observations.

The article is organized as follows; firstly, three commonly applied measures of sensitivity (d', P_r, A') and their assumptions are described. We then discuss previous simulation studies that have looked at the issue of main effects with point estimates of sensitivity and outline the rationale for the present simulations examining interactions. Following this, we describe type I error and power simulations for the three oft-used measures under different generative assumptions. Extending the previous simulations on main effects, the present work demonstrates that the interpretation of interaction effects poses an even greater problem if the selected measure is inappropriate for the data at hand. Specifically, problems with interactions can arise even in the absence of variation in response bias. Finally, we discuss ways in which researchers may choose a measure which is appropriate for their data, provide an empirical example of how conclusions are liable to be affected if they don't, and discuss approaches if no reasonable measure can be identified.

Commonly used Measures and their Assumptions

Two broad classes of model dominate the analysis of sensitivity in recognition or detection paradigms; namely, the signal detection and high-threshold accounts. We outline these models in turn along with their frequently used estimators of sensitivity.[1] In addition, we discuss a measure outside of these models that is often used in an attempt to avoid making assumptions about the underlying discrimination process, and the arguments that have been made against this measure.

[1]The present analysis deals entirely with commonly used point measures of sensitivity derived from accuracy data (i.e. hits and false-alarms). It should be noted that there are a range of models that allow the joint analysis of accuracy and reactions times, for example the diffusion model (Ratcliff, 1978; Ratcliff, Smith, Brown, & McKoon, 2016) and linear ballistic accumulator model (Brown & Heathcote, 2008). Where applicable these models provide a range of informative parameters and can address questions beyond mere analysis of accuracy (e.g. speed-accuracy trade offs). There are highly accessible tutorials, as well as simplified versions of these models geared towards investigating empirical effects, for interested readers to pursue (see Wagenmakers, Van Der Maas, & Grasman, 2007; van Ravenzwaaij, Donkin, & Vandekerckhove, 2017; Donkin, Brown, & Heathcote, 2011).

Signal Detection Theory. Accounts based on signal detection theory propose that items are evaluated on a continuous decision variable (Green & Swets, 1966; Tanner Jr & Swets, 1954). This evaluation is noisy (either due to the nature of the stimulus or the internal processing) but targets are expected to yield greater values on the decision variable than non-targets, producing two distributions (see Figure 1). The observer must establish a criterion, above which they respond *target* and below which they respond *non-target*. Applications of signal detection theory tend to assume that the underlying evidence distributions are Gaussian as depicted in Figure 1. In this parameterization, the difference between the means of the two distributions is given by d. This parameter captures the sensitivity of the observer, which is their ability to distinguish target and non-target items. As depicted in Figure 1, sensitivity is given in units of the standard deviation of the non-target distribution, which is fixed to 1 (without loss of generality). To account for the possibility that the two distributions have different variances, an additional parameter, s, controls the ratio of the non-target to target standard deviations. Consequently, the standard deviation of the target distribution is $1/s$ (as the non-target standard deviation is fixed to 1. See Figure 1). Here the criterion, k, is placed relative to the mid point between the means of the two distributions. According to this account, when a stimulus is presented, it elicits a value along the decision variable. As the observer does not know from which distribution this value has been sampled, they respond *non-target* if it is smaller than k and *target* if it is greater. Thus, negative k values indicate a liberal criterion (a bias toward responding *target*), whereas positive values indicate a more conservative criterion (a bias toward responding *non-target*).

Assuming the distributions are normal, the probability an observer makes a hit is given by $h = \Phi[(d/2 - k)/(1/s)]$, where Φ is the standard normal cumulative distribution function. Related to Figure 1, this gives the area of both the blue and green shaded areas (the area of the target distribution above the criterion). The probability of making a false-alarm is given by, $f = \Phi(-d/2 - k)$. This is the blue shaded area in Figure 1 (the area of the non-target distribution above the criterion). In

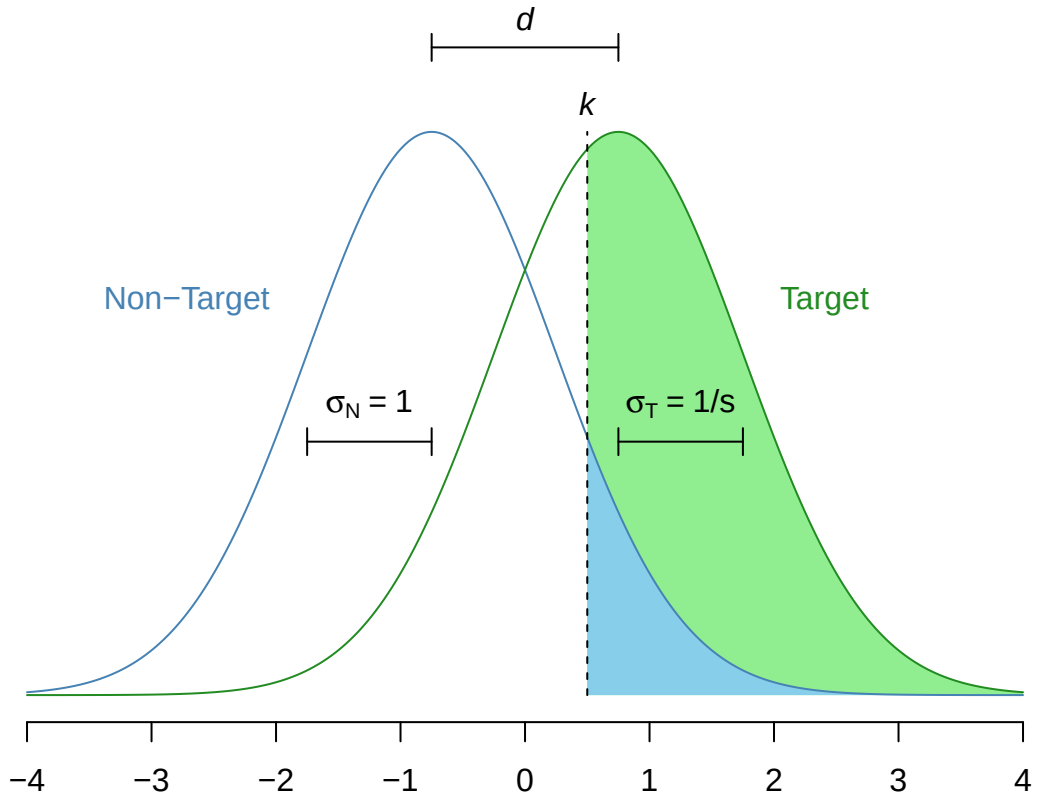

Figure 1. Illustration of Gaussian signal detection theory. The parameter d determines the distance between the means of the non-target and target distributions, while σ_N and σ_T give the standard deviations of these distributions. Finally, k is the criterion at which the observer goes from responding *non-target* to *target*. In this example, $d = 1.5$, $k = 0.5$, and $s = 1$ (equal variance). Hit rate is equal to the area of both the green and blue shaded areas. False-alarm rate is equal to the area of the blue area only.

order to derive a closed form estimator that translates pairs of false-alarm and hit rates into sensitivity, irrespective of response bias, it is necessary to restrict the s parameter (which controls the relative variance of the target and non-target distributions). Figure 1 depicts the case where $s = 1$ and, consequently, the distributions have equal variance. In this case a simple estimate of sensitivity can be determined by rearranging the above equations, $d' = z(h) - z(f)$, where z is the quantile function of the standard normal distribution ($z(x) = \Phi^{-1}(x)$). As the equation shows, d' only serves as an estimate of d provided that the two distributions are standard normals with equal variances, otherwise sensitivity and bias will be confounded (Swets, 1986b).

Each account of the discrimination process predicts different receiver operating characteristic or ROC curves. These theoretical ROC curves plot the relationship

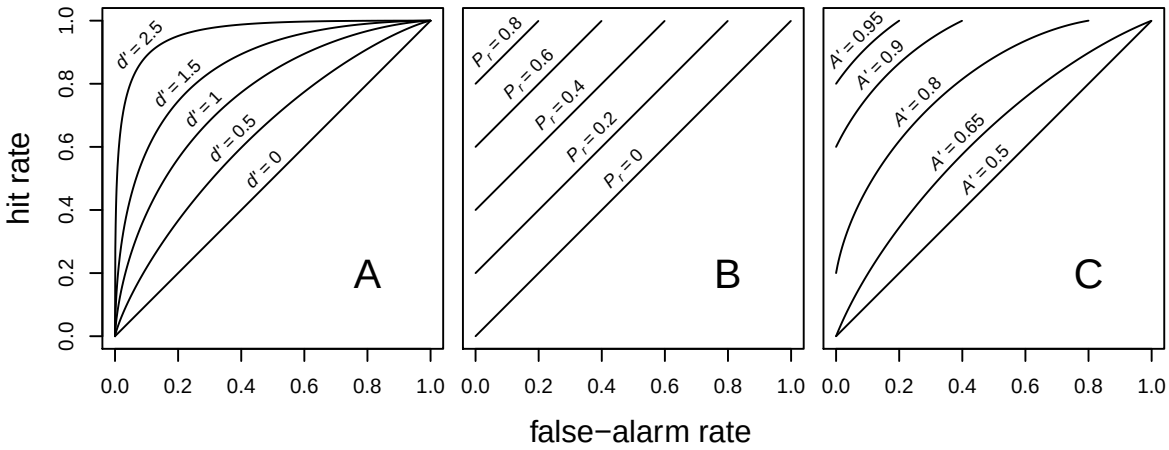

Figure 2. ROC curves predicted by (A) the Gaussian equal variance signal detection theory ($\sigma_N = \sigma_T$), (B) two high-threshold theory ($P_N = P_T$), and (C) the measure, A'.

between false-alarm rate and hit rate for a fixed level of sensitivity across all levels of bias. Figure 2A presents several ROC curves for different values of d'. This measure predicts symmetrical, curved ROCs that become more compressed to the top left of ROC space with incrementally increasing sensitivity. Relaxing the assumption of equal-variance ($s \neq 1$) allows for asymmetrical ROC curves, that have often been observed in the literature (Ratcliff, Sheu, & Gronlund, 1992; Swets, 1986a). The form of these ROC plots becomes particularly important when discussing evidence for interactions, as will be shown later in detail.

High-Threshold Theory. Accounts based on threshold theory propose that, rather than items being evaluated on a continuum, observers enter one of a handful of discrete states (Luce, 1963; Snodgrass & Corwin, 1988). Consequently, the crucial quantities of interest are the probabilities of entering these states. According to this account, observers are said to either detect a certain state of affairs or be in a state of complete information loss and must guess. The most popular variant of threshold theory states that observers cannot detect an incorrect state of affairs (i.e. they cannot enter a target detect state when presented with a non-target and vice versa); that is, the thresholds are 'high' (see Snodgrass & Corwin, 1988). Figure 3 presents the two-high threshold (THT) account of the detection process. The left tree indicates that, when a target is presented (as denoted by T), an observer has a specific probability, P_T, of detecting this state of affairs, in which case they certainly respond *target* and make a

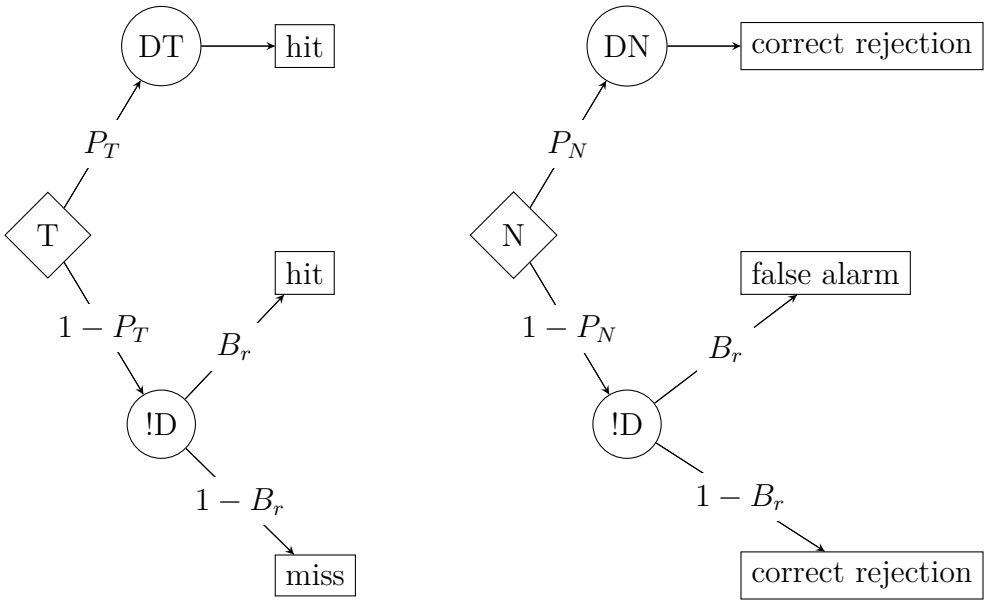

Figure 3. Illustration of the two-high threshold (THT) theory of detection. Diamonds refer to the nature of the item presented (T = target, N = non-target), circles refer to the internal state of the observer (DT = detect target, DN = detect non-target, !D = no detection), and rectangles refer to category of response made. Note that targets cannot lead to the detect non-target state and vice versa, making the model 'high threshold'.

hit. On some proportion of trials ($1 - P_T$), however, the observer does not enter the detect state, and must guess. The probability that the observer correctly guesses that the item is old is given by, B_r. Zero bias towards responding target over non-target is reflected in a B_r of 0.5. The probability of correctly responding *target*, then, is the sum of the branches in Figure 3 ending in a *target* response, $h = P_T + (1 - P_T)B_r$. On trials where a non-target is presented (the right tree of Figure 3 beginning N), observers detect this at a rate of, P_N. Therefore, a false-alarm can only occur if the detect threshold is not passed and participants incorrectly guess. The probability of this occurring is given by, $f = (1 - P_N)B_r$.

Constraints are needed to arrive at a point estimate of sensitivity from a pair of observed false-alarm and hit rates. Assuming that the probability of entering the target detection state is the same as the probability of entering the non-target detection state ($P_T = P_N$, referred to as P_r: Snodgrass & Corwin, 1988) results in a simple measure of discriminability, $P_r = h - f$, which is commonly referred to as hit rate *corrected* for guessing (or corrected recognition).

The two-high threshold model of recognition can also be used to justify the use of

proportion correct, $p(c)$, as an index of sensitivity. As noted by Macmillan and Creelman (2005), the formula for proportion correct can be written as a linear function of hits minus false alarms, $p(c) = \frac{1}{2}(h - f) + \frac{1}{2}$. Thus, using proportion correct to summarize performance (sensitivity) in a recognition task implies a threshold decision model (see also Swets, 1986b), although this is rarely acknowledged. The two-high threshold model can also be expressed in a signal detection framework with rectangular underlying evidence distributions (see Rotello et al., 2008), however the parameterization presented here more clearly outlines the rationale of this measure. The ROC curve predicted by the measure P_r is presented in Figure 2B. Unlike equal variance signal detection, this account predicts linear functions within ROC space with an intercept of P_r and a slope of 1.

The measure A'. As noted by Pastore, Crawley, Berens, and Skelly (2003), the assumptions underlying measures of sensitivity, in particular signal detection theory, have caused some concern among researchers, who have gravitated towards the use of ostensibly assumption free, 'non-parametric' measures of sensitivity. The most popular measure in this vein is A', which was developed by Pollack and Norman (1964) with the aim of estimating the area under the ROC curve. From the estimate of this area the researcher would obtain the expected proportion correct of an unbiased observer on a two-alternative forced choice task, irrespective of the underlying process (Green, 1964). This would achieve the goal of obtaining a truly assumption free measurement of an observers' ability to distinguish targets from non-targets (Bamber, 1975; Green & Moses, 1966). With a single hit and false-alarm pair, however, it is only possible to get an average of the minimum and maximum areas implied.

As A' was derived without specific reference to the shape of underlying evidence distributions, it gained the title 'non-parametric'; however, Macmillan and Creelman have shown that A' does indeed harbor parametric assumptions—in that the shape of the ROC curve (see Figure 2C) does imply underlying distributions of evidence. In fact, they demonstrated that the implied distributions change with the sensitivity of the observer (Macmillan & Creelman, 1996). As shown in Figure 2C, when performance is

high, A' is consistent with a threshold model assuming (roughly) rectangular distributions, as the ROCs are approximately linear. However, as performance lowers, the shape of the A' ROC curve increasingly mimics that of a signal detection model assuming logistic evidence distributions (see Macmillan & Creelman, 1990, for a critique of bias measures associated with A'). No currently offered theory of recognition or detection makes the assumption that the underlying evidence distributions change with greater sensitivity, making this a problematic characteristic of A'. Further, while the initial aim behind the development of A' was to obtain a compromise between the minimum and maximum areas implied by the single point, A' fails to do this (see Smith, 1995; Zhang & Mueller, 2005). Despite these issues, A' is still regularly used, so we include it in the present simulations. Grier (1971) and Aaronson and Watts (1987) provide the formulae for A' for above and below chance performance, respectively:

$$A' = \frac{1}{2} + \frac{(h-f)(1+h-f)}{4h(1-f)}, \quad h \geq f$$

$$A' = \frac{1}{2} - \frac{(f-h)(1+f-h)}{4f(1-h)}, \quad h < f.$$

Previous Simulations: Main Effects and Differences in Response Bias

Looking closely at Figure 2, one can see that for two conditions in which participants have equal sensitivity, the different sensitivity metrics will be in agreement *provided that* response bias is also equivalent. This is because, in this situation, the (f, h) pair from each condition occupies the same point in ROC space. As the effect of changing bias is to move the (f, h) point along the ROC curve, differences between conditions in bias produce disagreements between different measures of sensitivity. For example, imagine two conditions generated by a signal detection process with the same sensitivity (d') but where one condition has a neutral criterion and the other condition has a conservative response bias (i.e. $k > 0$). The more conservative condition is shifted along the same ROC curve as the neutral condition, but the effect of this is to produce a sensitivity difference when, say, P_r is calculated. This is an example of type I error, but an analogous situation can also be conceived where two conditions with truly

different sensitivities (i.e. points on different ROC curves) appear to result in equal performance when an inappropriate measure is applied (i.e. a type II error). To our knowledge, two studies have looked at the ramifications of this via simulation.

Schooler and Shiffrin (2005) explored the efficacy of sensitivity measures when the available data are sparse due to small trial numbers per participant. Their simulated data were generated using a equal-variance signal detection model, in which two conditions either did or did not differ in terms of sensitivity. This allowed them to assess power and type I error rate, respectively. Provided the two conditions did not differ in terms of criterion placement or bias, the high threshold measure, P_r, performed well, with type I error rates around the accepted value of 0.05. However, when conditions differed in criterion placement, type I error rates for this measure were high (up to 34%) and, unsurprisingly, d' performed better. When conditions truly differed in terms of sensitivity, power was greatly improved by using d' relative to P_r, again unsurprisingly as it matched the generative model.

Rotello et al. (2008) provided a more comprehensive set of simulations in which they generated data using either an underlying signal detection or two-high threshold structure. They assessed the type I error rates of repeated measures t-tests on multiple commonly used measures—including d', A' and proportion correct. They found that, with data simulated to have true differences in response bias between conditions, use of a measure not matching the assumptions of the generative model (e.g. applying P_r to data simulated from a signal detection model) resulted in type I error rates typically in excess of 20%. Critically, the error rate associated with A' was large regardless of the true underlying distributions. They also conducted power simulations for their measures where there were true differences between hypothetical conditions in sensitivity, but identical bias. All measures were found to perform fairly well, including A', especially with low overall sensitivity and small numbers of trials. However, given its unacceptably high type I error rates, Rotello et al. (2008) council against the use of A' in any situation.

The Present Simulations: Problems Interpreting Interactions, Even with Equal Bias

The work of Schooler and Shiffrin (2005) and Rotello et al. (2008) clearly shows that, in the case of a comparison between two experimental conditions (or two groups), a misguided choice of sensitivity measure can result in errors *provided there are differences in response bias*. The reason for this is clearly seen in the ROC plots depicted in Figure 2; variation in bias between conditions results in two different points along the same ROC curve and, consequently, different measures of sensitivity give discrepant results.

One contribution of the present work is to point out that, for *interactions* between experimental conditions in ANOVA, it is possible for measures of sensitivity to produce discrepant results *without* any variation in response bias. The only condition that must be met for this disagreement to occur is that each experimental factor has a main effect on sensitivity.

To illustrate this, the simplest situation to consider is a 2×2 design; here we consider one between-subjects factor (e.g. age group) and one within-subjects (experimental condition). Figure 4 depicts two situations where, in absence of variation in response bias, P_r and d' would give opposing answers to the question of a group $\times$ condition interaction effect in a standard analysis of variance. The top row presents the case where for P_r there are two main effects but no interaction. However, when d' is calculated using the (f, h) pairs, a clear interaction effect emerges, where the effect of condition for group 1 is smaller than it is for group 2. The reason for this distortion can again be seen in the ROC plots of Figure 2; at increasingly high levels of sensitivity (d') the curves implied by signal detection theory become more compressed and, thus, the equally spaced points in ROC space shown in Figure 2A imply increasingly large values of d', resulting in an *over-additive* interaction. This analysis would imply that the effect of condition is larger for the higher performing group.

The bottom row of Figure 4 presents the case where no interaction is present using d', but an *under-additive* interaction is present with P_r, such that the difference

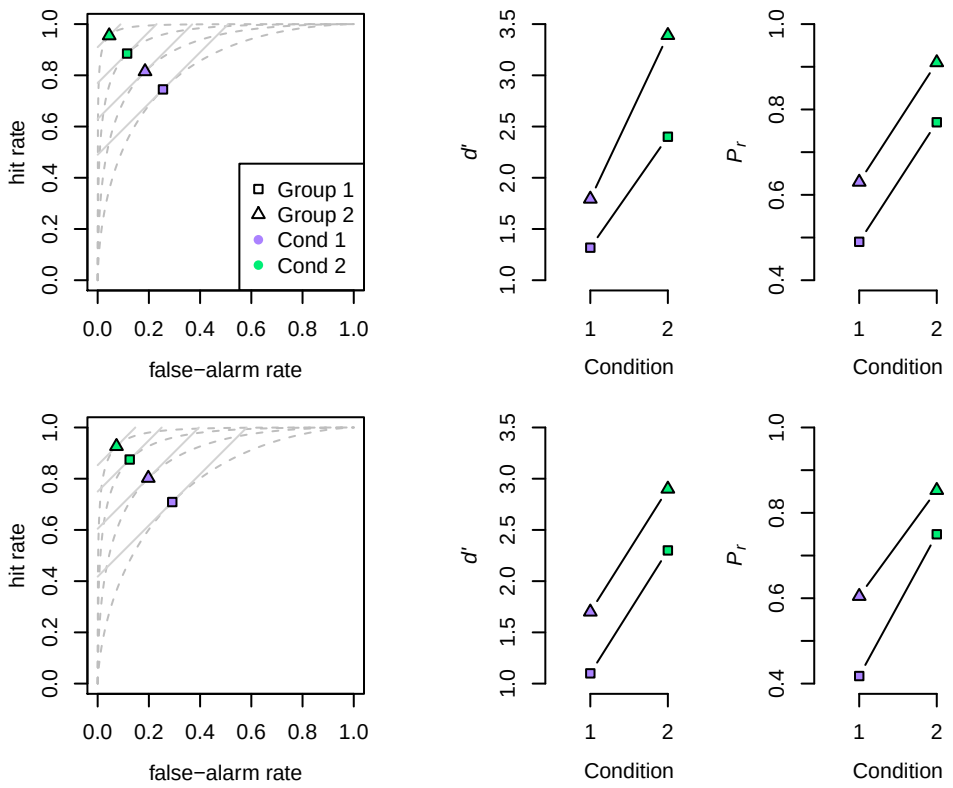

Figure 4. Plots showing that, when considering interactions, it is possible for the commonly used two-high threshold and signal detection measures, P_r and d', to disagree without variation in bias. Top row: non-interaction with P_r but interaction using d'. Bottom row: interaction using P_r but main effects only with d'.

between conditions is smaller for the higher performing group. It is important to reiterate that this disagreement arises without any difference in response bias (all the points lie on the negative diagonal indicating neutral responding) and occurs when each experimental factor produces a main effect. No variation in bias and no main experimental effects on sensitivity would result in four overlapping points in ROC space and, thus, no disagreement between the measures. Crucially, the data from this simple 2×2 design will not signal which is the correct measure to use. This decision must be based on empirical ROC data and other tests aimed at probing the generative process underlying the specific task in question (see the General Discussion).

As the above shows, the different scaling applied by commonly used measures of sensitivity has the potential to produce large disagreements regarding interactions in standard analysis of variance. Given the theoretical or explanatory weight often bestowed on interactions such as these (Wagenmakers et al., 2012), this should be of

great concern to researchers. The extent to which this has affected the literature on detection and recognition (especially that on group differences across different experimental conditions) is difficult to gauge (although see Rotello, Heit, & Dubé, 2015, for examples where the choice of outcome measure may have greatly affected conclusions). Simulation allows us to examine the extent to which these issues could conceivably cause problems in interpreting interactions for reasonable research designs. Consequently, in the present simulations we assess a wide range of parameter values, for both sensitivity and bias, across equal and unequal variance signal detection, and two-high threshold generative models as well as assessing the influence of number of simulated subjects per group or trials per condition. There are potentially limitless combinations of these factors, so the code used to produce the present simulations is available at `https://github.com/stephenrho/MeasuresAndErrors` for researchers interested in assessing specific parameter settings. We report a subset of the simulations here which serve to illustrate, and elaborate on, the main points made so far. The results of all simulations are presented in Supplementary Material.

Structure of Simulations

Simulated data sets contained two groups of N_S hypothetical subjects each performing in two within-subjects conditions with the same number (N_T) of target and non-target trials per condition. Each set of trials was drawn from independent Binomial distributions, with the probability of a hit or false-alarm determined by the underlying model parameters (i.e. d and k, or P_r and B_r) for a given subject in a given condition. For both signal detection and two-high threshold simulations the underlying parameter values, p, that controlled sensitivity and bias were determined using the same linear equation:

$$p_{i,b,w} = \beta_0 + \beta_1 x_b + \beta_2 x_w + \beta_3 x_b x_w + b_i,$$

where β_0 is the grand mean (for example, average true sensitivity, d) and β_1 and β_2 are deflections from the grand mean associated with the factors B (between subjects,

group) and W (within subjects, condition), respectively.[2] The indicator variables, x_b and x_w, are set to -1 if the observation comes from the first level of the factor or to 1 for the second level. Consequently, positive values of our deflection parameters mean lower parameter values at level 1 relative to level 2. The value β_3 is the multiplicative interaction between the two variables ($x_b x_w$); so as long as $\beta_3 = 0$ there is no interaction present in the simulated data set. The final component, b_i, reflects random variation in the parameter value due to the i^{th} simulated subject. This is assumed to be normally distributed with a mean of zero and standard deviation, σ_S. Note that this only affects the overall level of performance between hypothetical participants and there is no variability associated with the effect of the within subjects factor.

Both sensitivity (d or P_r) and bias (k or B_r) parameters were determined using this linear equation. These general parameters are referred to throughout this manuscript with superscripts denoting the model parameter to which they refer; for example, a main effect of group on sensitivity with a signal detection generative model would be indicated in the magnitude of $\beta_1^{(d)}$, whereas an effect of condition on two-high threshold response bias is determined by $\beta_2^{(B_r)}$. For the high threshold simulations, as the parameters are constrained to fall between 0 and 1 values outside this range were rounded up to 0 or down to 1. Each simulation proceeded as follows;

1. Parameter values for the generative model were derived for N_S hypothetical participants using the above equation with the present settings for β_0, β_1, β_2, and β_3. Participant variability was sampled from a normal distribution with a mean of zero and a standard deviation, σ_S.

2. Predicted h and f rates were determined according to the generative model (see description of models above).

3. N_T target and non-target trials were simulated for each condition for each subject as samples from a Binomial distribution, with success probability determined by h

[2] In many applications of signal detection theory the symbol β is used to refer to the likelihood ratio at the criterion, k (see Macmillan & Creelman, 2005, pp. 33). However, in this case we use β, with super- and sub-scripts, to refer to the intercept, main effect, and interaction parameters in the linear equation above, that in turn determined the settings for the generative model in our simulations.

for target trials and f for non-target trials.

4. d', A', and P_r were calculated with the simulated hit and false-alarm rates (see formulae above).[3]

5. Steps 1-4 were repeated 1000 times per combination of parameter settings and ANOVAs were conducted on the resulting data sets to obtain the frequency of $p < 0.05$ for the group $\times$ condition interaction.

Here we focus on a subset of the type I and type II error simulations where there are no true condition or group differences in response bias. Full results of all the simulations conducted are presented in Supplementary Material and the code to conduct them is available online: `https://github.com/stephenrho/MeasuresAndErrors`.

Type I Error Simulations

In these simulations, we were interested in assessing the type I error rates of the different measures in the absence of variation in response bias. To this end, we varied the overall sensitivity (controlled by the parameter β_0 in the linear equation above) of our hypothetical observers as well as the main effects of group (β_1) and condition (β_2). To reduce the number of simulations, we fixed the size of the two main effects to be equal. In addition, we also assessed the effect of increasing the number of participants per group (N_S) or trials per condition (N_T) on error rate.

Two-High Threshold

For the simulations with the threshold model, the grand mean bias parameter ($\beta_1^{(B_r)}$) was set to 0.5 (neutral) and there were no main effects or participant variability in bias (that is, $\beta_1^{(B_r)} = \beta_2^{(B_r)} = \sigma_S^{(B_r)} = 0$). Of primary interest here was variability in grand mean sensitivity (or probability of detection) and main effects of group and

[3] In calculating d', one encounters problems with undefined z scores with proportions of 0 or 1. Prior to calculating d', hit and false-alarm rates of 1 or 0 were adjusted by 0.01, which is the simplest and perhaps most commonly used approach. Previous simulations have suggested that the correction applied has little effect on error rates (Rotello et al., 2008; Schooler & Shiffrin, 2005) therefore we did not consider this further.

condition. Mean sensitivity ($\beta_0^{(P_r)}$) was varied from 0.4 to 0.8 in steps of 0.1 and random participant variability in sensitivity was simulated by setting the standard deviation parameter, $\sigma_S^{(P_r)}$, to 0.1. Our primary interest was in the influence of main effects on sensitivity, so the main effect parameters for group ($\beta_1^{(P_r)}$) and condition ($\beta_2^{(P_r)}$) took the values 0, 0.025, 0.05, and 0.1, with the constraint that they were fixed to the same magnitude. Scaled in terms of the random intercept parameter mentioned above, these values represent no, small, medium, and large effects, respectively. These settings allowed us to cover a wide range of parameter values and hit/ false-alarm rates. The range of hit and false-alarm rates across simulations was comparable across the two-high threshold and signal detection simulations (presented next). For both, the most sensitive group and condition were expected to achieve almost perfect performance ($h \approx 1$, $f \approx 0$; i.e. a ceiling effect), whereas the least sensitive group/ condition was close to chance ($h \approx .6$, $f \approx .4$). Note that the main effects were also scaled in the same way relative to the random participant standard deviation parameter for both the THT and SDT simulations (0%, 25%, 50%, 100% for none, small, medium, and large, respectively).

Finally, in separate simulation runs we varied the number of hypothetical subjects (N_S) or the number of trials (N_T) through 12, 24, 48. The results of increasing the number of participants per group or trials per condition were practically identical, therefore we present the former here and the latter in the supplementary material (see Simulation 2).

The results of this simulation with a threshold-generative model are presented in Figure 5. As can be seen in this figure, type I error rates remain under control until there are large main effects of group and condition. As shown in the rightmost panels, error rates for both A' and d' depend on the overall level of discriminability ($\beta_0^{(P_r)}$). The frequency of errors is more pronounced with d' relative to A' until the very highest grand mean sensitivity is reached, at which point the error rate for d' drops to a similar level as P_r. As noted above, erroneously applying d' to data conforming to the predictions of a two-high threshold model leads to the impression that sensitivity

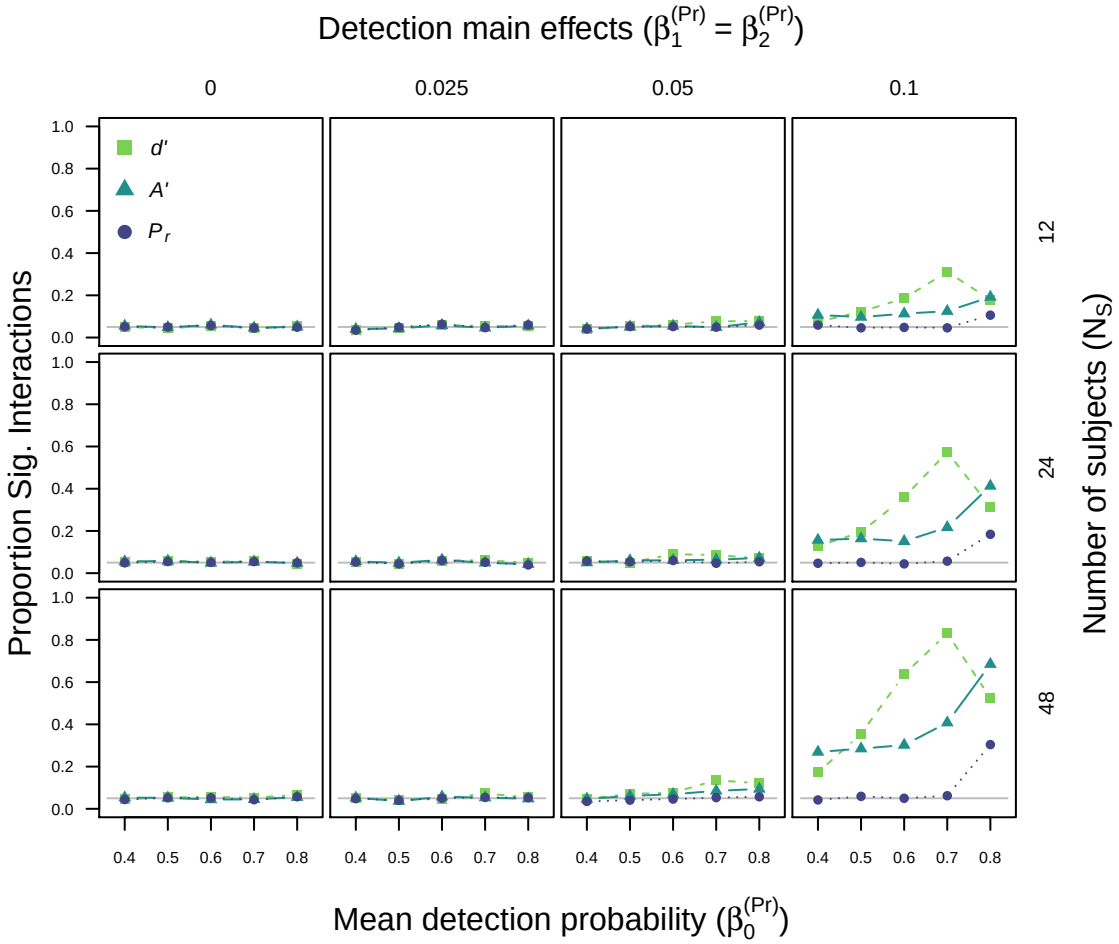

Figure 5. Type I error rates for d', A', and P_r with an underlying two-high threshold (THT) model. This simulation varied overall detection probability (x axis), magnitude of main effects on detection (left to right), and number of simulated participants (top to bottom).

differences between conditions are *larger* in groups that are more sensitive overall (an *over-additive* interaction effect). As overall performance improves (moving along the x-axis), this distorting effect becomes even larger, rising to an error rate in excess of 80% with groups of 48 hypothetical subjects. The ceiling effect encountered at high underlying levels of sensitivity (yeilding hit rates around 1 and false-alarm rates around 0 in the highest performing group) reduces this *over-additive* tendency, seen as a drop of error rate at the highest mean sensitivity in Figure 5. The effect of ceiling discriminability on type I error rate is also seen when the correct measure, P_r, is applied (see the rightmost panels of Figure 5). Finally, the effect of increasing the number of subjects is to exacerbate these error rates as the analysis converges on the wrong

pattern of means (Rotello et al., 2015).

In further simulations, we introduced individual differences in bias (B_r) and varied the overall grand mean bias, which did not affect the pattern of errors greatly (see Simulation 4 in supplement). Introducing main effects of both group and condition on response bias, along with the main effects on sensitivity, had a big effect on errors, however. This was especially true for d', whose errors were magnified by a conservative overall bias (i.e. $B_r < 0.5$) and large main effects on this parameter. As a logical extension of the simulations reported by Rotello et al. (2008) and Schooler and Shiffrin (2005), we also assessed the effect of manipulating bias between groups and conditions in the absence of true main effects on sensitivity. d' was the only measure to exhibit inflated type I errors in this simulation (around 20%) when the bias effects were large. Interested readers can find the full results of these simulations in the supplementary material accompanying this article.

Signal Detection

According to the formulation of the signal detection model described above (see Figure 1), a criterion of zero represents a neutral bias. Therefore, for these simulations the grand mean criterion ($\beta_0^{(k)}$) was fixed to 0 and there were no main or participant effects on criterion placement ($\beta_1^{(k)} = \beta_2^{(k)} = \sigma_S^{(k)} = 0$). Overall sensitivity was varied by having the grand mean parameter ($\beta_0^{(d)}$) take several values (1.5, 2, 2.5, 3, 3.5). Similarly, the magnitude of the main effects of group ($\beta_1^{(d)}$) and condition ($\beta_2^{(d)}$) on sensitivity were also varied (0, 0.125, 0.25, 0.5) with the two fixed to be the same value. The random participant effect had a standard deviation ($\sigma_S^{(d)}$) of 0.5. These settings allowed us to cover a wide variety of parameter values and hit/ false-alarm rates. As in the previous simulations with a threshold generative model, combinations of these parameter settings were assessed with varying numbers of hypothetical subjects or (in a separate run) varying numbers of trials. Once again the results of these two sets of simulations were remarkably similar (see Simulation 2 in the supplementary material).

Figure 6 presents the estimated type I error rates for d', A', and P_r with an

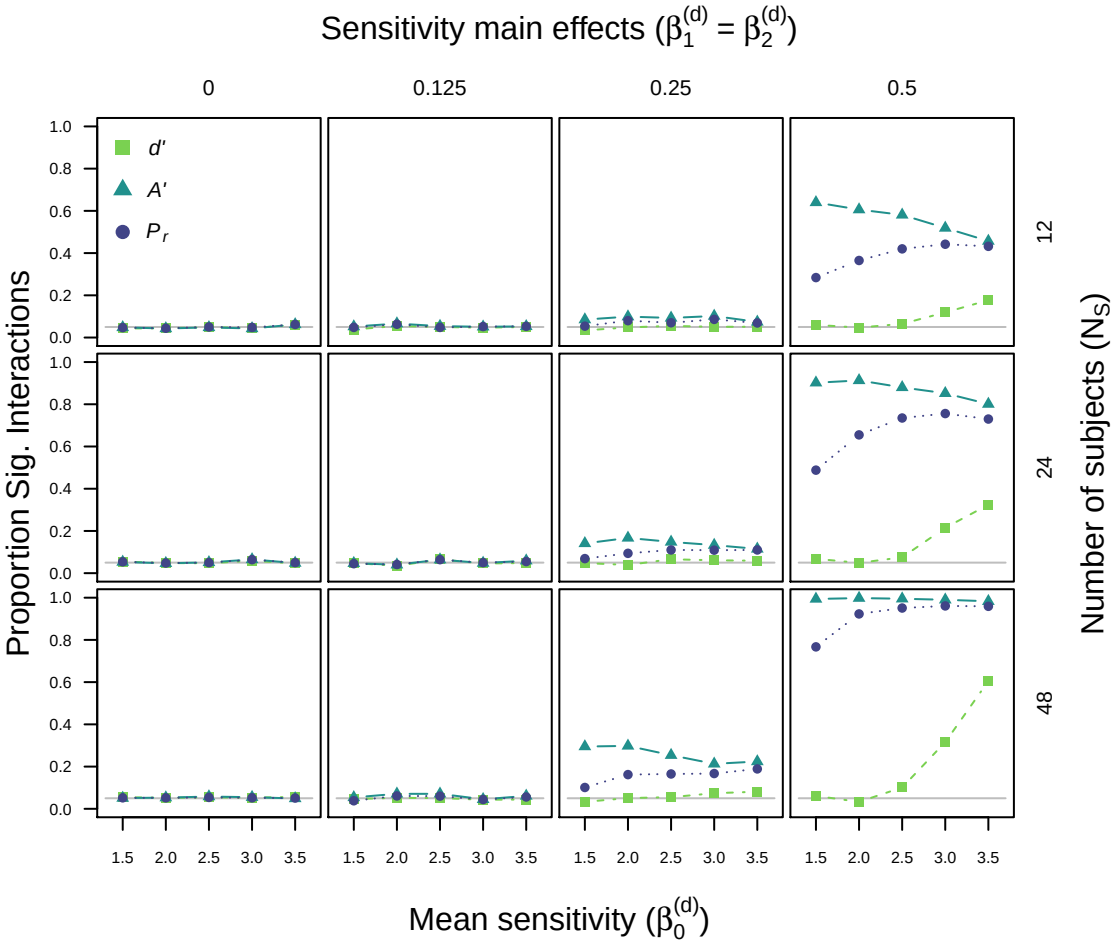

Figure 6. Type I error rates for d', A', and P_r with an underlying Gaussian equal-variance signal detection theory (EV-SDT) model. This simulation varied overall detection sensitivity (x axis), magnitude of main effects on sensitivity (left to right), and number of simulated participants (top to bottom).

underlying equal variance signal detection model. A couple of noteworthy patterns standout; firstly, error rates for all measures are at, or around, the conventionally accepted rate of 0.05 when there are no main effects or the true main effects of group (the between-subjects factor) and condition (the within-subjects factor) on sensitivity are small. When there are medium-to-large main effects, the error rates for A' and P_r depart from accepted levels, and increasing the number of participants per group exacerbates this (Figure 6 right-hand panels). Of the measures, A' clearly is more likely to erroneously give evidence for an interaction effect where none exists.

The effect of increasing the overall mean level of sensitivity is slightly more complex (moving along the x-axis of Figure 6). Using P_r on signal detection data

becomes increasingly problematic as the underlying sensitivity increases, whereas for A' errors become somewhat less pronounced as mean sensitivity increases. For both of these measures, however, when there are large main effects, the type I error rate is unacceptable (around 0.7 for P_r and 0.8 for A' with 24 participants per group). d' also exhibits exacerbated error rates for the interaction test when underlying sensitivity is high and there are large main effects, despite the fact that this measure is consistent with the generating model. This is clearly due to a ceiling effect, which restricts performance of the highest performing group, producing an under-additive interaction where the effect of condition appears smaller in the better performing group. This begins to appear in the simulations where grand mean sensitivity ($\beta_0^{(d)}$) is 3 and the main effects are large (0.5). In this case the expected hit and false-alarm rates of the highest performing group are 0.98 and 0.02, respectively.

We also conducted several simulations in which bias was allowed to vary, the results of which are presented in the supplement. To summarize, allowing for individual differences and overall shifts in criterion placement (i.e. non-neutral) had a small effect on error rates. Mainly, type I error rates for all three measures were less pronounced as the observers departed from neutral overall responding, either towards a liberal or conservative criterion placement. Introducing main effects of both variables had a pronounced effect on error rates (see Simulation 6 in supplement). As the size of these main effects increases, type I error rates rise for both P_r and A', particularly when the overall criterion placement is neutral. We discuss reasons for this pattern of results in the supplement. This tendency for main effects on criterion placement to inflate type I errors for interactions was also present when main effects on sensitivity were removed (see Simulation 7). Thus, in line with previous findings concerning type I error for main effects (Rotello et al., 2008; Schooler & Shiffrin, 2005), it is possible to generate spurious interactions purely from differences in response bias.

In these simulations, we made it so that the two Gaussian evidence distributions shared variance (that is, $s = 1$). However, when empirical ROC curves are plotted, this assumption is often violated (see, e.g., Ratcliff et al., 1992; Swets, 1986a). Therefore, we

also looked at situations where targets produced a more variable distribution over the decision variable than non-targets ($s = 0.8$) and vice versa ($s = 1.2$). For each simulation the same general patterns of type I error rates (and power, see below) were visible as in the equal variance case, although they were slightly less pronounced for $s = 0.8$ and slightly more pronounced for $s = 1.2$. Interested readers can find the full results of these simulations in the online supplementary material.

Summary of type I error simulations

The simulations reported above (and those in the supplement) confirm the argument made on the basis of Figure 4 and show that choosing a measure of sensitivity that mismatches the model underlying the decision process can lead to spurious evidence for interaction effects in detection or recognition experiments. Crucially, this happens without systematic variation in response bias when the factors in the experimental design produce main effects. In follow up simulations, we fixed the number of participants per group and trials per condition to 24 and orthogonality varied the magnitude of group and condition main effects. This confirmed that as long as one effect is large and the other is at least medium in size, inflated type I error rates occur (see Simulation 3 in the supplement).

Introducing systematic variation in response bias complicates the interpretation of interactions further and can lead to type I errors even when there are no true main effects on sensitivity. The effect of collecting more data, either in terms of numbers of subjects or numbers of trials, is to exacerbate these effects even further as the ANOVA settles on the incorrect interpretation of the data (see also Rotello et al., 2015). This should be of great concern to applied researchers who want to interpret interactions arising from detection or recognition tasks. In the discussion we discuss potential strategies to mitigate these problems, but first we move to the topic of power to detect sensitivity interactions when they do exist.

Power Simulations

Another implication of Figure 4 is that there are likely to be occasions where actual interaction effects on the correct scale of measurement are missed. Specifically, it seems that over-additive interactions with a signal detection underlying model may be missed by choosing P_r, whereas an under-additive interaction on two-high threshold's detection parameter may be missed with d'. The next set of simulations addressed this.

As described above in the *Structure of Simulations* section, the indicator variables x_b and x_w were set to -1 if the observation came from the first level of the factor or to 1 for the second level. Using positive values of our β_1 and β_2 coefficients is then useful for simulating interaction effects in the current context. A negative coefficient for the interaction (β_3) produces an *under-additive* interaction effect, where the effect of condition is less pronounced for the better performing group (or, equivalently, group differences are less pronounced in the easier condition). A positive coefficient, on the other hand, produces an *over-additive* interaction where there is a greater effect of condition in the higher performing group (or a greater group difference in the easier condition). We used this to assess the power (1− type II error) of the three measures under different generative models. There was no variation in response bias between groups, conditions, or individuals. We examined power under varying numbers of participants per group (presented here) and trials per condition (presented in the supplementary material, Simulation 9).

Two-High Threshold

This set of simulations used the same general parameter settings as the earlier type I error simulations, but this time introduced interaction effects that were either under- or over-additive. As mentioned above, negative values of $\beta_3^{(P_r)}$ produce under-additive interactions, whereas positive values produce over-additive interactions. This parameter was varied through, -0.05, -0.025, 0.025, 0.05.

As these simulations have the additional variation in the interaction coefficient, Figure 7 presents only the results from the simulation where the true underlying main

effects on sensitivity were large ($\beta_1^{(P_r)} = \beta_2^{(P_r)} = 0.1$). It is in this condition where the differences between measures were most pronounced, however we summarize the other results here. At a small main effect size ($\beta_1^{(P_r)} = \beta_2^{(P_r)} = 0.025$) power for the interaction effect was more-or-less identical for all three measures. Power began to diverge between measures for the medium main effect size (0.05), with d' slightly under-powered for under-additive interactions and somewhat overpowered (relative to the correct measure, P_r) for over-additive interactions, as we would predict from Figure 4. Figures depicting these findings are presented in the supplement under Simulation 8.

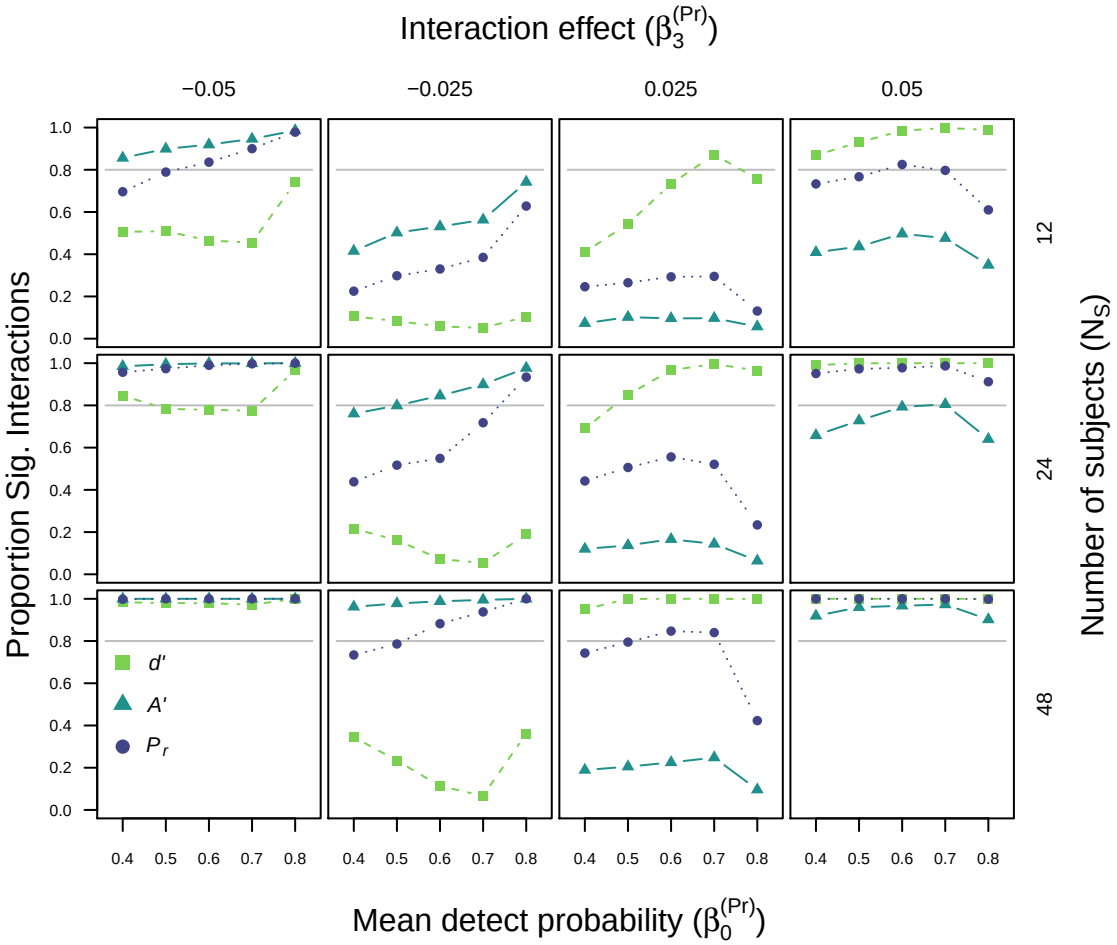

Figure 7. Power for d', A', and P_r with an underlying two-high threshold (THT) model. This simulation varied overall probability of detection (x axis), the magnitude and direction of interaction (left to right) and the number of participants per group (top to bottom). Note: this figure depicts the case where main effects on detection equal 0.1.

Figure 7 plots estimates of power for the interaction effect when there are two large main effects on detection probability. Comparing the two left columns to the two

right ones it is clear to see that, relative to the correct measure P_r, A' is overpowered[4] for under-additive interactions (left) and tends to miss over-additive interactions (right). This pattern is reversed for d', which tends to miss under-additive interactions, as indicated by its low power. Figure 4 provides the explanation for this, in that conversion of P_r onto d' scale produces an exaggeration of differences at higher levels of sensitivity; therefore, when the 'true' interaction is under-additive the two cancel each other out to a large extent. Finally, P_r appears to miss over-additive interactions at very high levels of performance due to a ceiling effect for the most sensitive group, which limits the visibility of the over-additive interaction (see rightmost panels).

Once again, in additional simulations, presented in full in the supplement, we also assessed the effect of manipulating bias on power. Introducing participant variability and different overall levels of bias did not affect estimates greatly. Adding large main effects on response bias complicated things further and was sufficient to cause distortions in power even without group or condition effects on sensitivity.

Signal Detection

For the signal detection simulations, we used the same parameter values as the initial type I error simulations with the additional interaction effect on sensitivity ($\beta_3^{(d)}$), which could either be under-additive ($-0.25, -0.125$) or over-additive (0.125, 0.25).

At a small main effect on sensitivity ($\beta_1^{(d)} = \beta_2^{(d)} = 0.125$) there is not much separating the measures in terms of power. With 24 participants in each group and a larger interaction effect (-0.25 or 0.25), power is always well above 80%. The measures diverge with two slightly bigger main effects (0.25), with A' and P_r producing more misses, relative to d', for under-additive interactions but exhibiting artificially inflated power for over-additive interactions. The clearest divergence is shown in Figure 8 which depicts the results of simulations with two large main effects (0.5) on sensitivity. The results not depicted here can be found in the supplementary material.

As can be seen in Figure 8, when the true generative model consists of a small

[4]It is worth noting that being 'overpowered' is not a good thing in this case, as it arises from an exaggeration of the true effect. Estimates of effect size with this measure will be incorrect and biased upwards.

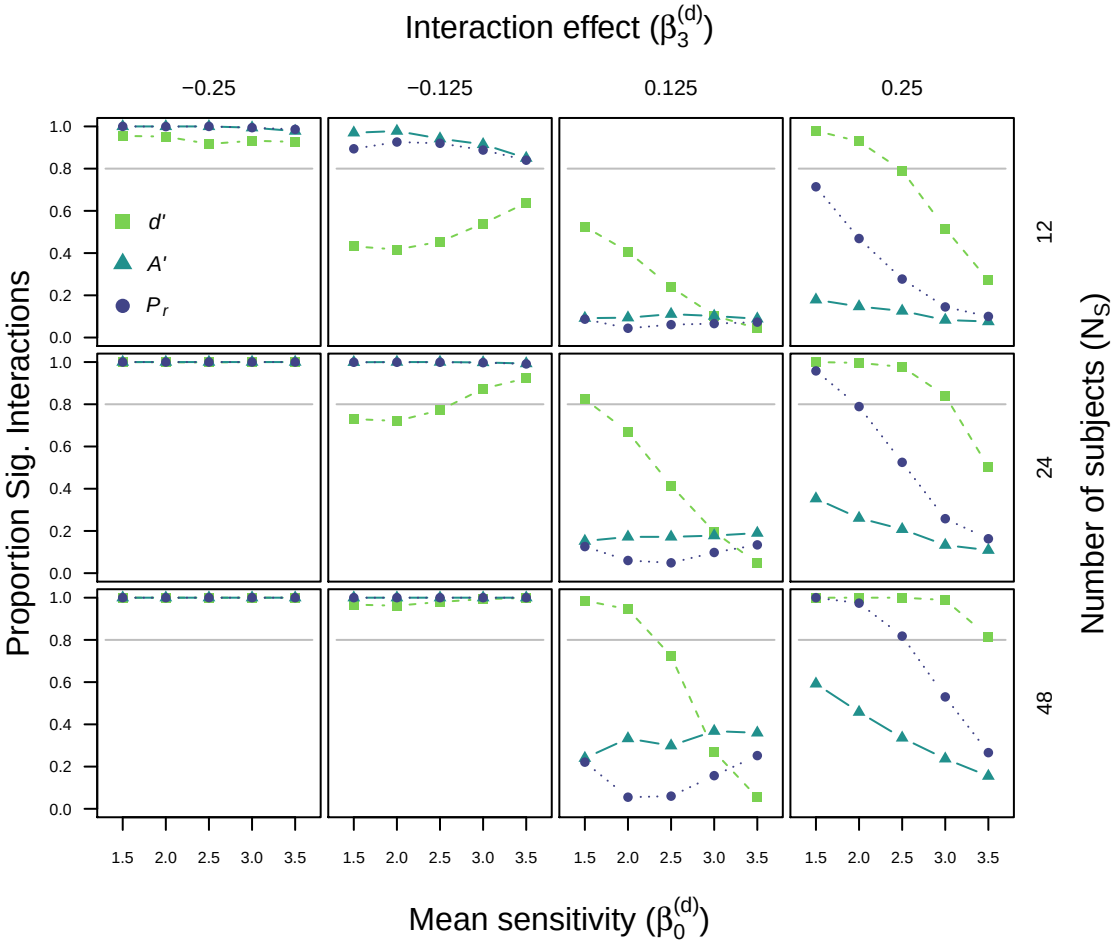

Figure 8. Power for d', A', and P_r with an underlying Gaussian equal variance signal detection theory (SDT) model. This simulation varied overall sensitivity (x axis), the magnitude and direction of interaction (left to right) and the number of participants per group (top to bottom). Note: this figure depicts the case where main effects on sensitivity equal 0.5.

under-additive interaction and a small sample size per group A' and P_r produce higher estimates of power than the correct measure, d', due to their tendency to distort data conforming to a signal detection model (see Figure 4). Their theoretical ROCs do not imply the same compression towards the top-left of ROC space that d' does and, consequently, they make differences at higher levels of performance appear less pronounced. This is inverted for over-additive interactions, for the exact same reason, and d' has consistently greater power, except for cases where ceiling level sensitivity quashes the interaction effect (as can be seen progressing along the x-axis). A' performs particularly badly for a large over-additive interaction (see right most panels of Figure 8).

Additional power simulations, in which overall criterion placement was varied, showed that for over-additive interactions the power of the correct measure, d' dropped as criterion deviated from neutral, whereas power with P_r was inflated. Adding main effects on criterion placement complicated things a great deal (see the supplement for in depth discussion) and was sufficient to distort estimates of power even in the absence of main effects on the sensitivity parameter.

Summary of power simulations

The power simulations reveal a complex set of results wherein evidence for interactions is either suppressed or erroneously magnified, depending on the nature of the true interaction effect (either under- or over-additive) and the generative model. Analyses of interactions conforming to a two-high threshold discrimination process with the measure d' tend to miss under-additive effects, where group differences are suppressed in higher performing conditions. Applying A' to this data, however, one would be liable to miss over-additive interactions where group differences are more pronounced at higher performance levels. With data conforming to a signal detection theory account, applying A' or P_r will lead to an increased likelihood of missing a true over-additive interaction effect. This happens because of the different scaling implied by the different measures and does not rely on variation in bias, although adding this variation complicates matters further (see Supplement).

General Discussion

Binary choice recognition or detection performance summarized in terms of hit and false-alarm rates confounds the observer's ability to perform the discrimination with their inherent bias towards one of the response options. In order to separate the contribution of these two factors, a model of the underlying decision process is needed. Commonly used measures, d' and P_r, come from two broad classes—those that assume a continuous decision variable, with observers comparing sampled values to a criterion, and those assuming a handful of discrete states—and are particularly restricted realizations of these models. The commonly applied measure A', on the other hand, does

not specify the basic process underlying the discrimination but, that being said, *does* imply specific evidence distributions and ROC curves (Macmillan & Creelman, 1996).

Previous work has established that, if a measure is applied that does not conform to the underlying structure of the data, two conditions may appear to differ in performance (i.e. sensitivity) where no true difference exists, *provided that* they differ in response bias (Rotello et al., 2008; Schooler & Shiffrin, 2005). In the present case we were interested in conclusions regarding interactions between variables assessed via analysis of variance. The simulations presented here (and in the online supplementary material) have shown that erroneous conclusions can arise regarding interactions between variables in the absence of any difference in response bias. In the examples used here, provided both group and condition resulted in moderate to large main effects themselves, the probability of erroneously encountering a significant interaction effect is quite large when an inappropriate measure is applied. Further, genuine interaction effects are likely to be missed; with a specific tendency for d' to miss under-additive interactions and for P_r to miss over-additive interactions (see Figure 4). These simulations show how various factors, such as the overall sensitivity of observers, differences between groups and conditions, and variation in response bias, combine to influence error rates and power.

The problem of understanding the theoretical importance of interactions, such as these, were well outlined by Loftus (1978) who distinguished 'interpretable' and 'uninterpretable' interactions. The present simulations underline the problems associated with model dependent interactions: it is possible to make interactions appear and disappear simply by changing the measure of sensitivity chosen as the outcome variable. Despite the possibility that this will change the way in which researchers interpret interaction effects, a relatively recent survey of the field suggests that Loftus' warnings have gone largely unnoticed by experimental psychologists, who still ascribe fairly large theoretical weight to significant interactions (Wagenmakers et al., 2012).

Recommendations

As we have described so far, assessing the evidence for interaction effects is more difficult than it may first appear. In the field of recognition and detection, however, there are ways that researchers may gain greater confidence in the veracity of their interaction effects. Table 1 summarizes a number of steps researchers might take to avoid the problems highlighted in the simulations presented here. The remainder of the article covers these recommendations in more detail.

Choosing an appropriate measure. The importance of selecting a measure consistent with the structure of the data has been highlighted several times before (Macmillan & Creelman, 2005; Rotello et al., 2008; Swets, 1986a). But, as recently noted by Rotello et al. (2015), there is a tendency for outcome measures to be chosen on the basis of what is typically used in the field, without consideration of whether the measure gives an accurate representation of the underlying process. As outlined in the *Introduction*, the source of the discrepancy between measures is in their different predictions as to how false-alarm and hit rates should relate over varying response bias; that is, their predicted ROC curves (see Figure 2). The form of these ROC curves need not remain a mystery, however. Researchers should search the existing literature for empirical ROC curves constructed for tasks similar to their own (Swets, 1986a).

This is not without its pitfalls, and researchers should be wary of the controversy surrounding ROC curves derived from ratings procedures, in which, instead of a yes/no decision, participants rate their decision on a scale from *sure target* to *sure non-target*. With such a ratings procedure, it is not necessarily the case that a threshold model predicts linear ROCs, as a change in the mapping of detection states onto ratings can introduce curvature (see Bröder, Kellen, Schütz, & Rohrmeier, 2013; Malmberg, 2002). Thus, it has been argued that analysts should focus on the form of binary choice ROC curves, derived from the manipulation of base rates (i.e. expectation of the frequency of target trials), or payoffs (i.e. offering more or less reward for correct *target* responses) (see Bröder & Schütz, 2009).

It is beyond the scope of this article to provide a review of ROC data from various

disciplines or tasks, and several such reviews already exist (Dube & Rotello, 2012; Macmillan & Creelman, 2005; Swets, 1986a; Wixted, 2007; Yonelinas & Parks, 2007). However, we are able to give an example of choosing an appropriate measure from our own field of interest (and what happens when an inappropriate measure is applied). Isella, Molteni, Mapelli, and Ferrarese (2015) recently used a change detection paradigm to assess younger and older adults' short-term recognition memory for simple objects made up of color and shape. They were interested in the extent to which healthy aging affected the ability to detect changes to color-shape pairing beyond any age effect observed for detecting changes to individual features. In this case, a group by condition interaction may suggest that older adults have a specific deficit in short-term memory binding.

To address this question, the authors applied A', claiming that it was "the most classical and appropriate parameter for a change detection task" (p. 38); however, it is not clear how this conclusion was reached. Nevertheless, Isella et al. (2015) found a significant group by condition interaction with this measure. In the online supplement to their paper, however, they report an additional analysis of proportion correct in which the crucial interaction is not significant. It is important to note that Isella et al. (2015) employed an adjustment for multiple comparisons which led them to dismiss the interaction contrast between the lowest performing individual feature condition (shape only) and the condition requiring color-shape binding ($p = .021$) as non-significant. As we argue below, this was likely a good move, however, given that this group by condition interaction was the component of primary interest in this study, we can equally envisage a situation where the correction was not applied and the significant interaction interpreted in a different manner. Thus, we are using this as a case where different conclusions *could* have been reached based on the choice of measure, not as a specific critique of the conclusions of Isella et al. (2015).

It is our contention that A' is an inappropriate measure in this situation, and a more appropriate measure can be justified in the basis of the extant literature. Previous research has shown that base rate manipulations with the change detection task,

achieved by varying the proportion of target and non-target trials and informing participants of this, yield linear ROC curves when categorically distinct stimuli are used (e.g. Rouder et al., 2008). Thus, it can be argued that proportion correct, which can be justified as an index of sensitivity according to a high-threshold model (see above), is a more appropriate measure for the data of Isella et al. (2015). Indeed, other studies on the same topic have applied the two-high threshold measure, P_r, and have provided evidence against the group by condition interaction (see Rhodes, Parra, & Logie, 2016; Rhodes, Parra, Cowan, & Logie, 2017).[5]

Nevertheless, it is often difficult to find truly diagnostic ROC curves (see, for example, the debate in the recognition memory literature; Bröder & Schütz, 2009; Dube & Rotello, 2012). Researchers should therefore consider additional sources of information to help them decide between measures. For instance, recent interest has shifted from comparing models on the basis of ROC data towards tests of critical predictions made by models of the discrimination process. An example of this is the prediction of conditional independence made by the two-high threshold model (see Chen, Starns, & Rotello, 2015; Province & Rouder, 2012, for recent tests of this prediction). This is the expectation that given a certain internal state (e.g. detection of a previously studied stimulus) has been reached, responses (e.g. confidence ratings) should be invariant and not depend on experimental manipulation. Assessments of the predictions of signal detection and high-threshold accounts regarding distributions of reaction times are also informative in arriving on an appropriate measure for a given situation (see Donkin, Nosofsky, Gold, & Shiffrin, 2013; Dube, Starns, Rotello, &

[5]When considering change detection it is worth noting that there are a range of other measures that are used with this paradigm. These attempt to measure the number of items held in working memory (k) and different measures are appropriate for different versions of the change detection task (depending on the type of probe used; see Cowan, Blume, & Saults, 2013, for discussion). Each of these models is related to the two-high threshold model, in that they assume discrete states (items are either in memory or not), but predict different precise forms of ROC curves. Thus, while we believe it is correct to say that proportion correct is the "more appropriate measure" given the extant data, there are likely to be even more appropriate measures for the specific change detection task used by Isella et al. (2015). The general findings of the present simulations will apply to these change detection measures; when the implied ROC curves of a measure disagree with the underlying process, sensitivity (in this case k) will be confounded with bias and scaling will effect the interpretation of interaction effects. Rouder, Morey, Morey, and Cowan (2011) and Rhodes, Cowan, Hardman, and Logie (2018) discuss issues surrounding the change detection task and measurement of k.

Ratcliff, 2012; Province & Rouder, 2012).

In the process of searching for a measure it is highly likely that a measure outside of the set considered here will be appropriate for the paradigm under consideration. It is then useful for applied researchers to be aware of other models and measures that they may add to their toolkit. For example, even if the data conform to a Gaussian signal detection model, it may be that the evidence distributions have different variance, making the use of d' inappropriate (Swets, 1986a). Previous estimates of the ratio of non-target to target standard deviations (s) can be used to correct d' to be more appropriate to the underlying structure of the data. In recognition memory paradigms, for example, ROC estimates of s tend to be smaller than 1 at approximately 0.8 (see Ratcliff et al., 1992; Wixted, 2007). It is possible, given a previous estimate of s, to obtain an appropriate estimate of sensitivity according to a signal detection account. One such measure is d_a (Simpson & Fitter, 1973), which gives the separation of the target and non-target distributions in terms of the root mean square of the two standard deviations. The formula for calculating this measure can be found on page 400 of Rotello et al. (2008).

Further, there is low threshold theory (Luce, 1963), which relaxes the assumption of high thresholds and allows for observers to erroneously enter detection states (for example, entering the 'detect target' state when presented with a non-target). This model predicts ROCs made of two line segments and has recently been shown to provide a reasonable fit to data from recognition memory tasks (Kellen, Erdfelder, Malmberg, Dubé, & Criss, 2016; Starns & Ma, 2018). Thus, researchers should be aware of this alternative model and may consider applying it in their work, although it should be noted that this model does not produce a single point measure of sensitivity.

The present simulations (see in particular the unequal variance simulations in the supplementary material) demonstrate that the fact that so often empirical ROC curves are asymmetrical (see Swets, 1986a; Wixted, 2007, for reviews) should be of great concern to researchers. Findings may appear replicable and robust when, in fact, they arise due to a distorting effect of an inappropriate measure (see Rotello et al., 2015, for

a detailed discussion of this important point). Knowledge of these alternative measures and models, often ignored in applied research, will better place researchers to avoid erroneous conclusions regarding interactions.

Avoid ceiling level performance. Avoiding perfect, or near perfect, performance is usually a concern for researchers. The present simulations reinforce this concern by clearly showing that high levels of performance can cause issues even when an appropriate measure is chosen. The simulations with variable response bias also reveal a problem when *either* h or f are at ceiling (or floor), as this results in the other estimated rate coming to dominate inference, which in turn can drive erroneous conclusions even when the correct measure is applied. Thus, while overall performance (e.g. proportion correct) may not be at ceiling, individual rates of 0 and 1 can still cause problems. The supplementary material contains further discussion of this point.

Consider the scale of measures. When there is no variation in response bias, the primary source of disagreement between measures in an analysis of variance stems from the way in which they are scaled. Equal increments along the negative diagonal imply increasingly smaller differences in d' (Figure 2A), whereas the way in which our two-high threshold simulations were set up reflects the general assumptions implied in analyzing estimates of P_r with ANOVA. That is, that P_r is on an interval scale so that mean differences are equally as meaningful at moderate and high levels of sensitivity. It seems unlikely that researchers would ascribe as great a meaning to a difference in P_r of 0.55 vs 0.6 as they would to the difference between 0.9 and 0.95. These scaling issues are often raised when discussing the analysis of categorical data, so analyses assuming a high threshold process may benefit from insights from this area.

As noted in the Introduction, proportion correct can be justified as a measure of sensitivity given its linear relation to P_r. To reflect the assumption that differences at higher levels of accuracy are more meaningful than differences at lower levels, researchers may analyze a transformation of proportion correct. Logit models offer one such method of analysis, which is preferable to the analysis of aggregated and transformed proportions with ANOVA (see Dixon, 2008). Here, the researcher would

estimate the log odds of a correct response (across both target and non-target trials) using the raw binary (correct and incorrect) responses. This analysis would still imply linear ROCs but would produce a similar scaling of effects to that produced with the signal detection theory measure, d' (the z, or probit, transformation used in calculating d' is very similar to the log odds, or logit, transformation). Consequently, one avoids the problem of treating P_r as if it were interval in scale (as is implied in the use of ANOVA), and scales effects similarly to a signal detection theory analysis.

To show this, we fit a mixed-effects logit model to each simulated data set from the first series of simulations that used an equal variance signal detection generative model. P-values for group by condition interactions were obtained using the `lme4` package in `R` (Bates, Maechler, Bolker, & Walker, 2014; R Core Team, 2015). As expected, relative to ANOVA on P_r, the logit model fared much better with type I error rates, on average, approximately 13% lower. Overall, there was no clear difference between the logit model and the correct measure d' (on average, type I error rates differed by 0.5%). Thus, changing the way in which one thinks about the scale of effects on P_r, which seems reasonable given that this measure is bounded by 0 and 1, reduces the disagreement between models in the absence of variation in response bias. Disagreement still arises when bias is varied between groups and conditions, as exemplified by the simulations in which there were no main effects on sensitivity (see Simulations 7 and 12 in supplementary material) (see also Rotello et al., 2008; Schooler & Shiffrin, 2005).

Present a range of measures. It is likely that researchers will be faced with a situation where there is no good reason to pick a particular measure over others available. For example, a researcher may design a task for which no good ROC evidence exists and they are unable, due to lack of time or resources, to probe this themselves. In this case what can they do? Many of the options open to researchers in this circumstance have already been outlined by Wagenmakers et al. (2012, pp. 157). Perhaps the simplest is that investigators may choose to report a range of different measures, such as d', P_r, logit(p), as well as signal detection measures that do not assume equal variance (see Rotello et al., 2008), and examine whether the interaction

survives (or whether additivity remains). If an interaction effect appears with multiple measures, one can gain confidence that their conclusion is not dependent on a single scale of measurement (although doubtlessly it will vary in magnitude, in which case a range of effect sizes may be given).

Alternative ways of assessing of sensitivity. The frequent use of the measure A' suggests that—possibly due to an (implicit or explicit) understanding of the issues above—researchers are not comfortable with the underlying assumptions made by other measures of sensitivity, and desire an assessment of performance that is not tied to a particular conception of the discrimination process. However, as outlined in the *Introduction*, A' does not live up to its 'non-parametric' name (Macmillan & Creelman, 1996) and there have been repeated calls for use of the measure to be reconsidered or abandoned (Pastore et al., 2003; Rotello et al., 2008). Our simulations add to the long list of reasons to avoid this measure. In type I error simulations, A' performed particularly badly when data conformed to the expectation of a signal detection model and did not fare well at detecting over-additive interactions for both generative models considered here. However, the desire to have a measure of performance which is tied to as few assumptions as possible remains.

A truly non-parametric assessment of discrimination sensitivity can be achieved by constructing an empirical ROC curve, either by collecting rating responses or by varying response bias or criterion (see above). From the empirical ROC, an estimate of area under the curve can be obtained providing a model-free estimate of an observer's sensitivity (Bamber, 1975; Green, 1964; Green & Moses, 1966). This can be done non-parametrically by summing the trapezoids created by each successive (f, h) pair (see Macmillan & Creelman, 2005, pp. 64; Pollack & Hsieh, 1969). This estimate of area is known to be biased downwards from the 'true' area, but a better approximation can be acheived by obtaining more points on the ROC or by using the method of Donaldson and Good (1996) to try and correct this bias. One may also assess the area under the curve parametrically assuming a signal detection model (see Stanislaw & Todorov, 1999, for guidence).

Alternatively, if the target/ non-target decision is not integral to the study design, researchers may consider a two-alternative forced choice (2AFC) task in which observers are presented with two items (one a target and one a non-target) and must select one of them as the target. In many applications it may be safely assumed that bias plays little role in responding and does not vary greatly between conditions or groups. Thus the 2AFC can reduce concerns regarding the separability of sensitivity and bias, but measurement decisions still need to be made (see Macmillan & Creelman, 2005, Chapter 7 for detailed discussion).

A summary of recommendations, along with their strengths and drawbacks, can be found in Table 1.

Concluding Remarks

We have shown that measures of sensitivity, as typically analyzed via standard methods like analysis of variance, can produce conflicting evidence for interactions even in the absence of differences in response bias. In considering a wide array of underlying parameter settings, the present simulations show some situations where we may expect this tendency to be especially pronounced. Without access to raw hit and false-alarm rates, it is difficult to know the true extent to which issues like these have biased the literature. However, the potential for conclusions to vary given a change of sensitivity measure is certainly shown by the example of Isella et al. (2015) given above (also see Rotello et al., 2015, for other case studies). While the realization that interactions may be created and taken away via a change of measurement scale is not new (Loftus, 1978), we hope that the simulations presented here underline the scale of the problem and bring it to the wider attention of psychologists (Wagenmakers et al., 2012). We also hope that the recommendations provided will help researchers state their conclusions regarding interactions in detection or recognition experiments more clearly, even if the conclusions themselves are less clear.

Table 1
A summary of recommendations for assessing interaction in detection and recognition experiments.

Action	Necessary Steps	Strengths	Drawbacks
(1) Determine an appropriate measure on the basis of empirical ROC curves and other model comparison evidence.	Search literature (published and unpublished) for previously constructed ROCs (see references in text) or perform bias manipulations to trace out curves.	This is the best approach to choosing an appropriate measure. As the simulations show, this will reduce type I error rate and give appropriate power	ROCs generated from confidence ratings are contentious. Creating your own ROC requires a large number of trials per condition and can be time consuming.
(2) If applying a two-high threshold model, consider the scaling of effects. A standard linear model is likely unsuited to assessing P_r.	Apply generalized linear mixed-effects modeling approaches to model proportion correct.	This eliminates much of the disagreement between measures when there are no differences in bias.	This approach does not account for the distortion caused by differences in response bias.
(3) If obtaining ROC data is not possible or impractical, calculate a number of different measures.	Assess the interaction effect with numerous measures reflecting different underlying processes (including equal and unequal variance signal detection).	The conclusion that the interaction is present (or absent) is strengthened when all measures yield similar results.	Estimates of effect size are likely to vary substantially. An inferential cost is incurred by not identifying an appropriate measure.
(4) Avoid using A'. If a nonparametric assessment of sensitivity is desired, consider using a rating task.	Require participants to produce a confidence rating following every binary response	Area under the curve gives a model free estimate of sensitivity	Adds a moderate time to experimental session and extra demand on the observer.

References

Aaronson, D., & Watts, B. (1987). Extensions of Grier's computational formulas for A' and B" to below-chance performance. *Psychological Bulletin*, *102*(3), 439–442.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, *12*(4), 387–415.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4 [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=lme4` (R package version 1.1-7)

Bröder, A., Kellen, D., Schütz, J., & Rohrmeier, C. (2013). Validating a two-high-threshold measurement model for confidence rating data in recognition. *Memory*, *21*(8), 916–944.

Bröder, A., & Schütz, J. (2009). Recognition ROCs are curvilinear—or are they? on premature arguments against the two-high-threshold model of recognition. , *35*(3), 587–606.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178.

Chen, T., Starns, J. J., & Rotello, C. M. (2015). A violation of the conditional independence assumption in the two-high-threshold model of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(4), 1215–1222.

Cowan, N., Blume, C. L., & Saults, J. S. (2013). Attention to attributes and objects in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 731–747.

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language*, *59*(4), 447–456.

Donaldson, W., & Good, C. (1996). $A'r$: An estimate of area under isosensitivity curves. *Behavior Research Methods, Instruments, & Computers*, *28*(4), 590–597.

Donkin, C., Brown, S., & Heathcote, A. (2011). Drawing conclusions from choice

response time models: A tutorial using the linear ballistic accumulator. *Journal of Mathematical Psychology*, *55*(2), 140–151.

Donkin, C., Nosofsky, R. M., Gold, J. M., & Shiffrin, R. M. (2013). Discrete-slots models of visual working-memory response times. *Psychological Review*, *120*(4), 873–902.

Dube, C., & Rotello, C. M. (2012). Binary ROCs in perception and recognition memory are curved. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(1), 130–151.

Dube, C., Starns, J. J., Rotello, C. M., & Ratcliff, R. (2012). Beyond ROC curvature: Strength effects and response time data support continuous-evidence models of recognition memory. *Journal of Memory and language*, *67*(3), 389–406.

Green, D. M. (1964). General prediction relating yes-no and forced-choice results [abstract]. *The Journal of the Acoustical Society of America*, *36*(5), 1042–1042.

Green, D. M., & Moses, F. L. (1966). On the equivalence of two recognition measures of short-term memory. *Psychological Bulletin*, *66*(3), 228–234.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Grier, J. B. (1971). Nonparametric indexes for sensitivity and bias: computing formulas. *Psychological Bulletin*, *75*(6), 424–429.

Isella, V., Molteni, F., Mapelli, C., & Ferrarese, C. (2015). Short term memory for single surface features and bindings in ageing: A replication study. *Brain and Cognition*, *96*, 38–42.

Kellen, D., Erdfelder, E., Malmberg, K. J., Dubé, C., & Criss, A. H. (2016). The ignored alternative: An application of Luce's low-threshold model to recognition memory. *Journal of Mathematical Psychology*, *75*, 86–95.

Loftus, G. R. (1978). On interpretation of interactions. *Memory & Cognition*, *6*(3), 312–319.

Luce, R. D. (1963). A threshold theory for simple detection experiments. *Psychological Review*, *70*(1), 61–79.

Macmillan, N. A., & Creelman, C. D. (1990). Response bias: Characteristics of detection theory, threshold theory, and "nonparametric" indexes. *Psychological Bulletin*, *107*(3), 401–413.

Macmillan, N. A., & Creelman, C. D. (1996). Triangles in ROC space: History and theory of "nonparametric" measures of sensitivity and response bias. *Psychonomic Bulletin & Review*, *3*(2), 164–170.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Malmberg, K. J. (2002). On the form of ROCs constructed from confidence ratings. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(2), 380–387.

Pastore, R. E., Crawley, E. J., Berens, M. S., & Skelly, M. A. (2003). "Nonparametric" A' and other modern misconceptions about signal detection theory. *Psychonomic Bulletin & Review*, *10*(3), 556–569.

Pollack, I., & Hsieh, R. (1969). Sampling variability of the area under the ROC-curve and of d'_e. *Psychological Bulletin*, *71*(3), 161.

Pollack, I., & Norman, D. A. (1964). A non-parametric analysis of recognition experiments. *Psychonomic Science*, *1*(1-12), 125–126.

Province, J. M., & Rouder, J. N. (2012). Evidence for discrete-state processing in recognition memory. *Proceedings of the National Academy of Sciences*, *109*(36), 14357–14362.

R Core Team. (2015). R: A Language and Environment for Statistical Computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.

Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review*, *99*(3), 518–535.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: current issues and history. *Trends in Cognitive Sciences*, *20*(4), 260–281.

Rhodes, S., Cowan, N., Hardman, K. O., & Logie, R. H. (2018). Informed guessing in change detection. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, online ahead of print.

Rhodes, S., Parra, M. A., Cowan, N., & Logie, R. H. (2017). Healthy aging and visual working memory: The effect of mixing feature and conjunction changes. *Psychology & Aging*, *32*(4), 354–366. doi: 10.1037/pag0000152

Rhodes, S., Parra, M. A., & Logie, R. H. (2016). Ageing and feature binding in visual working memory: The role of presentation time. *The Quarterly Journal of Experimental Psychology*, *69*(4), 654–668. doi: 10.1080/17470218.2015.1038571

Rotello, C. M., Heit, E., & Dubé, C. (2015). When more data steer us wrong: Replications with the wrong dependent measure perpetuate erroneous conclusions. *Psychonomic Bulletin & Review*, *22*, 944–954.

Rotello, C. M., Masson, M. E., & Verde, M. F. (2008). Type I error rates and power analyses for single-point sensitivity measures. *Perception & Psychophysics*, *70*(2), 389–401.

Rouder, J. N., Morey, R. D., Cowan, N., Zwilling, C. E., Morey, C. C., & Pratte, M. S. (2008). An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences*, *105*(16), 5975–5979.

Rouder, J. N., Morey, R. D., Morey, C. C., & Cowan, N. (2011). How to measure working memory capacity in the change detection paradigm. *Psychonomic Bulletin and Review*, *18*, 324–330.

Schooler, L. J., & Shiffrin, R. M. (2005). Efficiently measuring recognition performance with sparse data. *Behavior Research Methods*, *37*(1), 3–10.

Simpson, A. J., & Fitter, M. J. (1973). What is the best index of detectability? *Psychological Bulletin*, *80*(6), 481–488.

Smith, W. D. (1995). Clarification of sensitivity measure A'. *Journal of Mathematical Psychology*, *39*(1), 82–89.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: applications to dementia and amnesia. *Journal of Experimental Psychology:*

General, *117*(1), 34–50.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, and Computers*, *31*(1), 137–149.

Starns, J. J., & Ma, Q. (2018). Guessing versus misremembering in recognition: A comparison of continuous, two-high-threshold, and low-threshold models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(4), 527–539.

Swets, J. A. (1986a). Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance. *Psychological Bulletin*, *99*(2), 181–198.

Swets, J. A. (1986b). Indices of discrimination or diagnostic accuracy: their ROCs and implied models. *Psychological Bulletin*, *99*(1), 100–117.

Tanner Jr, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, *61*(6), 401–409.

van Ravenzwaaij, D., Donkin, C., & Vandekerckhove, J. (2017). The EZ diffusion model provides a powerful test of simple empirical effects. *Psychonomic Bulletin & Review*, *24*(2), 547–556.

Wagenmakers, E.-J., Krypotos, A.-M., Criss, A. H., & Iverson, G. (2012). On the interpretation of removable interactions: A survey of the field 33 years after Loftus. *Memory & Cognition*, *40*(2), 145–160.

Wagenmakers, E.-J., Van Der Maas, H. L., & Grasman, R. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, *14*(1), 3–22.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114*(1), 152–176.

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: a review. *Psychological Bulletin*, *133*(5), 800–832.

Zhang, J., & Mueller, S. T. (2005). A note on ROC analysis and non-parametric estimate of sensitivity. *Psychometrika*, *70*(1), 203–212.