The relative psychometric function: a general analysis framework for relating psychological processes

Authors:

Brian Maniscalco[1,2], Olenka Graham Castaneda[1,2], Brian Odegaard[3], Jorge Morales[4,5], Sivananda Rajananda[2], Rachel N. Denison[6], & Megan A. K. Peters[1,2,7,8,9,10]

Affiliations:

1 Department of Cognitive Sciences, University of California Irvine
2 Department of Bioengineering, University of California Riverside
3 Department of Psychology, University of Florida
4 Department of Psychology, Northeastern University
5 Department of Philosophy and Religion, Northeastern University
6 Department of Psychological and Brain Sciences, Boston University
7 Department of Logic & Philosophy of Science, University of California Irvine
8 Center for the Theoretical Behavioral Sciences, University of California Irvine
9 Center for the Neurobiology of Learning and Memory, University of California Irvine
10 Program in Brain, Mind, & Consciousness, Canadian Institute for Advanced Research

Correspondence should be directed to:

Brian Maniscalco (bmaniscalco@gmail.com)
Megan A. K. Peters (megan.peters@uci.edu)

Abstract

Psychophysics seeks to quantitatively characterize relationships between objective properties of the world and subjective properties of perception. However, traditional approaches investigate psychophysical dependencies of perception on stimulus properties on a case by case basis rather than seeking to identify quantitative relationships *among* these psychological processes themselves. This latter goal is particularly important when the processes in question likely depend on each other in some way, such as is the case for subjective experience and task performance: typically, stronger physical stimuli lead to better performance *and* stronger subjective experiences of clarity, vividness, or confidence. But is the relationship between performance and subjective experience fixed, or can it vary, e.g. by task or attentional demands? Such questions are key for better understanding psychological processes in general, and subjective experience in particular. Here, we develop and showcase a new psychophysical method designed to answer such questions: *relative psychometric function* (RPF) analysis, which characterizes the nonlinear psychometric relationships between psychological processes and how these relationships change with experimental manipulations or individual differences. We demonstrate the advantages of RPF analysis using a sample dataset in which human subjects discriminated random dot kinematogram stimuli which varied in dot motion coherence and overall dot density (dots per visual degree), and rated confidence. RPF analysis revealed systematic changes in the relationship between performance and two subjective measures (confidence and metacognitive sensitivity) due to dot density and task design choices. While these empirical results are intriguing in their own right, they also show how RPF analysis can reveal changes in quantitative relationships between any two psychological processes: performance, vividness, clarity, reaction time, confidence, and more. To encourage the scientific community to use RPF analysis on their data, we also present our open-source RPF toolbox.

Keywords: psychophysics; psychometric functions; relative psychometric function; subjective experience; quantitative psychology

Introduction

Arguably, the field of quantitative psychology began in the 1860s with Fechner's *Elemente der Psychophysik – Elements of Psychophysics* (Fechner, 1860; Fechner et al., 1966). Fechner's work set the foundation for what is now over 150 years of concerted effort to map objective properties of the world to properties of the brain and mind. Weber, Stevens, and others followed, seeking to establish the functional forms of these relationships: that an observer's "just noticeable difference" in discriminating two stimuli depends on their absolute magnitude (Weber's law), and that the perceived magnitude of a stimulus (brightness, loudness, painfulness) exhibits an exponential relationship to the objective stimulus magnitude (Stevens' power law). These "introductory psychology course" concepts are foundational pillars in the modern study of psychology.

The success of this framework underscores our deep motivation to build models of our minds, but standard psychophysical approaches represent one-to-one mappings between the physical and mental. Ultimately, we wish to understand not only how psychological processes relate to world properties, but also how psychological variables relate to *other psychological processes.* For example, increasing stimulus strength typically leads to faster, more accurate decisions, and increased sense of confidence in those decisions. Likewise, the subjective sense of clarity may also systematically vary with stimulus properties. But what is the relationship among all these psychological variables, and is it fixed across different attentional states, tasks, or individuals? While the relationship linking stimulus magnitude, discriminability, and absolute magnitude estimation has recently been described (Zhou et al., 2024), what about the relationships linking all these other psychological properties to each other?

Characterizing quantitative relationships between psychological variables is also especially important when those relationships themselves may change depending on properties of the world or other psychological processes. Perhaps nowhere is this more evident than in psychophysical studies of subjective experience, where there are clear, empirically observed relationships among stimulus intensity, task performance, and second-order judgments such as confidence (judgment of whether a given discrimination decision is likely to be correct) or subjective visibility or vividness (e.g., judgment of the clarity with which you saw a stimulus, regardless of its objective properties). In most instances, these subjective aspects covary with objective performance (Baranski and Petrusic, 1994): a higher probability of correctly identifying a stimulus is typically accompanied by higher confidence, higher vividness ratings, and a higher probability of reporting having seen the stimulus at all. This means that any neural or psychophysical measures of subjective experience are easily confounded by processes driving objective performance.

A standard approach to disentangling the neural correlates of subjective experience from those underlying objective performance has been to control for these 'performance confounds' through either experimental or analytic approaches (Lau, 2008; Lau and Passingham, 2006; Morales et al., 2022; Peters et al., 2017b). One popular approach is to create multiple experimental conditions in which performance (e.g., percent correct responses or the signal

detection theoretic metric *d'*) is held constant (e.g., through subject-specific staircasing) but subjective reports (confidence, vividness, visibility, clarity) vary. However, this approach is insufficient for several reasons. First, if performance is held constant across conditions, discovering a difference in subjective reports despite no difference in objective performance might rely on a statistical null effect, i.e. that the null hypothesis of equivalent performance could not be rejected given the data available. This situation might arise simply from situations in which a chosen manipulation had a smaller or noisier effect on performance than it did on subjective reports, rather than no effect at all. Second, the effect of an experimental manipulation on a subjective measure at a given matched performance level may strongly depend on the absolute level of performance (see also (Morales et al., 2022) for further discussion). Such condition-driven differences in confidence at matched performance have been observed in many different paradigms (Koizumi et al., 2015; Lau and Passingham, 2006; Maniscalco et al., 2016; Odegaard et al., 2018a, 2018b; Rahnev et al., 2011; Rouault et al., 2018; Samaha et al., 2016; Stolyarova et al., 2019), but the effect size (and sometimes even direction!) varies drastically across stimulus or task manipulations as well as the (matched) performance level itself.

For these reasons, performance matching is not enough. Neither is the qualitative characterization that performance and subjective reports both increase with stimulus strength. Instead, we wish to understand the precise quantitative relationship among these variables, including assessing the stability or generalizability of those relationships across experimental manipulations, changing brain states, or individuals.

To date, however, no framework exists for precisely characterizing such nonlinear relationships among psychological variables and how they may change in interesting ways. An impediment to this enterprise has been that these relationships are, by definition, nonlinear linkages between variables measured with error. Fitting any function linking two such variables constitutes a nonlinear “errors in variables” problem, for which there is no known closed form solution (see e.g., (Hausman et al., 1995; Huang et al., 2023; Li, 2002; Wolter and Fuller, 1982). Moreover, even if this problem were solved, the functional form linking two or more psychological variables is unlikely to be known a priori, requiring us to fall back on nonparametric methods (e.g. rank-based correlations) designed merely to reveal the *presence* and *strength* of a potential relationship, not its shape.

In short, we need a framework to (a) quantitatively characterize the nonlinear relationships among psychological variables measured with error; and (b) quantitatively characterize how much – and in what way – those relationships change with experimental manipulations, neural factors, or individual differences.

Here, we introduce an analytic framework to address these problems in the study of the neural and computational machinery underlying multiple psychological variables at once. The approach, which we term *relative psychometric function* (RPF) analysis, aims to systematically characterize how some aspects of perception, experience, or stimulus processing behave relative to the behavior of other aspects in response to stimulus or task manipulations, individual

differences, and so on. The framework is sufficiently general to be applied to investigation of the relationship between any psychological or neural processes P_1 and P_2 which can be expressed as psychometric functions of a common continuous variable such as stimulus intensity. And, for the study of perceptual metacognition specifically, the RPF offers a method for quantifying, parameterizing, and thus understanding the entire *relative psychometric function* linking various objective and subjective aspects of perception – including confidence, visibility, vividness, clarity, and any others that might be deemed relevant – across the whole range of performance that might be elicited in a given task. This approach thus provides a tool for precisely measuring how different subjective experiences might arise from equivalent intervals of objective processing capacity, sidestepping earlier challenges described above. The RPF method as applied to perceptual metacognition also answers recent calls for a 'metacognitive psychophysics' (Fleming, 2023), and builds upon the "metacognition as a step towards explaining phenomenology" (M-STEP) approach introduced by Peters (2022), which called for research to seek canonical metacognitive computations as a strategy for revealing how subjective experience in general may be generated.

In what follows we introduce this relative psychometric function, derive its parameterization, explore its behavior, develop and validate interpretable summary statistics, and discuss its interpretation using a sample dataset in which performance and confidence were independently manipulated across a large range of stimulus strengths.

We believe this framework will prove a highly flexible and powerful analysis tool in psychology and neuroscience for studying relationships between various psychological and neural processes. To facilitate this goal, all the methods, data, and analyses presented here are also used to introduce the RPF toolbox (https://github.com/CNClaboratory/RPF) – an open-source resource for the community to apply RPF analyses to any suitable dataset. Thus, we make reference to this toolbox throughout, and include additional details about implementation on the case study dataset presented here in the **Supplemental Material**.

Methods, Results, & Discussion

Deriving and interpreting the relative psychometric function (RPF)

Foundations of RPF analysis

The general form of the RPF

We define the *relative psychometric function*, or RPF for short, as the function describing the relationship between any two conventional psychometric functions that are expressed in terms of a common independent variable. More formally, suppose we have two psychometric functions

$$P_1 = F_1\left(x; \theta_1\right)$$
$$P_2 = F_2\left(x; \theta_2\right) \quad (1)$$

where x is stimulus strength and P_1 and P_2 are different measures of performance, such as p(correct) or average confidence[1,2]. Provided that F_1 is invertible such that $x = F_1^{-1}(P_1)$, we may express P_2 in terms of P_1 by writing $P_2 = F_2(F_1^{-1}(P_1))$. We may then define the relative psychometric function as

$$R = F_2 \circ F_1^{-1}$$
$$P_2 = F_2\left(F_1^{-1}\left(P_1\right)\right) \quad (2)$$
$$= R\left(P_1; \theta_1, \theta_2\right)$$

and write $P_2 = R(P_1)$ for short. Thus, R uses the known psychometric functions F_1 and F_2 of a common independent variable x in order to express P_2 as a function of P_1 (**Figure 1**).

R can be seen as the result of a coordinate transformation of F_2 in which the x input is replaced with a P_1 input derived from the mapping $x = F_1^{-1}(P_1)$. Thus, the plot of $P_2 = R(P_1)$ resembles a warped plot of $P_2 = F_2(x)$ in which the y-axis values are identical but the x-axis is warped according to the (likely nonlinear) transformation specified by $P_1 = F_1(x)$ (cf. the three panels of **Figure 1**, and **Figures S1** and **S2** in the **Supplemental Material**).

Importantly, deriving $P_2 = R(P_1)$ via the relationship of P_1 and P_2 to a common independent variable x, where x is known exactly rather than measured with error, bypasses difficulties that would arise from attempting to fit a function directly to the (P_1, P_2) data. Since both P_1 and P_2 are variables measured with error and likely have a nonlinear relationship, attempting to fit $P_2 = R(P_1)$ directly requires a nonlinear errors-in-variables model. However, there is no known solution for fitting such models directly, and existing approaches require incorporation of additional data and application of complex analysis methods tailored to specific cases (see e.g. (Hausman et al., 1995; Huang et al., 2023; Li, 2002; Wolter and Fuller, 1982).

In the case of analyzing the relationships between psychometric functions in experimental psychology research, the (P_1, P_2) data we might wish to relate are themselves already derived from systematic manipulation of the common independent variable x, and thus the information needed to estimate the relationships of P_1 and P_2 to x comes “for free” in the collection of the (P_1, P_2) data. Thus, the approach described in this work is a natural choice for conducting RPF analysis that easily bypasses thorny analysis issues with readily available data.

[1] Note that P is intended as shorthand for “performance” and does not necessarily connote a probability.
[2] Here we use F rather than the conventional ψ to denote psychometric functions for consistency with the RPF toolbox notation, in which it is more convenient to use F.

The metaperceptual RPF

As discussed in the introduction, one particular application of interest is the case where P_1 and P_2 correspond to objective and subjective measures of perception, respectively. Here, we conceive of objective measures of perception as pertaining to judgments about objective states of the world (e.g. detecting stimulus presence or discriminating stimulus features), and subjective measures as pertaining to judgments about one's own perceptual processing (e.g. assessing confidence in an objective judgment or reporting on the qualities of one's perceptual experience). This characterization of "objective" and "subjective" categories can be seen as a generalization of the classical distinction between type 1 and type 2 perceptual tasks, in which the type 1 task is to classify a stimulus event and the type 2 task is to classify one's type 1 judgment as correct or incorrect (Clarke et al., 1959; Galvin et al., 2003; Maniscalco et al., 2024).

Taking inspiration from the term "psychophysics," we call this special class of RPFs *metaperceptual RPFs* or *metaperceptual functions*. Just as the roots of the word "psychophysical" connote "relationship of perception (psycho-) to stimulus (physical)," so the roots of the word "metaperceptual" connote "relationship of judgments *about* perception (meta-) to perception (perceptual)." We may also use the term *type 2 psychometric function* to refer to more restricted cases where the RPF relates type 2 judgments about type 1 accuracy (typically confidence ratings) to type 1 accuracy itself (e.g. as in p(correct)).

Objective measures of perception include accuracy measures such as p(correct) and the signal detection theory (SDT) measure of sensitivity *d'*, and response bias measures such as p(response) and the SDT measure of criterion *c*. Subjective measures include ratings of confidence and reports of experiential qualities such as visibility, clarity, intensity, etc. Subjective measures may also characterize the relationship between subjective and objective judgments, e.g. by measuring how well confidence ratings track accuracy as in the SDT measure of metacognitive sensitivity meta-*d'* (Fleming, 2017; Maniscalco and Lau, 2014, 2012)).

Considerations for fitting the component psychometric functions of the RPF

Psychometric functions can be fitted to probability measures such as p(correct) and p(high confidence) with standard maximum likelihood estimation (MLE) procedures (Kingdom and Prins, 2016). However, MLE fitting of psychometric functions to non-probabilistic measures requires a different approach. Least square fits maximize likelihood when errors in the fit can be assumed to be normally distributed with constant variance (Burnham and Anderson, 2002), but this assumption may not always hold (e.g. as for *d'*; see (Miller, 1996)).

In the **Supplemental Material** we derive approaches to achieving MLE psychometric function fits to several variables of central interest for metaperceptual functions: *d'*, meta-*d'*, and mean rating (e.g. for confidence or visibility ratings). These approaches work by relating the variable in question to probabilities for single-trial outcomes, and thus only require the standard MLE assumption that outcome probabilities are independent across trials. For cases where specifying or fitting analytical psychometric functions is problematic, we also develop

nonparametric RPF analysis methods (see below and **Supplemental Material** for further discussion), and demonstrate that MLE and nonparametric approaches are comparable in their ability to retrieve certain characteristics of the true RPF (see **Supplemental Material**).

It is possible to take a radically modular approach to constructing the RPF from its component psychometric functions, in the sense that because the F_1 and F_2 fits can be treated independently, they can be applied to any variable and approached with any fitting method prior to being combined in an RPF. Thus e.g. if F_1 is fitted via MLE, this does not constrain the possibilities for fitting F_2 via MLE or least squares or nonparametric methods. In all cases, the approach described in Eq. 2 is sufficiently general to conduct RPF analysis, with the proviso that F_1 must be invertible. However, even if F_1 is not invertible, nonparametric analysis of the RPF may still be conducted, as discussed further below.

All of these approaches to RPF analysis – MLE fitting for probabilistic variables and certain non-probabilistic variables, least square fitting, and nonparametric analysis – can be implemented in the RPF toolbox (https://github.com/CNClaboratory/RPF).

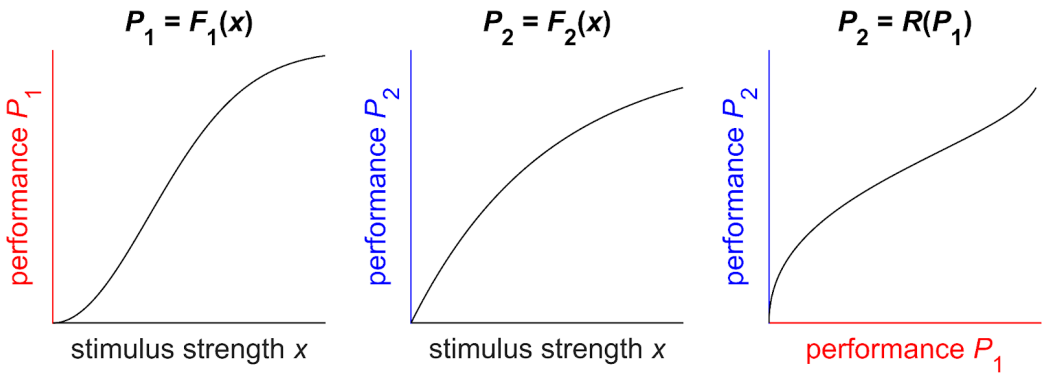

Figure 1. The relative psychometric function (RPF). (Left, middle) Conventional psychometric functions characterize the curve relating performance on a given task to stimulus strength. (Right) The *relative psychometric function* characterizes the curve relating performance on one task (P_2) to performance on another (P_1), given knowledge of how both relate to a common stimulus feature x. This function thus reveals how different measures of psychological processes relate to each other over a wide range of performance levels. In situations where RPFs differ across experimental conditions, analysis of the RPF can be used to help tease apart the behavior and underlying mechanisms of P_1 and P_2. In the special case where P_1 and P_2 correspond to objective and subjective aspects of perception (see main text), the RPF is a *metaperceptual function* which can be used to isolate subjective aspects of perception from potentially confounding aspects of objective task performance.

Probing RPF behavior: a case study using the Weibull RPF

How should we measure, summarize, and analyze the RPF? Can we summarize its behavior neatly with a small number of parameter values, similar to how conventional psychometric functions are typically analyzed in terms of location and slope parameters? Since the RPF depends on the mathematical forms of the two psychometric functions F_1 and F_2 from which it is composed, the answer to these questions requires specifying the equations for those functions.

Here, as a representative example, we consider the behavior of the RPF when P_1 and P_2 are probabilities (e.g. p(correct) and p(high confidence)) fitted by Weibull functions F_1 and F_2 (Kingdom and Prins, 2016).

The Weibull function for F_1 and F_2 takes on the form

$$P_n = F_n(x) = \gamma_n + \left(1 - \lambda_n - \gamma_n\right)\left[1 - e^{-(x/\alpha_n)^{\beta_n}}\right] \quad (3)$$

In this equation,

- n denotes the psychometric function to which all terms pertain, with n = 1 and 2 corresponding to F_1 and F_2 respectively
- P_n is performance (here, outcome probability)
- x is stimulus strength
- γ_n is the chance level of responding for P_n
- λ_n is the lapse rate, such that asymptotic performance for P_n is $1 - \lambda_n$
- α_n is the location parameter for $F_n(x)$
- β_n is the slope parameter for $F_n(x)$

Solving Eq. 3 for x gives

$$x = F_1^{-1}(P_1) = \alpha_1\left(\ln\left(\frac{1-\lambda_1-\gamma_1}{1-\lambda_1-P_1}\right)\right)^{\frac{1}{\beta_1}} \quad (4)$$

Substituting Eq. 4 into the general equation for R in Eq. 2 gives

$$P_2 = R_W(P_1) = \gamma_2 + \left(1 - \lambda_2 - \gamma_2\right)\left[1 - e^{-\left(\left(\frac{\alpha_2}{\alpha_1}\right)^{-\beta_2}\left(\ln\left(\frac{1-\lambda_1-\gamma_1}{1-\lambda_1-P_1}\right)\right)^{\frac{\beta_2}{\beta_1}}\right)}\right] \quad (5)$$

We name Eq. 5 the *Weibull RPF* (abbreviated R_W) as this is the mathematical form of the RPF in the case where both F_1 and F_2 are Weibulls. We can decompose the Weibull RPF into the following components:

- The F_2 guess rate γ_2 and lapse rate λ_2, which determine the minimum, chance level of performance and maximum, asymptotic level of performance for the R_W just as they do for F_2.
- The *performance ratio* $\frac{1-\lambda_1-\gamma_1}{1-\lambda_1-P_1}$, which characterizes performance P_1 relative to its possible range of values in $\left[\gamma_1, 1-\lambda_1\right]$. When P_1 is at the chance value of γ_1, the performance ratio = 1 and R_W is at the chance level of performance for P_2, i.e. γ_2. As P_1 approaches the asymptotic value of $1-\lambda_1$, the performance ratio approaches infinity and R_W approaches the asymptotic level of performance for P_2, i.e. $1-\lambda_2$.
- The *relative location* $\alpha_R = \frac{\alpha_2}{\alpha_1}$.
- The *relative slope* $\beta_R = \frac{\beta_2}{\beta_1}$.
- The F_2 slope β_2.

We explore how R_W depends on α_R, β_R, and β_2 in **Figure 2**. Without loss of generality, we set the scaling parameters $\gamma_1 = 0.5$, $\gamma_2 = 0$, and $\lambda_1 = \lambda_2 = 0$.

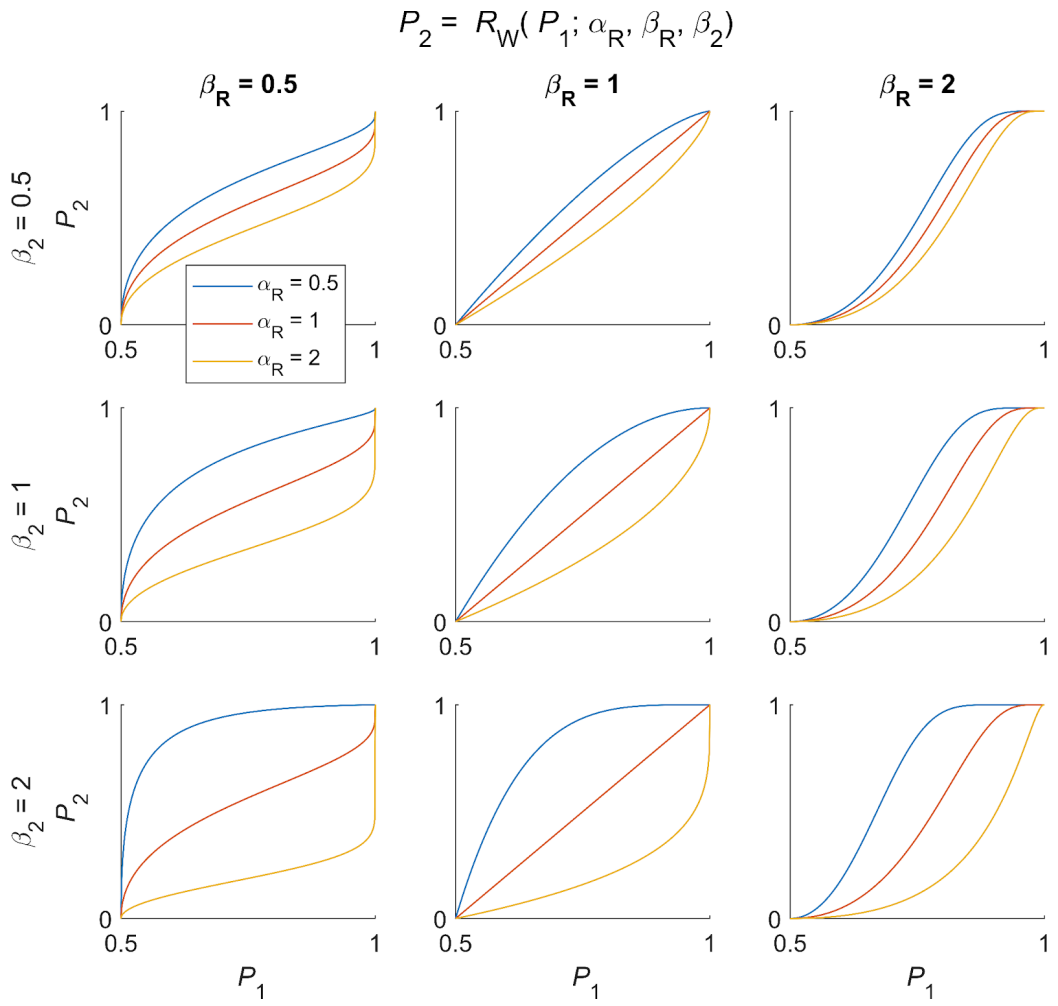

Figure 2. Behavior of the Weibull RPF (R_W) as a function of its three main parameters. Each plot shows how performance on one task (P_2) depends on performance on another task (P_1) according to the Weibull RPF specified by parameters for relative location $\alpha_R = \frac{\alpha_2}{\alpha_1}$ (separate lines within each plot), relative slope $\beta_R = \frac{\beta_2}{\beta_1}$ (columns), and F_2 slope β_2 (rows), as derived from the parameters of the component Weibull functions F_1 and F_2.

Overall, the behavior of the Weibull RPF is considerably more complicated than the standard Weibull. First, we can observe that R_W does not have a direct analogue to the Weibull's slope parameter β, but rather has a variable shape and degree of curvature depending on the relative slope β_R and relative location α_R. When $\beta_R > 1$, R_W is sigmoidal; when $\beta_R < 1$, R_W is inverse sigmoidal; and when $\beta_R = 1$, R_W is concave down, linear, or concave up depending on α_R. The complexities of how R_W's shape changes depending on combinations of parameter values obscures a straightforward and universally applicable interpretation of the function in terms of a

slope parameter, and so it is not clear that characterizing R_W in terms of a slope parameter is the best way to insightfully summarize its behavior.

Similarly, R_W does not have a simple analogue to the Weibull's location parameter α. In the Weibull (Eq. 3), α acts as a location parameter in the sense that the function takes on 63.2% of its maximal value above chance (i.e. 63.2% of the way between γ and $1 - \lambda$) when $x = \alpha$, since substituting this value of x into the formula entails $1 - e^{-\left(\frac{x}{\alpha}\right)^{\beta}} = 1 - e^{-1} = 0.632$. Thus, α tells us what value of x ("location") yields this threshold function value of 63.2% of the above-chance maximum. From Eq. 5, we see that R_W achieves 63.2% of its above-chance maximum when $\frac{\alpha_1}{\alpha_2}\left(\ln\left(\frac{1-\lambda_1-\gamma_1}{1-\lambda_1-P_1}\right)\right)^{\frac{1}{\beta_1}} = 1$. Solving for P_1 in this equation yields R_W's equivalent of the Weibull's location parameter α, which can be expressed as

$$P_1 = \gamma_1 + \left(1 - \lambda_1 - \gamma_1\right)\left[1 - e^{-\left(\frac{\alpha_2}{\alpha_1}\right)^{\beta_1}}\right] = F_1(\alpha_2) \quad (6)$$

Thus, R_W takes on its threshold value at the value of P_1 given by $F_1(x)$ evaluated at the location parameter of F_2, i.e. at $P_1 = F_1(\alpha_2)$. This result is intuitive in that the RPF derives from a transformation of the input variable of F_2 from x to P_1, while leaving F_2's output P_2 unchanged (Eq. 2). Thus, since α_2 is the value of x at which F_2 achieves its threshold value of P_2, R_W must achieve its threshold value of P_2 at whatever value of P_1 that α_2 maps onto in the RPF transformation, which is just $F_1(\alpha_2)$.

Although the value of $F_1(\alpha_2)$ provides a measure of what P_1 value yields R_W's threshold value, its interpretation is more complex than that of α for the Weibull function. Intuitively, lower and higher values of α in the Weibull function roughly correspond to the curve "shifting" or "tilting" left or right on the x-axis[3]. By contrast, the value at which R_W achieves its threshold value is strongly influenced by its curvature, which in turn depends on multiple parameters from F_1 and F_2. For instance, in the lower-left plot of **Figure 2**, the concave down and concave up curves achieve their threshold values at very low and high values of P_1, respectively, due primarily to their differences in curvature. This difference in threshold location cannot be attributed to a shift, tilt, or translation in an otherwise similar curve, as is the case for the Weibull, and thus $F_1(\alpha_2)$ cannot serve the same conceptual role as the Weibull's location parameter α.

[3] The slope of the Weibull function is actually controlled by both α and β, although when plotted against log x, α controls function translation and β controls slope (Kingdom and Prins, 2016).

Thus, it appears that while there are indeed aspects of the Weibull RPF's behavior that can be summarized with a small number of parameters – β_R and α_R control shape, and $F_1(\alpha_2)$ determines threshold – it is not clear that these parameters provide the same ease of interpretation and leverage for understanding the behavior of the RPF in terms of psychophysical performance as their counterparts α and β do for conventional psychometric functions. Furthermore, the exact mathematical formulation for such parameters depends on the psychometric functions used for F_1 and F_2, entailing that different choices for these functions may lead to different formulations for RPF summary parameters. These difficulties motivate the alternative approaches for comparing RPFs across conditions that we develop and discuss below.

Comparing RPFs across conditions

If using parameter values to summarize aspects of RPF behavior is not as straightforward and perhaps not as fruitful as it is for conventional psychometric functions, what alternatives are there for using RPFs to enrich our understanding of psychological processes?

One major goal of RPF analysis would be to investigate how the relationship between two target psychological processes changes across different conditions. For example, analysis of the metaperceptual RPF would be well-suited to address questions on how the relationship between objective and subjective aspects of perceptions are influenced by various factors, such as, "Is the relationship between subjective judgments and task accuracy the same in central versus peripheral visual field locations?" (Odegaard et al., 2018a; Winter and Peters, 2022) or "Does transcranial magnetic stimulation to a certain region of interest alter the relationship between confidence and task accuracy?" (Peters et al., 2017a; Rahnev et al., 2012; Rounis et al., 2010; Ruby et al., 2018).

The behavior of the RPF across conditions sheds light on the relationship between P_1 and P_2. Any across-condition changes in the RPF would indicate a differential effect of condition on P_1 and P_2, such that the changes in P_2 due to condition could not be solely attributed to changes in P_1 (or else the RPF would be identical), and would demonstrate that P_1 and P_2 are produced by at least partially separable processes. Alternatively, if P_1 and P_2 differ across conditions, but do so in such a way that preserves the RPF describing their relationship, this would be consistent with the possibility that the changes in P_2 are indeed attributable entirely to changes in P_1 (or vice versa), or that both are products of a single underlying process characterized by a constant RPF.

Below we consider two approaches to comparing RPFs across conditions: an AUC-based approach and a model comparison approach.

AUC approach: area under the RPF curve

As discussed in the Introduction, the motivating example behind this work is performance matching in the consciousness / metacognition literature, in which we seek to find conditions where objective task performance (P_1) is the same but subjective reports of awareness or

confidence (P_2) differ (Morales et al., 2022). Notice that this approach essentially attempts to compare a vertical slice of two RPFs at a particular P_1 value. Thus, a natural generalization of the performance matching approach is to compare two RPFs across a fixed *interval* of P_1 values rather than at a single fixed value. Within a given RPF, summing the values of P_2 across the entire interval of P_1 values amounts to computing the area under the curve (AUC) of the RPF, and dividing this AUC by the length of the P_1 interval yields the average value of P_2 over that interval. These AUCs and average P_2 values can then be compared across conditions to assess whether condition affects P_2 over and above any effects it may have on P_1.

More formally, the RPF AUC is given by

$$\mathrm{AUC} = \int_a^b R(P_1)dP_1 \tag{7}$$

This integral can be computed without an analytic solution, and indeed without specifying an equation for *R*, by using $x = F_1^{-1}\left(P_1\right)$ and $P_2 = F_2(x)$ to compute the RPF as $P_2 = F_2\left(F_1^{-1}\left(P_1\right)\right)$ (Eq. 2) and performing numerical integration.

Since the aim of this analysis approach is to compare AUCs across conditions for a fixed interval of P_1 values, it must be the case that all RPFs being analyzed fully span that interval. In general, this is not guaranteed to be the case unless the fixed P_1 interval is chosen appropriately. For instance, in a grating tilt discrimination task having conditions where the grating is attended or unattended across several levels of grating contrast spanning the full possible range of contrasts from 0 to 1, the fitted psychometric function for p(correct) in the attended condition may range from chance performance of 0.5 at zero contrast to a near-ceiling value (e.g. 0.98) at maximal contrast, whereas the fitted function for the unattended condition may range from chance performance at zero contrast to a level of performance at maximal contrast that is considerably lower than in the attended condition (e.g. 0.8). Thus, although the attended condition RPF spans a P_1 interval of [0.5, 0.98], its AUC can only be compared to that of the unattended condition for a fixed P_1 interval over [0.5, 0.8].[4]

Thus, the intervals of P_1 values exhibited by each RPF being compared jointly determine lower and upper bounds on possible intervals of P_1 values that are common to all RPFs. The lower bound *L* on the common P_1 interval is given by

$$L = \max_c \min_x P_{1\,c,x} \tag{8}$$

[4] Note that the P_1 interval over which two conditions are compared should be fixed across conditions (e.g. attended, unattended) within a subject, but can be allowed to vary across subjects in an experiment. Here we derive the within-subject comparison process.

where $P_{1\,c,x}$ denotes the value of P_1 at condition *c* and stimulus level *x*. In other words, the lower bound for a common P_1 interval across conditions is the minimal *within-condition* value of P_1 that is maximal *across* conditions. By similar reasoning, the upper bound U is given by

$$U = \min_c \max_x P_{1\,c,x} \tag{9}$$

i.e. the maximal within-condition value of P_1 that is minimal across conditions. For AUCs to be computed with a fixed P_1 interval [*a*, *b*] that is common to all conditions, it must be the case that

$$a \geq L,\ b \leq U \tag{10}$$

and the widest possible common P_1 interval is given by [*L*, *U*].

These considerations are illustrated in **Figure 3**. The psychometric functions for P_1 in conditions A and B have different values at the minimum and maximum values of *x* (left panel), which entails that their corresponding RPFs do not span the same range of P_1 values (right panel). Thus, to compare AUC for a fixed P_1 interval, this interval must be restricted to the set of P_1 values that is common to both functions (shaded region). The lower and upper bounds of this common interval are set by the largest across-condition minimum P_1 value and the smallest across-condition maximum P_1 value, respectively, as described in Eqs. 8 and 9.

In the example of **Figure 3**, RPF AUC is larger for condition A than for condition B. This indicates that condition influences P_2 over and above its influence on P_1, and that this differential effect holds over a wide range of P_1 values.

Normalizing the AUC by the length of the P_1 interval over which it is computed yields the average P_2 value over that interval:

$$\overline{P}_2 = \frac{1}{b-a}\int_a^b R(P_1)dP_1 = \frac{\mathrm{AUC}}{b-a} \tag{11}$$

The quantitative values of this metric are more intuitive to interpret than those for AUC. The normalization it provides may also be desirable in cases where the P_1 intervals used to compute AUC differ across subjects, since P_1 interval size influences AUC and the spirit of this analysis approach is to factor out or control for the influence of P_1. However, this consideration is mitigated somewhat given that the effect of interest pertains to within-subject differences in AUC as a function of condition, and the P_1 interval within each subject is constant across conditions.

$\overline{P}_2$ also has an intuitive connection to the performance matching approach discussed above. Whereas performance matching seeks to measure the difference between subjective reports at

a fixed value of task performance, comparing $\overline{P}_2$ for the metaperceptual RPF across conditions gives the *average* difference between subjective reports over a *range* of task performance levels.

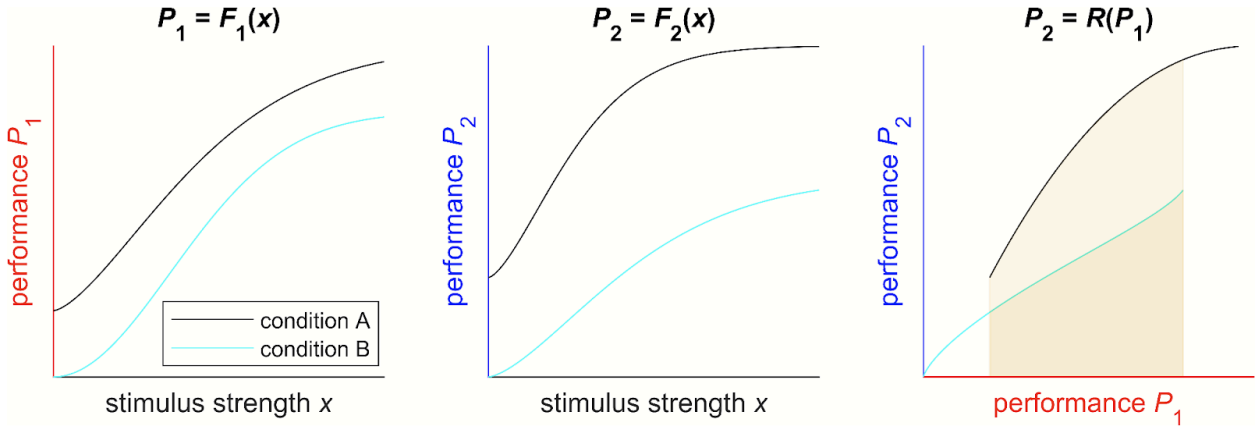

Figure 3. Comparing RPFs with the AUC approach. In this illustrative example, psychometric functions for P_1 and P_2 differ across condition, and so do their corresponding RPFs. The difference in the RPFs can be quantified by comparing the area under the curve (AUC) over the set of P_1 values that both RPFs share in common (shaded region). Here, condition A has the higher AUC, indicating higher levels of P_2 across the fixed P_1 interval. The AUCs can be divided by the length of this common P_1 interval to yield the average P_2 values over the interval.

Nonparametric computation of AUC

In the foregoing, we have assumed that P_1 and P_2 data are fitted with psychometric functions F_1 and F_2. However, there may be cases where fitting P_1 and/or P_2 encounters difficulties, such as:

- The researcher may be uncertain about the most appropriate functional form to choose for F_1 and/or F_2.
- The researcher may prefer to avoid making parametric assumptions about F_1 and/or F_2.
- For certain dependent variables, it may be unclear, complicated, and/or labor intensive to develop an MLE fitting approach and implement the fitting procedure in analysis code (consider e.g. the discussion of MLE fitting for *d'*, meta-*d'* and mean rating as discussed above and in **Supplemental Material**).
- When plotted against stimulus strength *x*, the data to be fitted may be monotonically decreasing (e.g. reaction time data) or non-monotonic (e.g. when rating confidence in a detection task, confidence may be high for "no" responses at low values of *x* and "yes" responses at high values of *x*, yielding a U-shaped function of confidence when collapsed across response type), whereas standard psychometric functions are monotonically increasing with *x*.
- Limitations and noise in the data may cause technical difficulties with the fitting procedure, or may yield fitted parameter values that are implausible or present analysis difficulties (e.g. infinite slope).

These difficulties can be circumvented by computing AUC nonparametrically. The simplest nonparametric approach is to perform linear interpolation between the data points in the plot of P_2 vs. P_1 and compute AUC from the resulting trapezoids, analogous to the nonparametric measure of area under the ROC curve A_g (Pollack and Hsieh, 1969). A hybrid approach can also be applied in which the RPF is constructed from a parametric fit of $P_1 = F_1(x)$ and a nonparametric estimation of $P_2 = F_2(x)$ via interpolation. (However, note that a hybrid approach where P_1 data are interpolated and P_2 data are fitted is not viable, since the function yielded by interpolation of P_1 will in general not be monotonic with x and so will not be invertible, preventing the computation of the RPF as described in Eq. 2.)

In **Supplemental Material** we discuss methodological considerations for nonparametric computation of RPF AUC in more detail, and in **Supplemental Material** we present simulations demonstrating that nonparametric methods are similarly effective to parametric methods at estimating the true AUC of a known generating RPF under data collection conditions typical of those used in psychophysical experiments.

Benefits and limitations of the AUC method

Summarizing RPFs with AUC (or $\overline{P}_2$) in this manner has a number of virtues:

1. **Ease of computation.** RPF AUC can be computed via numerical integration based on $F_1(x)$ and $F_2(x)$ without needing to find a closed form expression for $R(P_1)$.
2. **Ease of interpretation.** RPF AUC provides a single, easy to interpret measure (compare to the multiple, complex, interrelated parameters of the Weibull RPF, for example).
3. **Universality.** AUC computations are applicable to any RPF for any P_1 and P_2, regardless of the functional forms of F_1 and F_2.
4. **Robustness.** AUC is more robust to measurement error than general psychometric function parameter estimation. For instance, in certain cases small changes in the data can yield large differences in the fitted parameters without having large effects on the overall shape of the psychometric function, which in turn would lead to only small changes in the RPF AUC. In fact, AUC estimation can even be robust if $F_n(x)$ is constructed from piecewise linear interpolation rather than fitting a function, further simplifying the analysis approach; we explore this possibility in detail in the **Supplemental Material**.

The AUC method is most straightforward to interpret in cases where the RPFs do not intersect over the chosen fixed P_1 interval, since in such cases the values of P_2 in one condition are always higher than in the other for every value of P_1 in the interval. However, if the empirical RPFs *do* intersect in this interval, then the relationship between AUCs across conditions differs on either side of the intersection point, which complicates interpretation of AUC computed over the whole interval. Two possibilities must be considered: (1) the “true” generating RPFs are similar or identical over this interval, and the intersection in the empirical RPFs is due to statistical noise; or (2) the generating RPFs are distinct and do indeed intersect over this interval, as is validly reflected in the empirical RPFs. Since across-condition AUCs have opposite relationships on either side of the intersection point, computing AUC over the entire

interval will tend to wash out any across-condition differences. This behavior can be a virtue that accurately reflects the absence of an effect in case (1), but may underestimate or even fail to detect the presence of a true effect in case (2).

For instance, consider an idealized case where over a P_1 interval [0, 1], the empirical RPF in condition A has a constant value of 0.5, and the empirical RPF in condition B is linear with values [0, 1] at the endpoints of the P_1 interval. In this case, RPF A forms a rectangle with base 1 and height 0.5, and RPF B forms a triangle with base 1 and height 1 that intersects RPF A at $P_1 = 0.5$. Both RPFs have an AUC of 0.5 despite differing considerably in their shape, since A's AUC is larger than B's over $P_1 \in [0, 0.5]$ and the opposite is true over $P_1 \in [0.5, 1]$.

Thus, if the "true" generating RPFs for A and B are similar or identical, and the empirical RPFs A and B differ due to noise, their identical AUCs will accurately reflect the absence of a difference in the generating RPFs. Conversely, if the generating RPFs have forms that are well represented by the empirical RPFs A and B, then computing AUC over the interval [0, 1] will fail to quantify the difference between the generating RPFs due to their intersection over that interval. In cases where the generating RPFs intersect in this way, the model comparison approach described below can still detect the difference between them.

Model comparison approach

An alternative approach to comparing RPFs across conditions is to capitalize on the observation that if a functional form for the RPF is available, the parameters of this function can be constrained in such a way as to ensure that fitted RPFs across conditions are identical. The data can then be fitted with two different models, one of which allows parameters to vary freely in such a way that the fitted RPFs can differ across conditions ("free model"), and one of which constrains parameters in such a way that the fitted RPFs are constrained to be constant across conditions ("constrained model"). Standard model comparison analysis approaches can then be conducted to investigate whether the free or constrained model provides a better account of the data, taking into account how the greater degrees of freedom in the free model introduce the possibility of overfitting. This model comparison analysis can be performed e.g. with information theoretic measures such as Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) (Vrieze, 2012), or alternatively with cross validation methods to assess model generalizability (de Rooij and Weeda, 2020).

For instance, consider the functional form of the Weibull RPF R_W discussed above (Eq. 5). A trivial way to ensure that R_W is constant across conditions would be to constrain all parameters for F_1 and F_2 to be constant across conditions. However, a more artful approach would be to allow the parameters of F_1 and F_2 to have the maximal degree of freedom possible while still constraining the corresponding RPFs to be constant across conditions. Investigation of Eq. 5 shows that this latter goal can be achieved by constraining the following parameter values to be constant across conditions: γ_1, λ_1, β_1, γ_2, λ_2, β_2, and $\frac{\alpha_2}{\alpha_1}$. This set of constraints allows α_1 and α_2 to vary across conditions in the fitting of F_1 and F_2, so long as the resulting parameter values

conserve constant values for $\frac{\alpha_2}{\alpha_1}$ across conditions. The free model, by contrast, would relax some or all of these constraints and thus allow fitted RPFs to differ across conditions.

Generalizing the above discussion, there are actually multiple ways to define “constrained” and “free” models, depending on what constraints on across-condition parameter values are imposed over and above the key set of constraints determining whether RPFs can vary across conditions or not. An extended model comparison analysis could thus consider a family of models, some of which are constrained and others of which are free in the way defined above. Interpretation of the results of such an analysis could reveal findings such that e.g. the best fitting model constrains $\frac{\alpha_2}{\alpha_1}$ but not the β parameters, or similar patterns, which could provide a more nuanced understanding of how condition influences the RPFs and the mechanisms underlying their behavior.

Regardless, the most basic and foundational question would still be whether the empirical RPFs are best characterized by constrained or free models. If free models are best supported in the model comparison analysis, this would suggest that RPFs are modulated by condition, and thus that the psychological processes generating P_1 and P_2 are at least partially separable. Conversely, if constrained models are best supported, this would suggest that the observed RPFs are consistent with the possibility that P_1 determines P_2 (or vice versa), or that both are generated by a single underlying process characterized by a constant RPF.

The model comparison approach has the advantages over the AUC approach that it can detect differences in RPFs even in cases where RPFs intersect in a way that yields similar AUC values, and that it can more specifically pinpoint which aspects of RPF behavior are influenced by condition. However, it has the disadvantages that it is more complex and resource intensive to conduct, and requires deriving an analytic expression for the RPF.

Empirical case study

In this next section, we demonstrate the power and utility of the RPF method by applying it to an empirical dataset in which subjects made perceptual decisions about coherent dot motion and rated confidence. Seven levels of motion coherence were presented, allowing construction of psychometric functions for accuracy, confidence, and metacognitive sensitivity. Experimental conditions were contrived so as to attempt to modulate the relationship between confidence and accuracy, naturally inviting an RPF analysis approach.

Experimental methods

Twenty-one healthy adult human subjects viewed random dot kinematogram (RDK) stimuli which continuously filled the entirety of a computer monitor with random dot motion. In a two-alternative (2AFC) task design, on each trial of the experiment a circular patch of these dots to the left or right of a central fixation cross briefly displayed coherent motion in a downward direction. The observer’s task was to indicate which side of the display contained the coherent

downward motion and rate their confidence on a scale of 1-4; they reported both choices with a single keypress.

We varied three aspects of the task to examine their effects on the relationship between accuracy and confidence. First, we varied motion coherence by randomly selecting the coherence of the downward dot motion on each trial from a list of seven values evenly spaced between 10% and 80% coherence. Second, we varied Dot Density by setting the density of the dots across the whole display on each trial to one of three levels (Low = 1 dot/deg^2, Medium = 3 dots/deg^2, High = 9 dots/deg^2). Third, we varied changes in Dot Density across trials by either setting Dot Density randomly on each trial (Trial Structure: Interleaved), or by holding Dot Density constant with each block of trials (Blocked).

Please see **Supplemental Material** for full details of participants, stimuli, equipment, and experimental design.

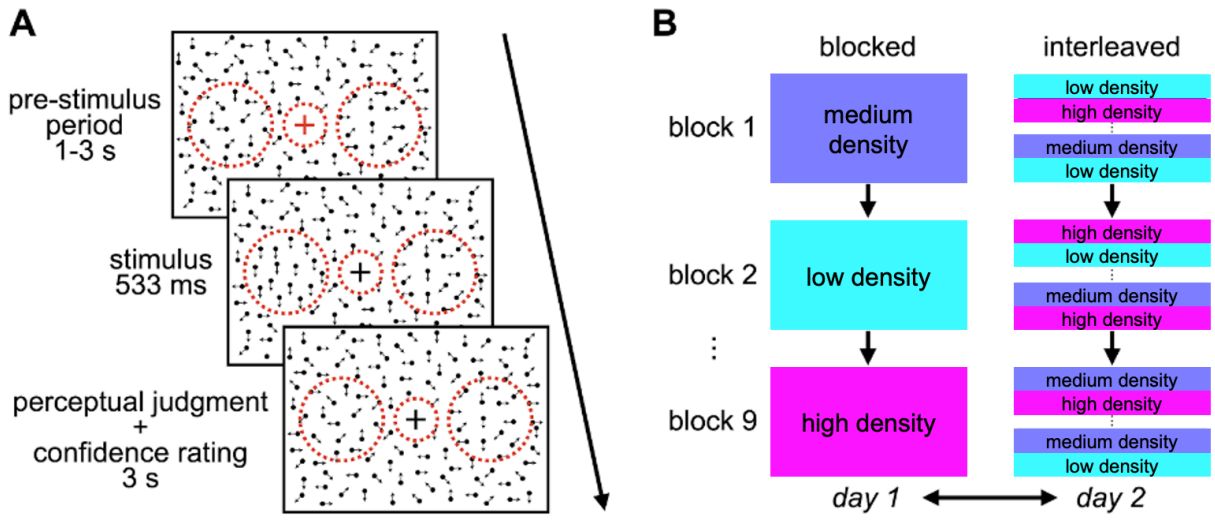

Figure 4. Behavioral task procedures. (A) Each trial began with a pre-stimulus period, during which full-field random dot motion was shown (black arrows illustrate dot motion direction). Subsequently, within one of two circular regions of the screen (indicated here by the red circles to the left and right of fixation—red circles were shown to participants only during preliminary practice trials but not during experimental trials), coherent downward dot-motion occurred, followed by a response period in which participants indicated on which side they saw the coherent motion and their decision confidence. The central red circle indicates an area around the fixation cross where no dots were presented; this red circle was not shown to participants and is used here for illustration purposes. (B) Participants underwent two Trial Structure conditions, Blocked and Interleaved, on two different days of testing. In the Blocked condition, Dot Density was constant across trials within a given block, whereas in the Interleaved condition, Dot Density varied randomly across trials. Blocked versus Interleaved days and order of density blocks was counterbalanced across all participants.

Data analysis

Following previous demonstrations (Koizumi et al., 2015; Odegaard et al., 2018b; Rollwage et al., 2020; Samaha et al., 2016; Stolyarova et al., 2019), we expected higher Dot Density conditions to yield higher confidence, even when task performance was similar. We also examined whether metacognitive sensitivity – quantified as meta-*d'* – would differ across Dot Density conditions. To explore these possibilities, we fit Weibull psychometric functions to *d'*, mean confidence, and meta-*d'* as a function of stimulus strength *x* (RDK coherence) for each subject in each condition (Dot Density: High, Medium, Low; Trial Structure: Blocked, Interleaved) using the methods for MLE fitting of these variables developed in the **Supplemental Material**. This allowed us to specify two categories of metaperceptual RPFs, one relating *d'* to mean confidence and another relating *d'* to meta-*d'*; we computed these for each subject and each condition. All RPF analyses were performed in the described manner through our open-source RPF toolbox, available at https://github.com/CNClaboratory/RPF.

For each of these objective-subjective pairs, we then computed the AUC and $\bar{P}_2$ for each Density and Trial Structure separately for each subject, and submitted these to 3 (Dot Density: High, Medium, Low) x 2 (Trial Structure: Blocked, Interleaved) repeated-measures analyses of variance (ANOVAs).

Empirical results and discussion

Having implemented the above-described analyses for each subject in each condition (blocked vs interleaved Trial Structure, and low, medium, and high Dot Density), we found that Dot Density did indeed affect both mean confidence judgments and metacognitive sensitivity (meta-*d'*) over and above any effects on *d'*, primarily in the Interleaved but not Blocked Trial Structure.

In the plotted data (**Figures 5 & 6**), for illustrative purposes we show RPF curves for each Trial Structure and Dot Density condition that are fitted to the entire combined group data concatenated across all subjects, rather than an average across the fitted curves for each subject individually. However, we remind the reader that all statistical measures were derived from single-subject fits.

For mean confidence versus *d'* (**Figure 5**), using the raw AUC in 3 (Dot Density: High, Medium, Low) x 2 (Trial Structure: Blocked, Interleaved) repeated-measures ANOVA we found a main effect of Dot Density ($F(2,40) = 7.6633$, $p = 0.0015$, $\eta_p^2 = 0.277$) but not Trial Structure ($F(1,40) = 2.0613$, $p = 0.1665$, $\eta_p^2 = 0.093$), and a significant Trial Structure x Dot Density interaction ($F(2,40) = 4.0675$, $p = 0.0247$, $\eta_p^2 = 0.169$) such that mean confidence increased with increasing Dot Density in the Interleaved but not Blocked Trial Structure. The pattern was similar when we used a second repeated-measures ANOVA to examine the normalized AUC measure $\bar{P}_2$, with a

main effect of Dot Density (F(2,40) = 6.4047, p = 0.0039, η_p^2 = 0.243) but not Trial Structure (F(1,40) = 0.8951, p = 0.3554, η_p^2 = 0.043) and a marginal Trial Structure x Dot Density interaction (F(2,40) = 3.2056, p = 0.0511, η_p^2 = 0.138) – again suggestive that mean confidence increased with increasing Dot Density primarily in the Interleaved but not Blocked Trial Structure.

For metacognitive sensitivity (meta-*d'*) versus *d'* (**Figure 6**), we observed a similar pattern. Using raw AUC, a repeated-measures ANOVA revealed a main effect of Dot Density (F(2,40) = 5.6903, p = 0.067, η_p^2 = 0.221) but not Trial Structure (F(1,40) = 0.0092, p = 0.9244, η_p^2 = 0), and again a Trial Structure x Dot Density interaction (F(2,40) = 4.0087, p = 0.0259, η_p^2 = 0.167) such that meta-*d'* was significantly higher with increasing Dot Density in the Interleaved but not Blocked Trial Structure. A final repeated-measures ANOVA on the normalized AUC measure $\bar{P}_2$ for meta-*d'* revealed again a main effect of Dot Density (F(2,40) = 4.4246, p = 0.0184, η_p^2 = 0.181) but not Trial Structure (F(1,40) = 0.8020, p = 0.3811, η_p^2 = 0.039), and a trending interaction between Trial Structure and Dot Density (F(2,40) = 2.9437, p = 0.0642, η_p^2 = 0.128) – again suggestive that metacognitive sensitivity was higher in higher Dot Density conditions primarily in the Interleaved trials.

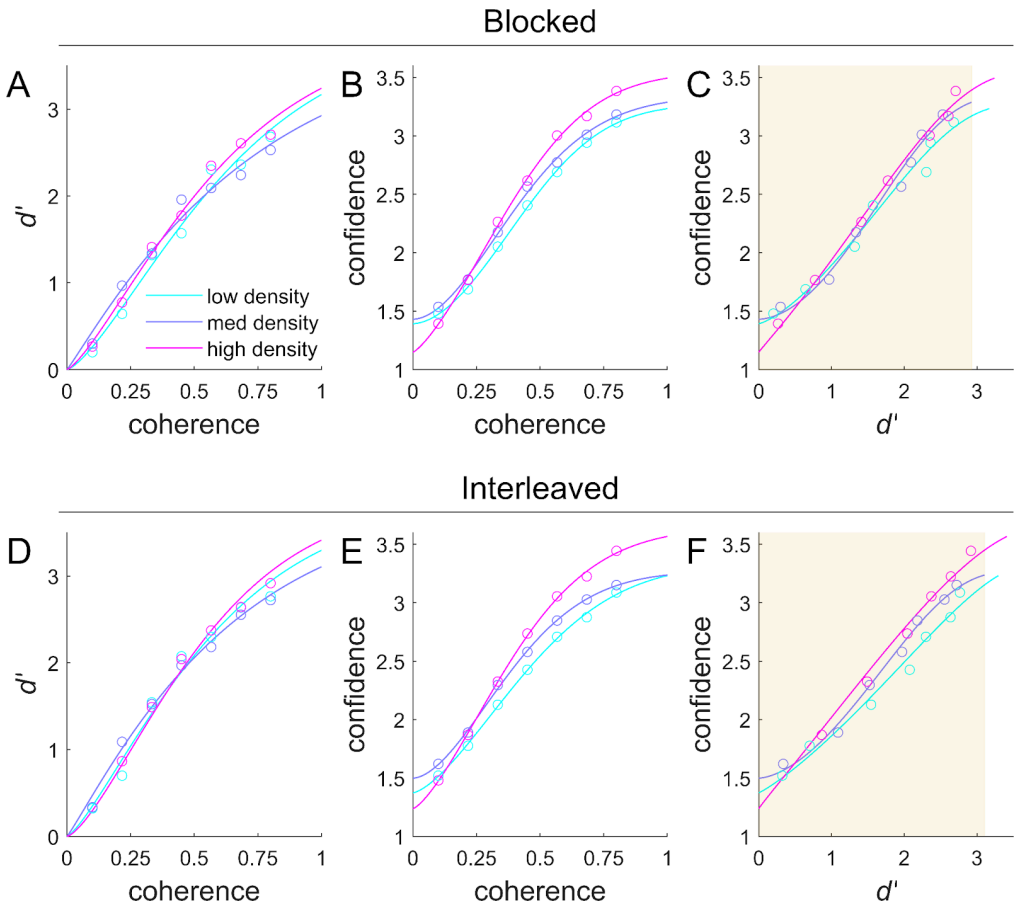

Figure 5. Results of the empirical case study showing the metaperceptual RPF relating *d'* and mean confidence ratings. Plots here visualize the statistical effects across subjects (see **Figure 7**) via direct fits to the group-level data. (A-C) show F_1 (*d'* versus RDK coherence), F_2 (mean confidence versus RDK coherence), and the RPF *R* (mean confidence versus *d'*) for the Blocked trials; (D-F) show the same for the Interleaved trials. Fitted RPFs for the Blocked (C) and Interleaved (F) Trial Structure show visually that the Blocked trials resulted in no apparent differences in mean confidence as a function of Dot Density, while the Interleaved trials show RPF separation with higher mean confidence in higher Dot Density conditions over the same interval of task performance.

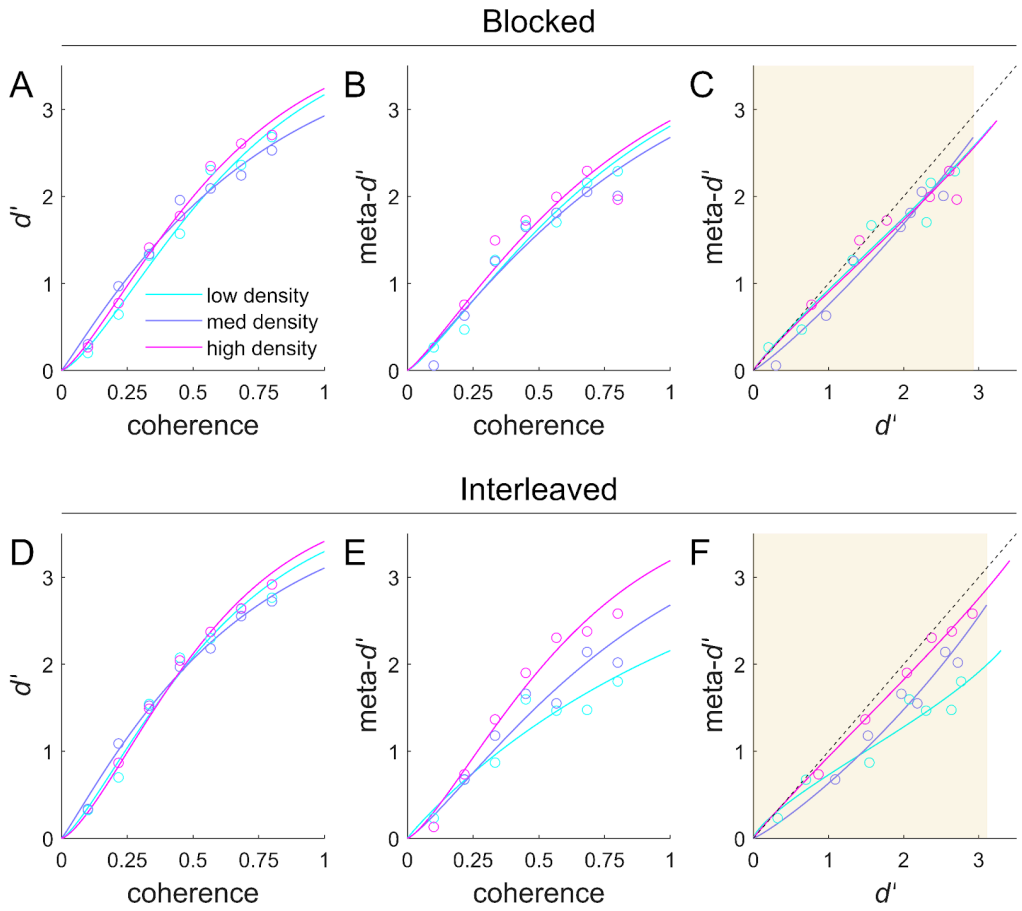

Figure 6. Results of the empirical case study showing the metaperceptual RPF relating *d'* and metacognitive sensitivity, measured with meta-*d'*. Plots here visualize the statistical effects across subjects (see **Figure 7**) via direct fits to the group-level data. Similar to the plots for mean confidence, (A-C) show F_1 (*d'* versus RDK coherence), F_2 (meta-*d'* versus RDK coherence), and the RPF *R* (meta-*d'* versus *d'*) for the Blocked trials; (D-F) show the same for the Interleaved trials. Fitted RPFs for the Blocked (C) and Interleaved (F) Trial Structure show visually that the Blocked trials resulted in no apparent differences in meta-*d'* as a function of Dot Density, while the Interleaved trials show RPF separation with higher meta-*d'* in higher Dot Density conditions over the same interval of task performance.

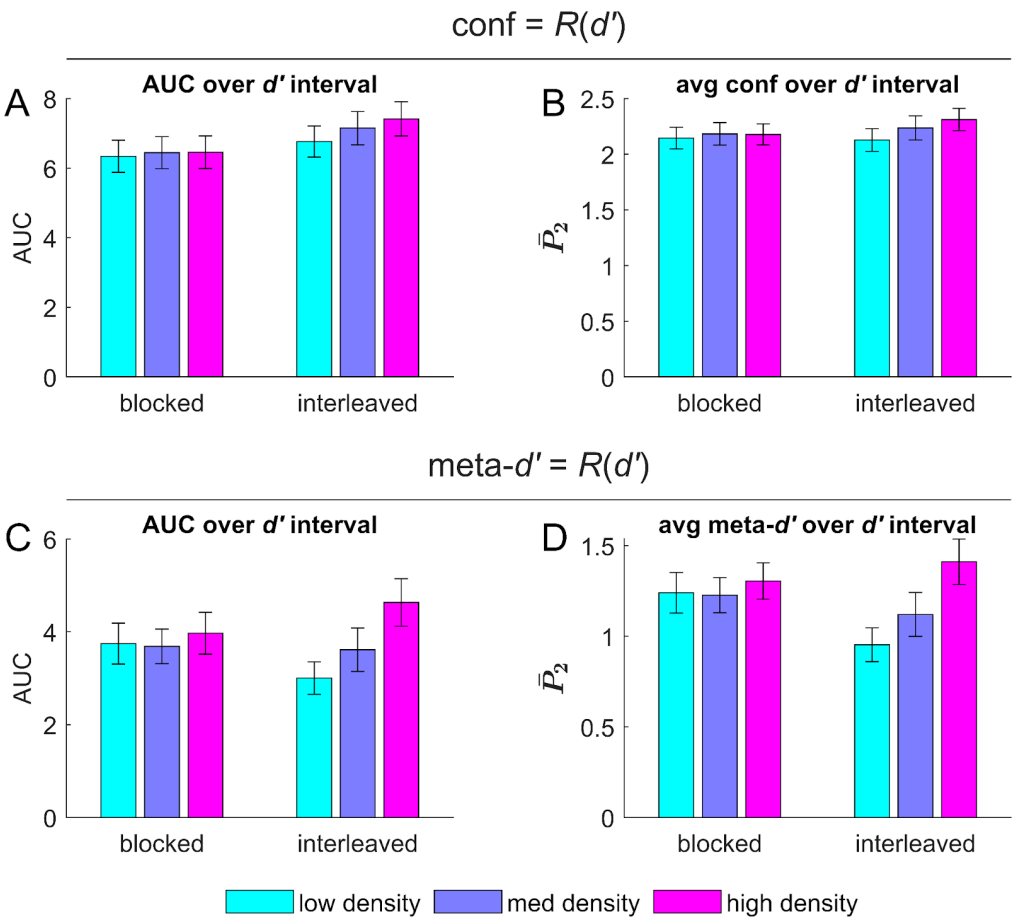

Figure 7. Results of the empirical case study showing the AUC-based analyses using raw AUC and $\bar{P}_2$. Both raw AUC (A, C) and its normalized variant, $\bar{P}_2$ ("avg", B, D) confirm visual inspection of the RPFs in the previous figures, showing significant main effects of Dot Density and interactions between Trial Structure (Blocked vs Interleaved) and Dot Density. That is, increasing Dot Density led to both higher mean confidence and higher metacognitive sensitivity (meta-*d'*) over a matched performance interval, especially in the Interleaved trials. See main text for statistical details.

Together, these results demonstrate that Dot Density does indeed affect both mean reported confidence rating and metacognitive sensitivity (meta-*d'*), especially when density is varied pseudorandomly on every trial.

These findings are of utility to the community in several ways. First, from a basic science perspective, the observation that a manipulation as simple as the density of an RDK can induce changes in overall mean confidence over and above any effect on task performance capacity is consistent with findings in the literature on the so-called "positive evidence" or "response-congruent evidence" bias in metacognition (Rollwage et al., 2020; Samaha and Denison, 2022). In a number of empirical investigations, it has been shown that higher amounts of absolute magnitude of evidence available to the observer to make a perceptual decision are associated with increased subjective confidence reports; these manipulations of evidence can

take the form of contrast or luminance (e.g., (Koizumi et al., 2015; Rausch et al., 2017; Samaha et al., 2019, 2016)), dot motion coherence (Zylberberg et al., 2012), or even more cognitive type evidence such as facial attractiveness (Ceja et al., 2022). Models have been proposed to account for these and similar findings, placing constraints on how confidence might be (neurally) computed in perceptual decisions (e.g., (Maniscalco et al., 2021, 2016; Peters et al., 2017c)).

Interestingly, here we meaningfully add to these previous findings by showing that in addition to Dot Density influencing overall confidence ratings separately from its influence on task performance, Dot Density *also* influences metacognitive sensitivity – here measured with meta-*d'*. This new empirical finding meaningfully adds to understanding of the computational relationship between performance and confidence ratings even on a trial-by-trial basis, providing rich opportunities for future computational studies to use these and similar kinds of data to arbitrate among candidate process models giving rise to metacognitive judgments (e.g., (Adler and Ma, 2018a, 2018b; Aitchison et al., 2015; Denison et al., 2018; Kiani et al., 2014; Maniscalco et al., 2016; Miyoshi and Lau, 2020; Peters et al., 2017a, 2017c; Peters and Lau, 2015; Winter and Peters, 2022; Zylberberg et al., 2016, 2014)).

However, our intent with this empirical demonstration was not only to demonstrate that manipulations of dot motion evidence may influence confidence in separate ways than influences on task performance, which had previously been established. Here, we were also concerned with showing how full RPF analyses can provide benefit over measuring differences in confidence or metacognitive sensitivity at one or two levels of (matched) performance.

Importantly, one can see clearly through the RPF approach that the size of the effect on subjective experience depends strongly on the performance level at which the effect is measured: for both mean confidence (**Figures 5 & 7**) and metacognitive sensitivity (**Figures 6 & 7**), the difference in the subjective report appears to grow as a function of task performance; this occurs with specific quantitative relationship to task performance in both cases. By extension, if a researcher were to try to measure the effect size of a manipulation's influence on mean confidence or meta-*d'*, but were unable to precisely match task performance across conditions or subjects, the effect size of interest would be at best poorly estimated, or at worst entirely missed (e.g., at lower levels of *d'*). By measuring and fitting the entire RPF and engaging in the AUC-based analyses presented here, such differences due to nuisance variables can be minimized, revealing a robust and quantitatively precise measure of subjective experience differences independent of task performance. Importantly, process models of metacognition or subjective experience in general – such as those mentioned above – become much more highly constrained if they must explain behavioral data across the entire RPF in multiple conditions, opening an exciting new set of questions for the community (Fleming, 2023; Peters, 2022).

While the data presented here perhaps consist of fewer subjects than would be ideal, we view this limitation as a strength because they nevertheless demonstrate the robustness of the RPF method to small numbers of subjects or even to few trials per condition. Here, we collected only 36 trials per condition (7 levels of RDK motion coherence, Blocked vs Interleaved Trial

Structure, and 3 levels of Dot Density). The entire dataset was collected across only approximately two hours of testing per subject, meaning that even as few as 36 trials per condition can be sufficient for conducting robust and precise RPF analyses comparing across conditions with the AUC-based metrics. (Of course, more trials are better, and following best practices for fitting d' or meta-d' in any dataset would suggest at least 100 trials per condition for robust estimates of these metrics.)

Overall, this empirical case study highlights an exciting direction for the study of subjective experience and for use of the RPF analytic approach in general. We believe these results and the analytic approach to be of great value both within the metacognition and subjective experience community (Michel et al., 2019; Rahnev et al., 2022) and beyond.

General discussion and future directions

Summary

In this piece we have laid out a novel framework for investigating, in general, the quantitative relationship between two psychological processes measured under noisy conditions and how these relationships may vary with any experimental manipulation or intervention that is of interest to the researcher. This framework includes the derivation of the *relative psychometric function* (RPF) under parametric assumptions, including special considerations for fitting customized psychometric functions to non-standard psychometric variables such as task performance capacity measured with the signal detection theoretic metric *d'*, confidence ratings, and metacognitive sensitivity (meta-*d'*) (Maniscalco and Lau, 2014, 2012). We also developed and tested a series of metrics and algorithms designed to provide intuitive insight into how the RPF may change across experimental condition, including the area under the RPF (AUC) method and its normalized variant, $\bar{P}_2$. These metrics provide a clear, precise, and interpretable approach for interpreting variations in RPFs across experimental conditions. And for those who wish to precisely evaluate other relationships among RPFs or how they may be captured by process models (e.g. signal detection or Bayesian decision theoretic, evidence accumulation models, etc.), we also lay out a model comparison approach in which the RPF can be constrained to be equivalent across conditions or free to vary in different ways. This model-based approach can provide important nuance and context to supplement the AUC-based analyses developed here.

We demonstrated the utility of the RPF framework by way of example, showing how the RPF approach can facilitate quantifying precisely how a manipulation of interest impacts subjective processing independent of (or over and above effect on) objective processing. In this case study on the metaperceptual RPF, we found that our dot density manipulation led to changes in mean confidence and also changes in metacognitive sensitivity (meta-*d'*) that were separable from the influence of this manipulation on task performance in this two-alternative forced-choice task.

Although these empirical results are intriguing, our primary excitement lies in the promise of the RPF framework to study the quantitative relationship between any pair of psychological variables the researcher may desire. Thus, we emphasize that the RPF framework can be used not only to study the relationship between objective processing capacity and subjective experience, but for characterizing the quantitative relationship among any two (likely nonlinearly) related psychological processes – including those for which no functional form relating each process to objective stimulus properties is known or presumed (see **Supplemental Material** for details). This is why we have developed the RPF toolbox as an open-source community resource, available for download and extension from https://github.com/CNClaboratory/RPF. We hope that these tools will be of great utility to the community.

Advantages of the RPF method over standard to performance-matching for the study of subjective experience

A primary use for RPF analysis for isolating the neural or computational correlates of subjective aspects of perception from those giving rise to task performance. Since Lau first articulated this need (Lau, 2008; Lau and Passingham, 2006), many groups have sought to control for 'performance confounds' by finding one or two levels of matched performance across various experimental manipulations, and then examining how subjective measures differ (Morales et al., 2022; Peters et al., 2017b). However, as introduced in the introduction, this performance matching approach is unsatisfactory for several major reasons: it relies on a statistical null effect (finding conditions where subjective experience differs but performance *fails* to differ), and it is likely that differences in subjective experience depend on the level of matched performance selected by the experimenter.

As we have seen, RPF analysis circumvents these challenges by revealing differences in P_2 (e.g., confidence) over an entire matched interval of P_1 (e.g., performance). Importantly, however, we can also relate components of RPF analysis directly back to more traditional performance-matching approaches to facilitate direct comparison with existing literature. For example, we can see that if one measures the entire RPF for each condition of interest, RPF AUC analyses can be tuned to any intervals within the available common P_1 interval across conditions of interest. In the limit as this interval approaches zero, computing RPF AUC reduces to "reading off" the P_2 (subjective) values given a particular P_1 (performance) value, i.e. selecting *exactly* the matched-performance level desired through relying on the fitted functions. Doing so avoids the methodological and statistical disadvantages of using staircasing or other methods to discover conditions where subjective measures vary but performance measures *fail* to vary. The freedom to select one or two levels of exactly matched performance also evokes performance-matching studies which have used two or more levels of performance-matched conditions (e.g., hard and easy, (Koizumi et al., 2015); see (Rahnev et al., 2020) for other potential datasets). RPF AUC analyses could be used to reexamine such data using RPF AUC analyses, potentially providing a more principled analytic approach; this might also be possible through the interpolation-based nonparametric approach (described in more detail in **Supplemental Material**) even if fitting a parametric RPF is not possible. Thus, RPF analysis

provides a natural extension to more traditional performance-matching approaches in a way that facilitates direct comparison to previous empirical and theoretical literature.

Relationship to other recent work linking relative and absolute judgments

The study of psychophysics has a long and clever history, spanning 150 years of quantitative psychological research. A large literature has developed documenting the relationship between small changes in physical stimulus magnitude and either humans' (or non-human animals') ability to discriminate or detect such differences, as well as the relationship between physical stimulus magnitude and absolute stimulus magnitude judgments – even of a subjective nature (brightness, loudness, painfulness, and so on). Weber's law, Fechner's law, Stevens' power law – these are all well-known, foundational examples that collectively support quantitative psychology across nearly countless domains of study.

Recently, a unifying framework linking such relative and absolute psychometric judgments – i.e. the relationship between questions such as "Was the left light brighter than the right one?" versus "How bright is this light?" – was proposed by Zhou and colleagues (2024). In this work, the authors combined generalizations of work by Fechner and classic signal detection theory to show how internal noise properties that accompany stimulus representation can explain so-called "power law" intensity percepts. This unifying framework thus elegantly links both relative and absolute psychophysical judgments to stimulus properties in the environment.

Here we have gone one step beyond such a framework to discover how to relate *any* two psychological processes – not just relative and absolute intensity judgments. Specifically, we have through our case study focused on the "metaperceptual" RPF. This form of the RPF is thus not limited to judgments about the subjective evaluation of *stimuli in the world* (absolute magnitude estimation judgments), but is also capable of handling *introspective* or *metacognitive* judgments. In other words, the metaperceptual RPF is sufficiently general so as to evaluate the relationship between the world and first-order internal representations of the world, and between those first-order internal representations and higher-order metacognitive or introspective evaluation of them (Brown et al., 2019; Overgaard and Mogensen, 2017). Thus, the metaperceptual RPF directly addresses recent calls for a psychophysical introspective research program (Fleming, 2023; Kammerer and Frankish, 2023) as a targeted technique for understanding phenomenological experience in general (Peters, 2022), building upon previous research programs seeking to isolate subjective experience for scientific study by holding performance constant (Lau, 2008; Lau and Passingham, 2006; Morales et al., 2022; Peters et al., 2017b). We expect that other field- and question-specific variants of the RPF will emerge, e.g. relating confidence judgments to reaction times, clarity assessments to criterion bias, or even extension to triads of variables or more.

Final thoughts

In sum, the RPF framework holds great promise as a foundation for the next generation of psychophysics. To facilitate the exploration and use of this framework across disciplines and

psychological areas of study, we encourage interested readers to make use of and extend the open-source RPF toolbox (https://github.com/CNClaboratory/RPF).

Acknowledgements

We thank Karen Tian, Michael Epstein, Angela Shen, and Emil Olsson for helpful discussions in the development of this methodology. We also acknowledge support from the Templeton World Charity Foundation (“An Adversarial Collaboration to Test Predictions of First-Order and Higher-Order Theories of Consciousness”, TWCF Number: 0567, to M.A.K.P. and R.D.) and the Canadian Institute for Advanced Research (Fellowship in the Brain, Mind, & Consciousness Program, to M.A.K.P.). The funders had no role in the design or execution of this project.

References

Adler WT, Ma WJ. 2018a. Limitations of Proposed Signatures of Bayesian Confidence. *Neural Comput* 1–28.

Adler WT, Ma WJ. 2018b. Comparing Bayesian and non-Bayesian accounts of human confidence reports. *PLoS Comput Biol* **14**:e1006572.

Aitchison L, Bang D, Bahrami B, Latham PE. 2015. Doubly Bayesian Analysis of Confidence in Perceptual Decision-Making. *PLoS Comput Biol* **11**:e1004519–e1004519.

Baranski JV, Petrusic WM. 1994. The calibration and resolution of confidence in perceptual judgments. *Percept Psychophys* **55**:412–428.

Brown GS, White KG. 2005. The optimal correction for estimating extreme discriminability. *Behav Res Methods* **37**:436–449.

Brown R, Lau H, LeDoux JE. 2019. Understanding the Higher-Order Approach to Consciousness. *Trends Cogn Sci* **23**:754–768.

Burnham KP, Anderson D. 2002. Model selection and multi-model inference: A Practical information-theoretic approach. Berlin; Heidelberg, Germany; New York: Springer.

Ceja V, Ezzeldine Y, Peters MAK. 2022. Models of confidence to facilitate engaging task designsCognitive Computational Neuroscience. doi: 10.32470/CCN.2022.1150-0

Clarke F, Birdsall T, Tanner WP. 1959. Two types of ROC curves and definitions of parameters. *J Acoust Soc Am* **31**:629–630.

Denison RN, Adler WT, Carrasco M, Ma WJ. 2018. Humans incorporate attention-dependent uncertainty into perceptual decisions and confidence. *Proc Natl Acad Sci U S A* **115**:11090–11095.

de Rooij M, Weeda W. 2020. Cross-validation: A method every psychologist should know. *Adv Methods Pract Psychol Sci* **3**:248–263.

Fechner GT. 1860. Elemente der psychophysik. Leipzig: Breitkopf und Härtel.

Fechner GT, Howes DH, Boring EG. 1966. Elements of psychophysics. Holt, Rinehart and Winston New York.

Fleming SM. 2023. Metacognitive Psychophysics in Humans, Animals, and AI. *J Conscious Stud* **30**:113–128.

Fleming SM. 2017. HMeta-d: hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neurosci Conscious* **2017**:nix007.

Galvin SJ, Podd JV, Drga V, Whitmore J. 2003. Type 2 tasks in the theory of signal detectability: discrimination between correct and incorrect decisions. *Psychon Bull Rev* **10**:843–876.

Hausman JA, Newey WK, Powell JL. 1995. Nonlinear errors in variables Estimation of some Engel curves. *J Econom* **65**:205–233.

Hautus MJ. 1995. Corrections for extreme proportions and their biasing effects on estimated values ofd′. *Behav Res Methods Instrum Comput* **27**:46–51.

Huang Z, Meng S, Ye Z. 2023. Effective estimation of nonlinear errors-in-variables models. *Commun Stat Simul Comput* 1–19.

Kammerer F, Frankish K. 2023. What Forms Could Introspective Systems Take? A Research Programme. *Journal of Consciousness Studies* **30**:13–48.

Kiani R, Corthell L, Shadlen MN. 2014. Choice certainty is informed by both evidence and decision time. *Neuron* **84**:1329–1342.

Kingdom FAA, Prins N. 2016. Psychophysics: A Practical Introduction, 2nd ed. San Diego, CA: Academic Press.

Koizumi A, Maniscalco B, Lau H. 2015. Does perceptual confidence facilitate cognitive control? *Atten Percept Psychophys* **77**:1295–1306.

Lau H. 2008. Are We Studying Consciousness Yet? In: Weiskrantz L, Davies M, editors. Frontiers of Consciousness. Oxford University Press. pp. 2008–2245.

Lau H, Passingham RE. 2006. Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proceedings of the National Academy of Sciences* **103**:18763–18768.
Li T. 2002. Robust and consistent estimation of nonlinear errors-in-variables models. *J Econom* **110**:1–26.
Macmillan NA, Creelman CD. 2004. Detection Theory: A User's Guide. Taylor & Francis.
Maniscalco B, Charles L, Peters MAK. 2024. Optimal Metacognitive Decision Strategies in Signal Detection Theory. *Psychon Bull Rev*.
Maniscalco B, Lau H. 2014. Signal Detection Theory Analysis of Type 1 and Type 2 Data: Meta-d′, Response-Specific Meta-d′, and the Unequal Variance SDT Model In: Fleming SM, Frith CD, editors. The Cognitive Neuroscience of Metacognition. Berlin, Heidelberg: Springer Berlin Heidelberg. pp. 25–66.
Maniscalco B, Lau H. 2012. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Conscious Cogn* **21**:422–430.
Maniscalco B, Odegaard B, Grimaldi P, Cho SH, Basso MA, Lau H, Peters MAK. 2021. Tuned normalization in perceptual decision-making circuits can explain seemingly suboptimal confidence behavior. *PLoS Comput Biol* **17**:e1008779.
Maniscalco B, Peters MAK, Lau H. 2016. Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. *Atten Percept Psychophys*. doi:10.3758/s13414-016-1059-x
Michel M, Beck D, Block N, Blumenfeld H, Brown R, Carmel D, Carrasco M, Chirimuuta M, Chun M, Cleeremans A, Dehaene S, Fleming SM, Frith C, Haggard P, He BJ, Heyes C, Goodale MA, Irvine L, Kawato M, Kentridge R, King J-R, Knight RT, Kouider S, Lamme V, Lamy D, Lau H, Laureys S, LeDoux J, Lin Y-T, Liu K, Macknik SL, Martinez-Conde S, Mashour GA, Melloni L, Miracchi L, Mylopoulos M, Naccache L, Owen AM, Passingham RE, Pessoa L, Peters MAK, Rahnev D, Ro T, Rosenthal DM, Sasaki Y, Sergent C, Solovey G, Schiff ND, Seth A, Tallon-Baudry C, Tamietto M, Tong F, van Gaal S, Vlassova A, Watanabe T, Weisberg J, Yan K, Yoshida M. 2019. Opportunities and challenges for a maturing science of consciousness. *Nat Hum Behav* **3**:104–107.
Miller J. 1996. The sampling distribution ofd′. *Percept Psychophys* **58**:65–72.
Miyoshi K, Lau H. 2020. A decision-congruent heuristic gives superior metacognitive sensitivity under realistic variance assumptions. *Psychol Rev* **127**:655–671.
Morales J, Odegaard B, Maniscalco B. 2022. The Neural Substrates of Conscious Perception without Performance Confounds In: De Brigard & W. Sinnott-Armstrong F, editor. Neuroscience and Philosophy. MIT Press. pp. 285–323.
Odegaard B, Chang MY, Lau H, Cheung S-H. 2018a. Inflation versus filling-in: why we feel we see more than we actually do in peripheral vision. *Philos Trans R Soc Lond B Biol Sci* **373**. doi:10.1098/rstb.2017.0345
Odegaard B, Grimaldi P, Cho SH, Peters MAK, Lau H, Basso MA. 2018b. Superior colliculus neuronal ensemble activity signals optimal rather than subjective confidence. *Proceedings of the National Academy of Sciences* **115**:E1588–E1597.
Overgaard M, Mogensen J. 2017. An integrative view on consciousness and introspection. *Rev Philos Psychol* **8**:129–141.
Peters MAK. 2022. Towards Characterizing the Canonical Computations Generating Phenomenal Experience. *Neurosci Biobehav Rev*.
Peters MAK, Fesi J, Amendi N, Knotts JD, Lau H, Ro T. 2017a. Transcranial magnetic stimulation to visual cortex induces suboptimal introspection. *Cortex* **93**:119–132.
Peters MAK, Kentridge RW, Phillips I, Block N. 2017b. Does unconscious perception really exist? Continuing the ASSC20 debate. *Neurosci Conscious* **2017**. doi:10.1093/nc/nix015
Peters MAK, Lau H. 2015. Human observers have optimal introspective access to perceptual processes even for visually masked stimuli. *Elife* 10.7554/eLife.09651.

Peters MAK, Thesen T, Ko YD, Maniscalco B, Carlson C, Davidson M, Doyle W, Kuzniecky R, Devinsky O, Halgren E, Lau H. 2017c. Perceptual confidence neglects decision-incongruent evidence in the brain. *Nature Human Behaviour*.

Pollack I, Hsieh R. 1969. Sampling variability of the area under the ROC-curve and of d'e. *Psychol Bull* **71**:161–173.

Prins N, Kingdom FAA. 2018. Applying the Model-Comparison Approach to Test Specific Research Hypotheses in Psychophysical Research Using the Palamedes Toolbox. *Front Psychol* **9**:1250.

Rahnev D, Balsdon T, Charles L, de Gardelle V, Denison R, Desender K, Faivre N, Filevich E, Fleming SM, Jehee J, Lau H, Lee ALF, Locke SM, Mamassian P, Odegaard B, Peters M, Reyes G, Rouault M, Sackur J, Samaha J, Sergent C, Sherman MT, Siedlecka M, Soto D, Vlassova A, Zylberberg A. 2022. Consensus Goals in the Field of Visual Metacognition. *Perspect Psychol Sci* 17456916221075615.

Rahnev D, Desender K, Lee ALF, Adler WT, Aguilar-Lleyda D, Akdoğan B, Arbuzova P, Atlas LY, Balcı F, Bang JW, Bègue I, Birney DP, Brady TF, Calder-Travis J, Chetverikov A, Clark TK, Davranche K, Denison RN, Dildine TC, Double KS, Duyan YA, Faivre N, Fallow K, Filevich E, Gajdos T, Gallagher RM, de Gardelle V, Gherman S, Haddara N, Hainguerlot M, Hsu T-Y, Hu X, Iturrate I, Jaquiery M, Kantner J, Koculak M, Konishi M, Koß C, Kvam PD, Kwok SC, Lebreton M, Lempert KM, Ming Lo C, Luo L, Maniscalco B, Martin A, Massoni S, Matthews J, Mazancieux A, Merfeld DM, O'Hora D, Palser ER, Paulewicz B, Pereira M, Peters C, Philiastides MG, Pfuhl G, Prieto F, Rausch M, Recht S, Reyes G, Rouault M, Sackur J, Sadeghi S, Samaha J, Seow TXF, Shekhar M, Sherman MT, Siedlecka M, Skóra Z, Song C, Soto D, Sun S, van Boxtel JJA, Wang S, Weidemann CT, Weindel G, Wierzchoń M, Xu X, Ye Q, Yeon J, Zou F, Zylberberg A. 2020. The Confidence Database. *Nat Hum Behav* **4**:317–325.

Rahnev D, Maniscalco B, Graves T, Huang E, de Lange FP, Lau H. 2011. Attention induces conservative subjective biases in visual perception. *Nat Neurosci* **14**:1513–1515.

Rahnev D, Maniscalco B, Luber B, Lau H, Lisanby SH. 2012. Direct injection of noise to the visual cortex decreases accuracy but increases decision confidence. *J Neurophysiol* **107**:1556–1563.

Rausch M, Hellmann S, Zehetleitner M. 2017. Confidence in masked orientation judgments is informed by both evidence and visibility. *Atten Percept Psychophys*. doi:10.3758/s13414-017-1431-5

Rollwage M, Loosen A, Hauser TU, Moran R, Dolan RJ, Fleming SM. 2020. Confidence drives a neural confirmation bias. *Nat Commun* **11**:2634.

Rouault M, Seow T, Gillan CM, Fleming SM. 2018. Psychiatric symptom dimensions are associated with dissociable shifts in metacognition but not task performance. *Biol Psychiatry*. doi:10.1016/j.biopsych.2017.12.017

Rounis E, Maniscalco B, Rothwell JC, Passingham RE, Lau H. 2010. Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cogn Neurosci* **1**:165–175.

Ruby E, Maniscalco B, Peters MAK. 2018. On a "failed"attempt to manipulate visual metacognition with transcranial magnetic stimulation to prefrontal cortex. *Conscious Cogn*.

Samaha J, Barrett JJ, Sheldon AD, LaRocque JJ, Postle BR. 2016. Dissociating Perceptual Confidence from Discrimination Accuracy Reveals No Influence of Metacognitive Awareness on Working Memory. *Front Psychol* **7**:851.

Samaha J, Denison R. 2022. The positive evidence bias in perceptual confidence is unlikely post-decisional. *Neurosci Conscious* **2022**:niac010.

Samaha J, Switzky M, Postle BR. 2019. Confidence boosts serial dependence in orientation estimation. *J Vis* 590.

Stolyarova A, Rakhshan M, Hart EE, O'Dell TJ, Peters MAK, Lau H, Soltani A, Izquierdo A.

2019. Contributions of anterior cingulate cortex and basolateral amygdala to decision confidence and learning under uncertainty. *Nat Commun* **10**:4704.
Vrieze SI. 2012. Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychol Methods* **17**:228–243.
Winter CJ, Peters MAK. 2022. Variance misperception under skewed empirical noise statistics explains overconfidence in the visual periphery. *Atten Percept Psychophys* **84**:161–178.
Wolter KM, Fuller WA. 1982. Estimation of nonlinear errors-in-variables models. *Ann Stat* **10**:539–548.
Zhou J, Duong LR, Simoncelli EP. 2024. A unified framework for perceived magnitude and discriminability of sensory stimuli. *Proc Natl Acad Sci U S A* **121**:e2312293121.
Zylberberg A, Barttfeld P, Sigman M. 2012. The construction of confidence in a perceptual decision. *Front Integr Neurosci* **6**:79–79.
Zylberberg A, Fetsch CR, Shadlen MN, Frank MJ. 2016. The influence of evidence volatility on choice, reaction time and confidence in a perceptual decision. *eLife* **5**:e17688.
Zylberberg A, Roelfsema PR, Sigman M. 2014. Variance misperception explains illusions of confidence in simple perceptual decisions. *Conscious Cogn* **27C**:246–253.

Supplemental Material

Supplementary information for fitting and assessing the RPF

Maximum likelihood estimation (MLE) fitting of non-probabilistic psychometric functions

When a psychometric function ψ(x; θ) is fitted to response probabilities (e.g. p(correct)), the likelihood of the data under ψ is derived by treating trial-level outcomes as Bernoulli trials. This entails assuming that for each level of stimulus strength *x*, the probability of a "success" on a given trial (e.g. a report of stimulus detection, or a correct discrimination response) is constant and independent of outcomes on other trials. Under these assumptions, the joint probability of all trial outcomes is the product of the probability of each individual trial outcome. Thus, if the probabilities of a trial outcome t being 0 or 1 at stimulus strength x are given by

$$p_{\theta\, t=1,x} = \psi(x;\theta)$$
$$p_{\theta\, t=0,x} = 1 - \psi(x;\theta) \quad \text{(S1)}$$

then the likelihood of all trial outcomes $\boldsymbol{t}$ according to ψ(x;θ) is given by

$$L(\theta \mid \boldsymbol{t}) = \prod_{t,x} p_{\theta\, t,x}^{\;n_{t,x}} \quad \text{(S2)}$$

where $n_{t,x}$ is the number of occurrences of trial outcome *t* for stimulus strength *x*. Using the more convenient log likelihood,

$$\log L(\theta \mid \mathbf{t}) = \sum_{t,x} n_{t,x} \log p_{\theta\, t,x} \quad \text{(S3)}$$

The MLE estimate of θ is then the value of θ that maximizes likelihood (or equivalently, log likelihood).

However, since this approach to MLE fitting assumes a probabilistic psychometric function, it cannot be applied to non-probabilistic psychometric functions fitted to variables such as *d'*, which raises the question of how fitting should proceed in such cases.

A simple approach would be to minimize the sum of the squared errors of the fit. When the errors of the fit can be assumed to be normally distributed with constant variance, there is a direct relationship between SSE and log likelihood given by

$$\log L(\theta) = -\frac{1}{2} n \log\left(\frac{SSE_\theta}{n}\right) \quad \text{(S4)}$$

and the parameters θ that minimize SSE also maximize likelihood (Burnham and Anderson, 2002).

However, such assumptions may not always hold. For instance, while the sampling distribution of *d'* is approximately normal, its variance is not constant (Macmillan and Creelman, 2004; Miller, 1996). For a fixed, unbiased criterion, the sampling variance of *d'* increases as the true value of *d'* increases. Additionally, for a fixed value of true *d'*, sampling variance depends on the true hit rate and false alarm rate, with values closer to 0 or 1 for either variable leading to higher variance in the estimated *d'*. For cases such as this, SSE cannot be used to compute likelihood, and minimizing SSE will not give a maximum likelihood estimate. Not being able to compute likelihood also hinders the ability of the fit to be assessed in conventional model comparison analyses, which require knowledge of the fit's likelihood (Burnham and Anderson, 2002).

Thus, where possible, it is always preferable to have an expression for the likelihood of the data given the model which makes minimal assumptions about the data being fitted. The likelihood function can then be used as a basis for MLE fitting.

Below we derive methods for MLE fitting of psychometric functions to three variables of interest: *d'*, meta-*d'*, and mean rating. In each of these cases, we use likelihood functions that assign probabilities to trial-level data using an appropriate model, which only requires making the standard assumption that trial outcome probabilities are independent across trials. All of these methods are implemented in the RPF toolbox (https://github.com/CNClaboratory/RPF).

Scaled psychometric functions

In considering how to approach fitting a psychometric function to non-probabilistic dependent variables, we first note that the function to be fitted cannot be a function that ranges from 0 to 1 to model response probabilities, e.g. as in the Weibull function given in Eq. 1 of the main manuscript. In the general case, the dependent variable to be fitted cannot be assumed to have an upper bound on its possible values, and so the fitted psychometric function cannot assume an *a priori* maximum value.

However, just as in the modeling of response probabilities we use the lapse rate parameter λ to allow for the possibility that asymptotic performance may not reach the maximal value of 1 even for maximal or arbitrarily large values of *x*, similarly it may be reasonable in certain cases to posit a practical upper bound for a non-probabilistic dependent variable to which it asymptotes

as *x* increases. This asymptotic upper bound can be captured using a parameter of the fitted psychometric function. To differentiate this parameter for asymptotic performance from the probabilistic lapse rate λ, we here give it the general name ω.

Thus, a simple way to adapt any probabilistic psychometric function ψ(x; α, β, γ, λ) to fit a non-probabilistic dependent variable is to adapt its formula so that it ranges over [γ, ω] rather than [γ, 1-λ]. For instance, for the probabilistic Weibull function

$$W(x;\alpha,\beta,\gamma,\lambda) = \gamma + (1 - \lambda - \gamma)\left[1 - e^{-\left(\frac{x}{\alpha}\right)^{\beta}}\right] \tag{S5}$$

we can rewrite this as a *scaled* Weibull function

$$W_S(x;\alpha,\beta,\gamma,\omega) = \gamma + (\omega - \gamma)\left[1 - e^{-\left(\frac{x}{\alpha}\right)^{\beta}}\right] \tag{S6}$$

Similar approaches can be used to adapt any probabilistic psychometric function. We use the general term *scaled psychometric functions* to refer to functions that have been adapted to apply to non-probabilistic dependent variables by virtue of ranging from γ to ω.

d'

In line with the above discussion, here we consider the question of how to perform an MLE fit of a scaled psychometric function ψ_S to a set of *d'* data over a set of *x* values.

Using the upper bound on computed *d'* to inform ω

The classical signal detection theory (SDT) model (Macmillan and Creelman, 2004) models a task in which an observer is repeatedly presented with stimuli from two stimulus classes denoted S1 and S2, and on each trial must categorize the presented stimulus by responding "S1" or "S2". The SDT measure of sensitivity *d'* measures the signal-to-noise ratio of perceptual evidence occurring under presentations of S1 and S2, and is computed from empirical hit rate (HR) and false alarm rate (FAR) data in accordance with the following equations:

$$\begin{aligned} \text{HR} &= p(\text{response} = \text{"S2"} \mid \text{stimulus} = \text{S2}) = \frac{n(\text{response="S2"} \cap \text{stimulus=S2})}{n(\text{stimulus=S2})} \\ \text{FAR} &= p(\text{response} = \text{"S2"} \mid \text{stimulus} = \text{S1}) = \frac{n(\text{response="S2"} \cap \text{stimulus=S1})}{n(\text{stimulus=S1})} \\ d' &= z(\text{HR}) - z(\text{FAR}) \end{aligned} \tag{S7}$$

where *n* denotes a trial count function returning the number of trials for which the specified condition is true, and *z* is the inverse of the standard normal CDF. Hereafter, we use “resp” and “stim” to abbreviate occurrences of “response” and “stimulus” in equations.

If either of the empirical HR or FAR equals 0 or 1, then *d’* computed in the above way is infinite. Typically, such cases can be considered to be an artifact of noisy data for which the “true” HR and FAR are both greater than 0 and less than 1, rather than a veridical measurement of infinite *d’*. To circumvent this numerical issue, it is standard practice to use an adjustment to the computed HR and FAR to ensure a finite value for estimated *d’* (Macmillan and Creelman, 2004). For instance, one approach is to add 0.5 to every response count cell before computing *d’* (Brown and White, 2005; Hautus, 1995), which can be thought of as adding one “dummy” trial for each of S1 and S2, treating each of these as counting halfway towards “S1” and “S2” responses. More formally,

$$\begin{aligned} \mathrm{HR}_{\mathrm{adj}} &= \frac{n(\text{resp="S2"} \cap \text{stim=S2}) + 0.5}{n(\text{stim=S2}) + 1} \\ \mathrm{FAR}_{\mathrm{adj}} &= \frac{n(\text{resp="S2"} \cap \text{stim=S1}) + 0.5}{n(\text{stim=S1}) + 1} \\ d'_{\mathrm{adj}} &= z\left(\mathrm{HR}_{\mathrm{adj}}\right) - z\left(\mathrm{FAR}_{\mathrm{adj}}\right) \end{aligned} \tag{S8}$$

This adjustment scheme has the effect of imposing a maximum value for computed *d’*, which occurs in the case where $\mathrm{HR}_{\mathrm{adj}}$ and $\mathrm{FAR}_{\mathrm{adj}}$ take on their maximum and minimum possible values, respectively:

$$\begin{aligned} \mathrm{HR}_{\mathrm{adj\ max}} &= \frac{n(\text{stim=S2}) + 0.5}{n(\text{stim=S2}) + 1} \\ \mathrm{FAR}_{\mathrm{adj\ min}} &= \frac{0.5}{n(\text{stim=S1}) + 1} \\ d'_{\mathrm{adj\ max}} &= z\left(\mathrm{HR}_{\mathrm{adj\ max}}\right) - z\left(\mathrm{FAR}_{\mathrm{adj\ min}}\right) \end{aligned} \tag{S9}$$

Thus, when fitting d'_{adj} data with a scaled psychometric function, $d'_{\mathrm{adj\ max}}$ presents a natural upper bound for the value of ω. It may be desirable to constrain ω to equal $d'_{\mathrm{adj\ max}}$ in certain cases where it can reasonably be assumed that asymptotic performance is very high, e.g. to improve the stability of psychometric function fits for data with low trial counts. More complex approaches might involve making assumptions about how factors like lapse rate and asymptotic criterion determine alternative values for $\mathrm{HR}_{\mathrm{adj\ max}}$ and $\mathrm{FAR}_{\mathrm{adj\ min}}$, and thus arrive at an alternative value for $d'_{\mathrm{adj\ max}}$.

MLE fitting of a psychometric function to *d'*

The general idea behind the following is to use fitted *d'*, in conjunction with empirical values for the criterion *c*, to compute a "fitted" HR and FAR at each level of *x*. These can then be used to assign probabilities to single trial outcomes, which in turn can be used to compute likelihood in the general way described above. This approach is conceptually similar to the MLE estimation of meta-*d'*, in which fitted meta-*d'* is used in conjunction with the empirical *c'* to assign single-trial probabilities to type 2 outcomes (Maniscalco and Lau, 2014, 2012).

A psychometric function ψ with parameters θ fitted to *d'* data gives fitted values at each level of *x* via

$$\widehat{d'}_x = \psi(x;\, \theta) \tag{S10}$$

We may consider these as being related to "fitted" HRs and FARs via

$$\widehat{d'}_x = z\left(\widehat{\mathrm{HR}}_x\right) - z\left(\widehat{\mathrm{FAR}}_x\right) \tag{S11}$$

but this does not provide sufficient information for computing unique values for $\widehat{\mathrm{HR}}_x$ and $\widehat{\mathrm{FAR}}_x$. However, we can make progress by taking the empirically computed criteria c_x as givens, where

$$c_x = -\frac{1}{2}\left(z\left(\mathrm{HR}_x\right) + z\left(\mathrm{FAR}_x\right)\right) \tag{S12}$$

Using the empirical d'_x and c_x, one can solve for the exact values of HR_x and FAR_x. If one instead used the *estimated* $\widehat{d'}_x$ and the empirical c_x in this calculation, one would compute *estimated* values of $\widehat{\mathrm{HR}}_x$ and $\widehat{\mathrm{FAR}}_x$ whose degree of error depends on the error in $\widehat{d'}_x$. These estimated $\widehat{\mathrm{HR}}_x$ and $\widehat{\mathrm{FAR}}_x$ are given by

$$\begin{aligned} \widehat{\mathrm{HR}}_x &= \Phi\left(\frac{\widehat{d'}_x}{2} - c_x\right) \\ \widehat{\mathrm{FAR}}_x &= \Phi\left(-\frac{\widehat{d'}_x}{2} - c_x\right) \end{aligned} \tag{S13}$$

where Φ is the standard normal CDF. These imply estimated miss rates ($\widehat{\mathrm{MR}}_x$) and correct rejection rates ($\widehat{\mathrm{CRR}}_x$):

$$\begin{aligned} \widehat{\mathrm{MR}}_x &= 1 - \widehat{\mathrm{HR}}_x \\ \widehat{\mathrm{CRR}}_x &= 1 - \widehat{\mathrm{FAR}}_x \end{aligned} \tag{S14}$$

For convenience of notation, we can re-express these as probabilities for every kind of stimulus classification outcome:

$$\begin{aligned} p_{\theta\, r=\text{"}S2\text{"}|s=S2,x} &= \widehat{\mathrm{HR}}_x \\ p_{\theta\, r=\text{"S2"}|s=\mathrm{S1},x} &= \widehat{\mathrm{FAR}}_x \\ p_{\theta\, r=\text{"S1"}|s=\mathrm{S2},x} &= \widehat{\mathrm{MR}}_x \\ p_{\theta\, r=\text{"S1"}|s=\mathrm{S1},\ x} &= \widehat{\mathrm{CRR}}_x \end{aligned} \tag{S15}$$

and then write the likelihood function as

$$\log L(\theta \mid \text{data}) \;=\; \sum_{r,s,x} n_{r|s,x} \log p_{\theta\, r|s,x} \tag{S16}$$

where $n_{r|s,x}$ is shorthand for the trial count $n(\text{resp} = r \mid \text{stim} = s \cap \text{strength} = x)$, *r* can take on values “S1” and “S2”, and *s* can take on values S1 and S2.

Thus, by considering estimated *d’* alongside empirical *c*, we are able to assign probabilities to each trial outcome and use those trial outcome probabilities to compute likelihood in the usual way. We can interpret the MLE fit derived from this approach as telling us what psychometric function parameters θ, taken in conjunction with the empirical values of c_x, maximize the likelihood of the subject’s stimulus classification responses for each stimulus type.

An alternative way of thinking about this approach is that the full parameter set for the model is

$$\theta = \left\{\alpha, \beta, \gamma, \omega, \widehat{\boldsymbol{c}}\right\} \tag{S17}$$

where $\alpha, \beta, \gamma, \omega$ are parameters of the scaled psychometric function determining each $\widehat{d'}_x$, and $\widehat{\boldsymbol{c}}$ is a vector of modeled $\widehat{c}_x$ values. This parameter set enables calculation of each $p_{\theta\, r|s,x}$. However, since we are only interested in estimating the psychometric function of *d'*, we constrain all elements of $\widehat{\boldsymbol{c}}$ to be equal to their corresponding empirical c_x values, leaving only $\alpha, \beta, \gamma, \omega$ as free parameters in the model fit. The ML estimate of the full model is then given by solving

$$\theta^* = \arg\max_{\theta} L(\theta \mid \text{data}), \;\; \text{subject to: } \forall x \; \widehat{c}_x = c_x \tag{S18}$$

and taking $\alpha, \beta, \gamma, \omega \in \theta^*$ to be the ML parameter estimates of the psychometric function for *d'*. Another approach to fitting *d'* might involve directly fitting the HR_x and FAR_x data using the usual MLE fitting method for probabilistic variables, and using these fitted curves to compute a fitted *d'* curve using the standard equation for *d'*. However, in our experience this method is ineffective and gives very unstable and problematic fits to the *d'* data, necessitating the alternative approach described above.

meta-d'

Let us first consider the approach to MLE fitting of meta-*d'* for a single set of data (Maniscalco and Lau, 2014, 2012), and then adapt this approach for use with fitting a psychometric function.

MLE fitting of meta-*d'*

We characterize the likelihood of type 2 responses (e.g. confidence ratings) conditional on type 1 outcomes (correct and incorrect "S1" and "S2" responses) as

$$\log L(\theta \mid \text{data}) = \sum_{y,s,r} n_{y|s,r} \log p_{\theta\, y|s,r} \tag{S19}$$

where $n_{y|s,r}$ is shorthand for the trial count $n(\text{conf} = y \mid \text{stim} = s \cap \text{resp} = r)$, *y* can take on values in {1, 2, …, N_{ratings}} where N_{ratings} is the number of available ratings on the confidence rating scale, *r* can take on values "S1" and "S2", and *s* can take on values S1 and S2.

$p_{\theta\, y|s,r}$ are type 2 response probabilities according to some model parameterized with θ, where the subscript notation is the same as that of $n_{y|s,r}$. These probabilities correspond to points on the type 2 ROC curves for "S1" and "S2" responses and thus characterize type 2 sensitivity. In the meta-*d'* model, the type 2 probabilities $p_{\theta\, y|s,r}$ are characterized in terms of a type 1 SDT model whose parameters are

$$\theta = \left\{\text{meta}-d', \text{meta}-c, \text{meta}-\boldsymbol{c}_{2,\text{"S1"}}, \text{meta}-\boldsymbol{c}_{2,\text{"S2"}}\right\} \tag{S20}$$

where the "meta-" prefix emphasizes that these type 1 SDT model parameters are used to characterize type 2 probabilities. Meta-*d'* and meta-*c* correspond to *d'* and *c* in the standard SDT model, and $\text{meta}-\boldsymbol{c}_{2,\text{"S1"}}$ and $\text{meta}-\boldsymbol{c}_{2,\text{"S2"}}$ are vectors of type 2 criteria for producing type 2 responses, each of length $N_{\text{ratings}} - 1$. Taken together, these parameters determine all $p_{\theta\, y|s,r}$ via standard probability calculations using the SDT model.

In fitting this model to the data, we constrain meta-*c* such that it yields a relative criterion in the meta-*d'* model equal to the empirical *c'* computed from the data. The relative criterion is defined as $c' = c/d'$, and so this constraint amounts to setting $\text{meta}-c = c' \cdot \text{meta}-d'$. We also constrain the type 1 and type 2 criteria so that they stand in the appropriate ordinal relationships to each other on the SDT decision axis, as summarized by a Boolean function $C(\theta)$ which returns 1 if the criteria are in appropriate ordinal relationships and 0 otherwise.

Fitting the meta-*d'* model thus consists in solving the optimization problem

$$\theta^{*} = \arg\max_{\theta} L(\theta \mid \text{data}), \text{ subject to: } \text{meta}-c' = c',\ C(\theta) = 1 \tag{S21}$$

and taking meta-*d'* ϵ θ* to be the ML estimate of meta-*d'*.

Thus, the meta-*d'* model characterizes the type 2 sensitivity exhibited by a set of type 2 data in terms of what *d'* value in a standard SDT model would maximize the likelihood of those type 2 data, provided that the model's relative criterion *c'* is identical to the *c'* computed from the same data set.

For more details on these methods, please see (Maniscalco and Lau, 2014).

MLE fitting of a psychometric function to meta-*d'*

The approach for MLE fitting of meta-*d'* described above can be straightforwardly adapted to an MLE fit of a psychometric function describing meta-*d'* as a function of stimulus strength *x*. Essentially, this adaptation consists in performing the same fit of the meta-*d'* model to every

level of *x* as described above, with the exception that each $\text{meta}-d'_x$ is determined not via N_x separate meta-*d'* parameters, but rather via the parameters of a fitted psychometric function.

First, we expand the likelihood function to take into account type 2 responses conditional on type 1 outcomes at each level of *x*:

$$\log L(\theta \mid \text{data}) = \sum_{y,s,r,x} n_{y|s,r,x} \log p_{\theta\, y|s,r,x} \tag{S22}$$

where $n_{y|s,r,x}$ is shorthand for the trial count $n(\text{conf} = y \mid \text{stim} = s \cap \text{resp} = r \cap \text{strength} = x)$, and $p_{\theta\, y|s,r,x}$ employs similar notation.

Second, we expand the model such that it is characterized by parameters

$$\theta = \left\{\alpha, \beta, \gamma, \omega, \text{meta}-\boldsymbol{c}, \text{meta}-\boldsymbol{c}_{2,\text{"S1"}}, \text{meta}-\boldsymbol{c}_{2,\text{"S2"}}\right\} \tag{S23}$$

where $\text{meta}-\boldsymbol{c}$ is a vector of length N_x containing values of meta-*c* at every level of *x*, $\text{meta}-\boldsymbol{c}_{2,\text{"S1"}}$ and $\text{meta}-\boldsymbol{c}_{2,\text{"S2"}}$ are matrices of size $(N_x, N_{\text{ratings}} - 1)$ containing type 2 criteria for "S1" and "S2" responses at every level of *x*, and $\alpha, \beta, \gamma, \omega$ are parameters of the psychometric function for meta-*d'* such that

$$\widehat{\text{meta}-d'}_x = \psi(x;\ \alpha, \beta, \gamma, \omega) \tag{S24}$$

At every level of *x* we apply the same constraints as in the standard meta-*d'* model fit, such that $\text{meta}-c'_x = c'_x$ and $C(\theta_x) = 1$ for all *x*, where θ_x indicates the subset of the parameters in θ applying to the specified value of *x*. This latter constraint ensures that within each level of *x*, the criteria of the meta-*d'* model stand in appropriate ordinal relationships.

Fitting the meta-*d'* model for a psychometric function thus consists in solving the optimization problem

$$\theta^* = \arg\max_{\theta} L(\theta \mid \text{data}),\ \text{subject to: } \forall x\ \text{meta}-c'_x = c'_x,\ C(\theta_x) = 1 \tag{S25}$$

and taking $\alpha, \beta, \gamma, \omega \in \theta^*$ to be the ML parameter estimates of the psychometric function for meta-*d'*.

Mean rating

In experiments using a rating scale (e.g. as of confidence, visibility, etc.) with three or more rating options, it may be of interest to fit a psychometric function to the mean rating across levels of stimulus strength *x*. Below we demonstrate that the psychometric function for mean rating can be expressed as a simple sum of psychometric functions fitted to cumulative rating probabilities of the form $p(\text{rating} \geq y)$, each of which can be fitted with standard MLE methods for probabilistic variables.

Expressing mean rating in terms of cumulative rating probabilities

Consider a rating scale consisting of options {1, 2, …, N_{ratings}} where N_{ratings} (or N_R for short) is the number of available ratings on the rating scale, and $N_R \geq 2$. Let *R* be a random variable for the rating observed on any given trial. Then the mean rating across trials is given by

$$\overline{R} = \sum_{y=1}^{N_R} y \cdot P(R = y) \tag{S26}$$

The $P(R = y)$ terms can be expressed in terms of cumulative probabilities via

$$P(R = y) = \begin{cases} 1 - P(R \geq 2), & y = 1 \\ P(R \geq y) - P(R \geq y + 1), & 2 \leq y \leq N_R - 1 \\ P\left(R \geq N_R\right), & y = N_R \end{cases} \tag{S27}$$

Combining Eqs. S26 and S27 and simplifying, we can express mean rating in terms of cumulative probabilities as

$$\overline{R} = 1 - P(R \geq 2) + \left[\sum_{y=2}^{N_R - 1} y \cdot (P(R \geq y) - P(R \geq y + 1))\right] + N_R P\left(R \geq N_R\right)$$

$$\overline{R} = 1 + \sum_{y=2}^{N_R} y \cdot P(R \geq y) - (y - 1) \cdot P(R \geq y)$$

$$\overline{R} = 1 + \sum_{y=2}^{N_R} P(R \geq y) \tag{S28}$$

Thus, mean rating across all trials can be expressed as the sum of the cumulative probabilities $P(R \geq y)$ over all values *y* in the rating scale.

MLE fitting of a psychometric function to mean rating

Now consider an experiment using several levels of stimulus strength *x*. We can compute the probability that ratings reach some threshold value *y* at each level of *x* as $P(R \geq y \mid x)$, and since each individual trial has a Boolean outcome for whether its rating reaches threshold or not, the $P(R \geq y \mid x)$ data can be fitted by a psychometric function with standard MLE methods for probabilistic variables as

$$\widehat{P}_{y,x}(R \geq y \mid x) = \psi\left(x; \theta_y\right) \tag{S29}$$

By the above logic, we may then express a psychometric function for mean rating $\widehat{\overline{R}}_x$ in terms of the sum of fitted psychometric functions for $\widehat{P}_{y,x}$ at each level of $y \geq 2$ as

$$\begin{aligned} \widehat{\overline{R}}_x &= 1 + \sum_{y=2}^{N_R} \widehat{P}_{y,x}(R \geq y \mid x) \\ &= 1 + \sum_{y=2}^{N_R} \psi\left(x; \theta_y\right) \end{aligned} \tag{S30}$$

Because each $\widehat{P}_{y,x}$ is derived via MLE, $\widehat{\overline{R}}_x$ gives an MLE fit to the mean rating data $\overline{R}_x$.

Note that due to the link between Eqs. S26 and S28, the MLE approach in Eq. S30 is equivalent to fitting the rating probabilities $\widehat{P}_{y,x}(R = y \mid x)$ for all but one value of y (since the estimated rating probability for the final y can be inferred from the others). This is intuitive, as an MLE fit to mean rating requires assigning a probability to the rating outcome on each trial, which in turn requires estimation of each $\widehat{P}_{y,x}(R = y \mid x)$. However, for the sake of fitting psychometric functions, it is more convenient to work with Eq. S30 and estimate $\widehat{P}_{y,x}(R \geq y \mid x)$ since these

functions will tend to be monotonically increasing with x, in agreement with the behavior of typical psychometric functions, whereas this is not the case for $\widehat{P}_{y,x}(R = y \mid x)$.

Nonparametric RPF AUC analysis: methodological considerations

As discussed in the main manuscript, there may be cases where it is desirable to estimate RPF AUC nonparametrically. Here we discuss methodological considerations for this approach in more detail.

Full interpolation

If neither P_1 nor P_2 data are fitted with parametric functions, then RPF AUC can be estimated nonparametrically by summing the areas of the trapezoids formed by linear interpolation over the plot of P_2 vs. P_1. The RPF toolbox (https://github.com/CNClaboratory/RPF) performs this interpolation with the `interp1` function of Matlab, which requires the input list of x-values to be unique and sorted in ascending order. Thus, the following preprocessing of the data is conducted:

1. The P_1 data is sorted in ascending order, with the P_2 data subject to the same re-ordering such that all (P_1, P_2) data pairs remain intact.
2. For any P_1 value that occurs more than once, the corresponding ($P_{1,i}$, $P_{2,i}$) pairs are replaced with a single (P'_1, P'_2) pair where P'_1 is the recurring P_1 value and P'_2 is the average of the $P_{2,i}$ values.

Partial interpolation

Alternatively, it is possible to construct an RPF by fitting a psychometric function F_1 to P_1 data and estimating the function F_2 of P_2 data by interpolation. This allows the standard computation of the RPF as $P_2 = F_2(F_1^{-1}(P_1))$, treating the interpolation of the P_2 data as the function F_2. However, it is in general not possible to interpolate P_1 and fit P_2. In this case, the interpolated P_1 function will in general not be monotonic with x, which prevents computation of $F_1^{-1}(P_1)$ since not every P_1 value will map onto a unique x value.

Constraints on P_1 intervals, and possible expansions

Under full interpolation, since interpolation cannot extrapolate beyond the P_1 data, the widest possible P_1 interval over which the RPF can be computed is the one defined by the minimum and maximum values in the P_1 data. Under partial interpolation, although the fitted F_1 function can extrapolate beyond the P_1 data, the interpolated P_2 data are still constrained to range over the minimum and maximum x values used in the experiment (call them $x_{\text{expt min}}$ and $x_{\text{expt max}}$). Thus, the RPF can only be computed over the fitted P_1 interval corresponding to this experimentally constrained x interval, i.e. over the P_1 interval $[F_1(x_{\text{expt min}}), F_1(x_{\text{expt max}})]$.

However, in some cases it may be reasonable to assume theoretical (P_1, P_2) data pairs corresponding to the minimum and/or maximum possible values for the variables in question, which can then be used to expand the P_1 interval beyond the constraints imposed by the data. For instance, if P_1 and P_2 correspond to *d'* and meta-*d'*, it is reasonable to assume that when

stimulus strength is minimal (x = 0), it must also be the case that $P_1 = P_2 = 0$ since both *d'* and meta-*d'* have chance values of 0. In this case, the lower bound of the empirically constrained P_1 interval for the fully interpolated RPF could be extended by appending the theoretical (0, 0) data pair to the empirical (P_1, P_2) data. Similarly, if P_1 and P_2 correspond to p(correct) and p(high rating), theoretical or empirical considerations might justify the assumption that when x takes on its maximal value (or an arbitrarily high value if there is no maximum), both p(correct) and p(high rating) should be expected to be near their ceiling values of 1, in which case the upper bound of the empirically constrained P_1 interval for the fully interpolated RPF could be extended by appending the theoretical (1, 1) data pair to the empirical (P_1, P_2) data.

Such assumptions are similar to the assumptions one might sometimes make in setting *a priori* values for γ (chance rate) and/or λ (lapse rate) in fitting psychometric functions. A similar approach is used in the nonparametric estimation of area under the ROC curve in the calculation of A_g (Pollack and Hsieh, 1969), in which the theoretical (false alarm rate, hit rate) data points (0, 0) and (1, 1) are appended to the empirical data to allow calculation of the area over all possible values of false alarm rate.

In certain cases, experiments may include presentation of stimuli at x = 0 (e.g., zero contrast grating stimuli or equivalently, grating-absent stimuli). For such stimuli, accuracy measures such as p(correct) and *d'* are undefined if they pertain to discrimination of stimulus features (e.g. grating tilt), since there is no such feature present to begin with. However, other measures such as p(high confidence) and reaction time can still have defined values for x = 0 stimuli. If constructing an RPF using one of each type of variable – say, p(correct) for P_1 and p(high confidence) for P_2 – then the variable that is defined at x = 0 (e.g. p(high confidence)) will have one more data point than the one that is not (e.g. p(correct)). However, since the variable that is undefined at x = 0 must be so by virtue of being an accuracy measure, it should also have a natural chance value that can be assumed to hold at x = 0 (e.g. 0.5 for p(correct) in a two-alternative task). (Although accuracy measures are undefined at x = 0, assuming a chance value here can be justified by considering this to be the limiting value that the accuracy measure approaches as x becomes arbitrarily close to 0.) Thus, in such cases it is natural to append a theoretical chance value for the accuracy measure at x = 0 (e.g. p(correct)) and use this in conjunction with the empirical data collected at x = 0 for the other variable (e.g. p(high confidence)) to even out the number of data points in P_1 and P_2 and form a new (P_1, P_2) data pair that extends the lower bound of the P_1 interval to its minimal possible value corresponding to x = 0.

Nonparametric RPF AUC analysis: statistical considerations

As discussed in the main manuscript, there may be cases where it is desirable to estimate RPF AUC nonparametrically. However, this invites the question of whether nonparametric methods are as effective as parametric methods in estimating the true RPF AUC value. Here we investigate this question using computational simulations.

Our overall approach is to repeatedly simulate data from a diverse range of known RPFs, and then use the simulated data to estimate RPF AUC with the methods of fitting, partial interpolation, and full interpolation, as described in the previous section. We can then compare the estimated AUCs from each method to the known true AUC to assess how well each method performs. We also investigate the influence of trial counts and P_1 interval size on the results.

We take our simulated task to be a simple binary classification of a stimulus whose strength can range over [0, 1] (e.g. discriminate whether a grating of contrast *x* is tilting left or right) along with a binary confidence rating (low or high). We assume the task has equal stimulus priors. We take the dependent variables for RPF analysis, P_1 and P_2, to be p(correct) and p(high rating), respectively.

We began by defining four sets of "true" Weibull functions (see Eq. 1 in the main text) for each of F_1 and F_2. We defined the parameters $\theta = (\alpha, \beta, \gamma, \lambda)$ of the four F_1 functions by taking every possible combination of $\alpha_1 \in \{0.3, 0.6\}$ and $\beta_1 \in \{1, 3\}$, with $\gamma_1 = 0.5$ and $\lambda_1 = 0.01$ for each function. Similarly, we defined θ for the four F_2 functions by taking every possible combination of $\alpha_2 \in \{0.3, 0.6\}$ and $\beta_2 \in \{1, 3\}$, with $\gamma_2 = 0.1$ and $\lambda_2 = 0.1$ for each function. This provided a diverse set of functional forms for both $F_1(x;\theta)$ and $F_2(x;\theta)$ (**Figure S1**). Each $F_1(x;\theta)$ and $F_2(x;\theta)$ function from these sets can then be combined to define an RPF, leading to 16 RPFs in total that exhibit a wide range of behaviors (**Figure S2**).

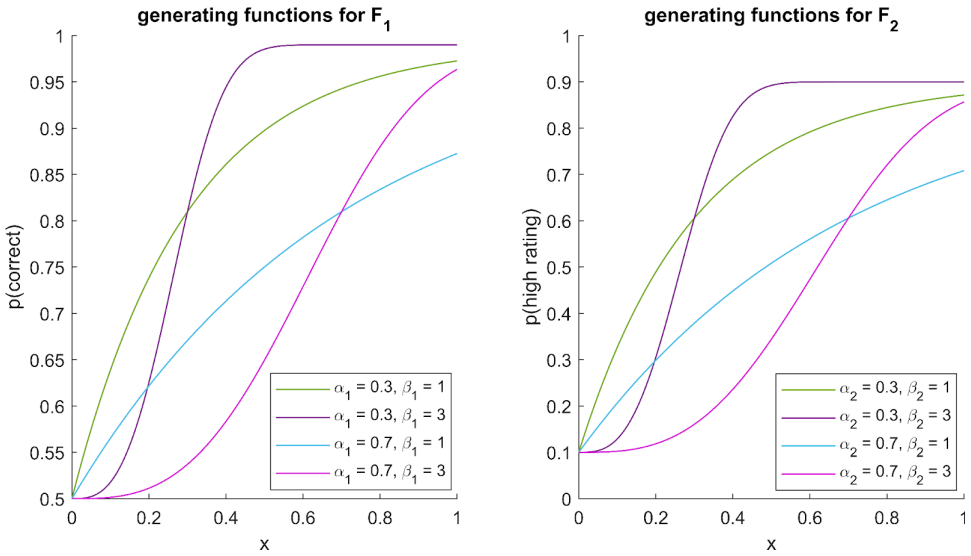

Figure S1. Generating functions for $F_1(x)$ and $F_2(x)$ in the simulations.

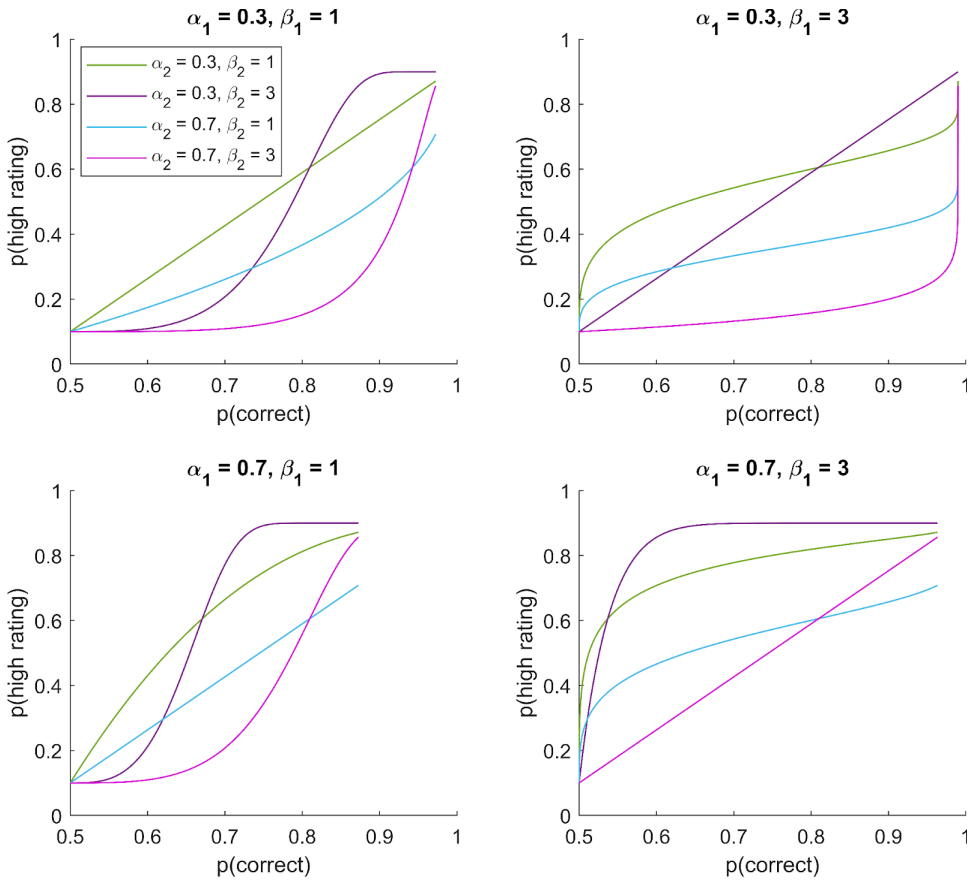

Figure S2. RPFs constructed from the generating functions for $F_1(x)$ and $F_2(x)$ in the simulations.

For each simulated experiment, we defined 6 values of *x* evenly spaced between 0.1 and 0.9, i.e. $x \in \{0.1,\ 0.26,\ 0.42,\ 0.58,\ 0.74,\ 0.9\}$. For each value of *x*, we simulated accuracy on each trial by setting accuracy to 1 if a pseudorandom number drawn from the standard uniform distribution was less than $F_1(x;\theta)$, and 0 otherwise. Similarly, we simulated rating on each trial by setting rating to 2 if a pseudorandom number drawn from the standard uniform distribution was less than $F_2(x;\theta)$, and 1 otherwise. Different simulations used different numbers of trials per level of *x*, with values $N_{\text{trials per x}} \in \{30,\ 50,\ 100,\ 200,\ 500,\ 1000\}$.

We then used the simulated data to estimate the RPF using three methods:

1. **fitting**: fit Weibull functions to both P_1 and P_2 using MLE; use these fits to construct the fitted RPF
2. **partial interpolation**: fit a Weibull function to P_1 using MLE, and use linear interpolation on P_2; use these to constructed the partially interpolated RPF
3. **full interpolation**: perform linear interpolation directly on the plot of P_2 vs. P_1 to construct the fully interpolated RPF

In fitting P_1 (i.e. p(correct)), we constrained γ_1 to the chance value of 0.5, and allowed α_1, β_1, and λ_1 to be free parameters. In fitting P_2 (i.e. p(high rating)), we allowed all of α_2, β_2, γ_2 and λ_2

to be free parameters. We performed these fits with the RPF toolbox (https://github.com/CNClaboratory/RPF), which also uses functions sourced from the Palamedes toolbox version 1.11.11 (Prins and Kingdom, 2018).

For each simulated data set, we computed true RPF AUC and the estimated RPF AUC for each method over three p(correct) intervals:

1. maximum p(correct) interval over the presented x values
2. p(correct) interval = [0.7, 0.8]
3. p(correct) interval = [0.74, 0.76]

We consider the maximum possible p(correct) interval *over the presented x values* because full interpolation cannot extrapolate beyond the empirical P_1 data[5], which itself is limited to the range of presented *x* values. Thus, we constrain *all* analyses for the maximum p(correct) interval to operate over the presented *x* values to make for a fair comparison among all methods.

The maximum p(correct) interval over the presented *x* values depends on the case being considered. For the true RPF, this p(correct) interval is given by the true values of $P_1 = F_1(x; \theta)$ occurring at the minimum and maximum presented *x* values, i.e. $\left[F_1\left(x_{min}; \theta\right),\ F_1\left(x_{max}; \theta\right)\right]$. Under fitting or partial interpolation, wherein p(correct) is fitted by $\widehat{P_1} = F_1\left(x; \widehat{\theta}\right)$, the p(correct) interval is given by the fitted $\widehat{P_1}$ values at the minimum and maximum presented values of x, i.e. $\left[F_1\left(x_{min}; \widehat{\theta}\right),\ F_1\left(x_{max}; \widehat{\theta}\right)\right]$. Under full interpolation, the p(correct) interval is given by the minimum and maximum empirical P_1 values, i.e. $\left[P_{1\,\min},\ P_{1\,\max}\right]$.

For each of these cases, we computed RPF AUC over each interval via numerical integration using the methods contained in the RPF toolbox (https://github.com/CNClaboratory/RPF).

Thus, in total, we simulated data from 16 RPFs for 6 different values of number of trials per level of *x*, for a total of 96 settings for simulated experiments. For each simulation setting, we estimated RPF AUC and $\overline{P}_2$ over 3 p(correct) intervals with the 3 methods of fitting, partial interpolation, and full interpolation. For each simulation setting, we ran 1000 simulated experiments.

Occasionally it occurred that for a simulated data set, the analysis could not proceed due to invalid AUC data occurring for at least one method applied to at least one p(correct) interval. This could occur due the MLE fit to either F_1 or F_2 data producing invalid parameter estimates, such as infinite slope, due to noise in the data. It could also occur due to at least one RPF

[5] In the previous section, we discussed cases where the P_1 interval under interpolation can be expanded by appending theoretical (P_1, P_2) data pairs. For the present simulation, although we can safely assume p(correct) = 0.5 at *x* = 0, there is no theoretical basis for assuming an *a priori* value for p(high rating) at *x* = 0, nor for assuming values for either variable at *x* = 1. Thus, in this case interpolation of the RPF is limited to the empirical P_1 data.

estimation method not having a maximal p(correct) window that fully spanned the pre-specified p(correct) interval of [0.7, 0.8]. This could occur e.g. if the maximum empirical p(correct) value in the simulated data was below 0.8, or if the maximum fitted p(correct) value at x_{max} was below 0.8. In these instances, the simulated data were discarded to ensure that for all data being analyzed, all methods had valid AUC results for all p(correct) intervals. When data was discarded in this way, we conducted extra simulation repetitions to ensure that every simulation setting wound up with 1000 repetitions containing fully valid data. **Table S1** summarizes the proportion of total repetitions that had fully valid data for each trial count setting.

$N_{\text{trials per x}}$	30	50	100	200	500	1000
p(valid)	0.894	0.926	0.961	0.991	0.999	1

Table S1. Proportion of simulation repetitions containing fully valid data for each level of $N_{\text{trials per x}}$.

In **Figures S3 - S7**, we show that over different P_1 intervals and RPF parameter settings, partial interpolation and full interpolation exhibit comparable overall performance to MLE fitting with regards to retrieving the true AUC, with some methods performing slightly better in some contexts than others. We observed similar AUC results in simulations using the P_1 interval [0.74, 0.76], and similar $\overline{P}_2$ results across all simulation settings (data not shown).

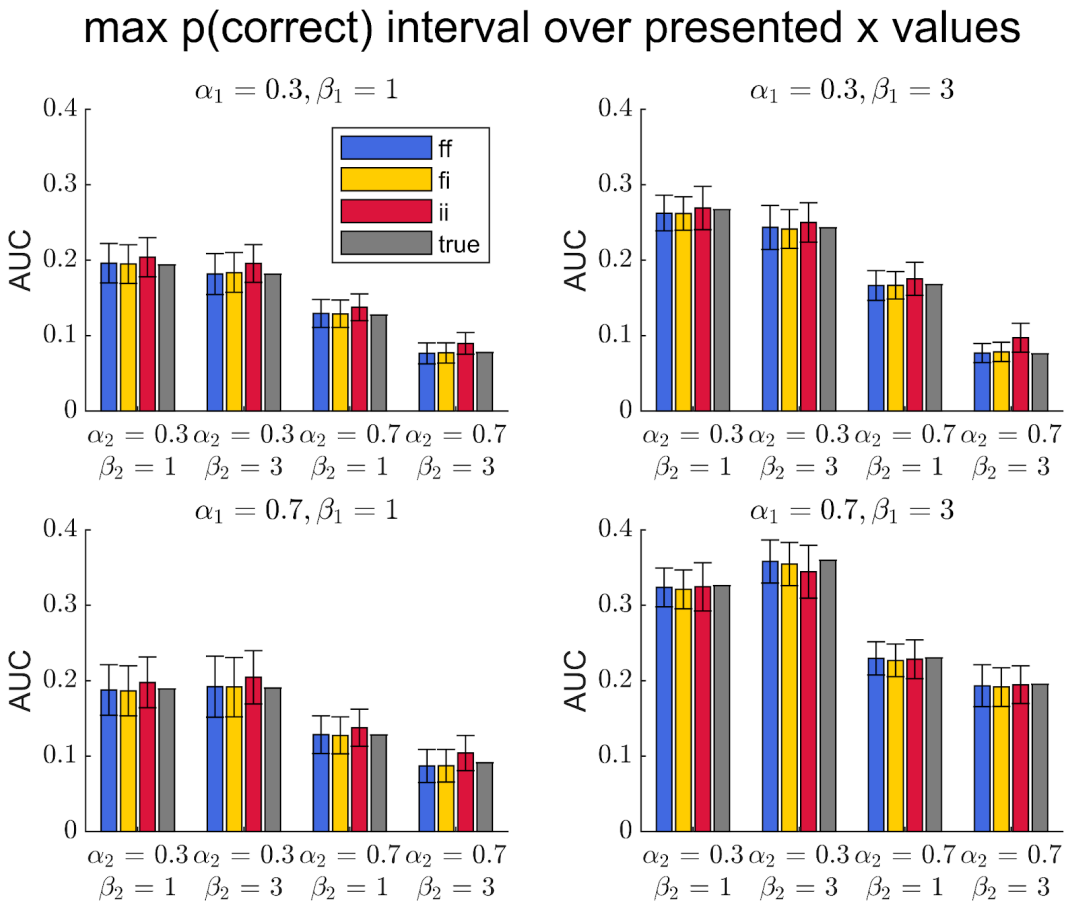

Figure S3. Accuracy and precision of RPF AUC estimation are robust to estimation method over maximal P_1 intervals. RPF AUCs estimated from simulated data using methods ff ~ fitted, fi ~ partial interpolation, ii ~ full interpolation, and compared to true AUC computed from known generating RPF. Error bars show standard deviation across simulations.

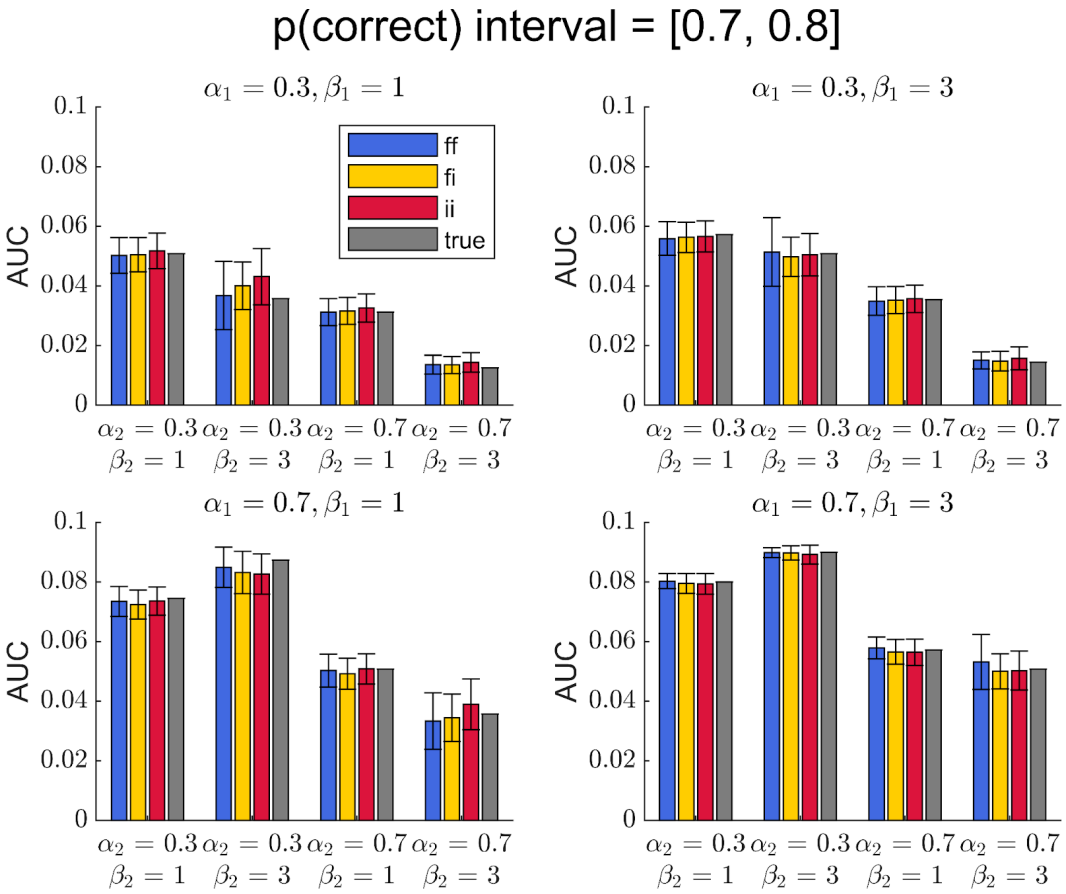

Figure S4. Accuracy and precision of RPF AUC estimation are robust to estimation method over narrower P_1 intervals. ff ~ fitted, fi ~ partial interpolation, ii ~ full interpolation. Error bars show standard deviation across simulations.

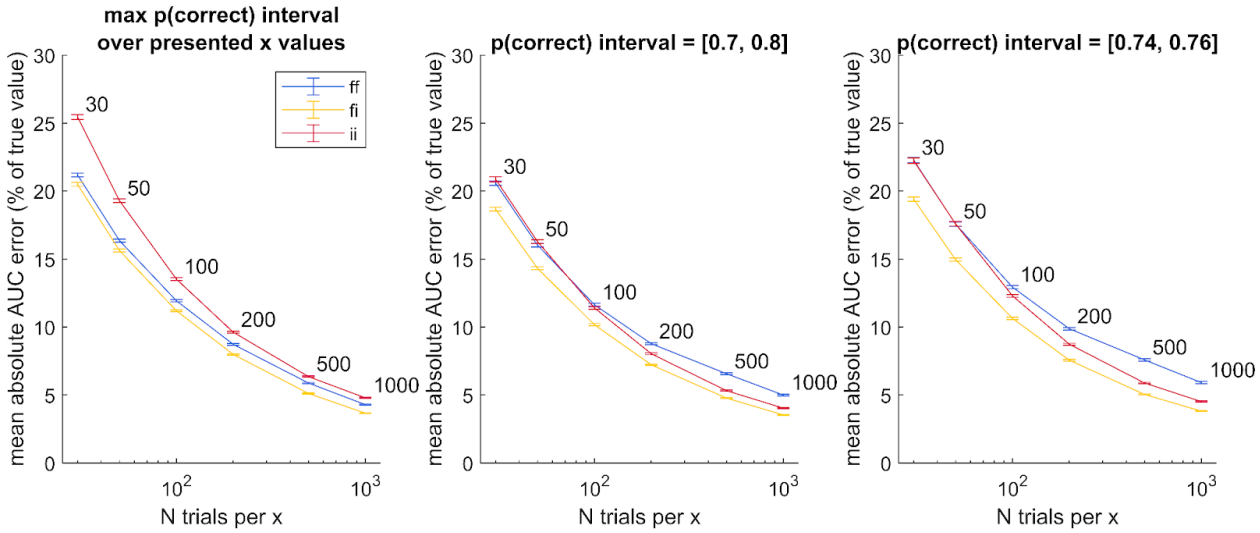

Figure S5. Mean absolute error in AUC estimation across all parameter settings for F_1 and F_2 as a function of fitting method, number of trials, and P_1 interval. ff ~ fitted, fi ~ partial interpolation, ii ~ full

interpolation. The $N_{\text{trials per x}}$ *x*-axis is displayed on a $\log_{10}$ scale, and numbers next to each data point show the corresponding $N_{\text{trials per x}}$ for clarity. Error bars show standard error of the mean across simulations.

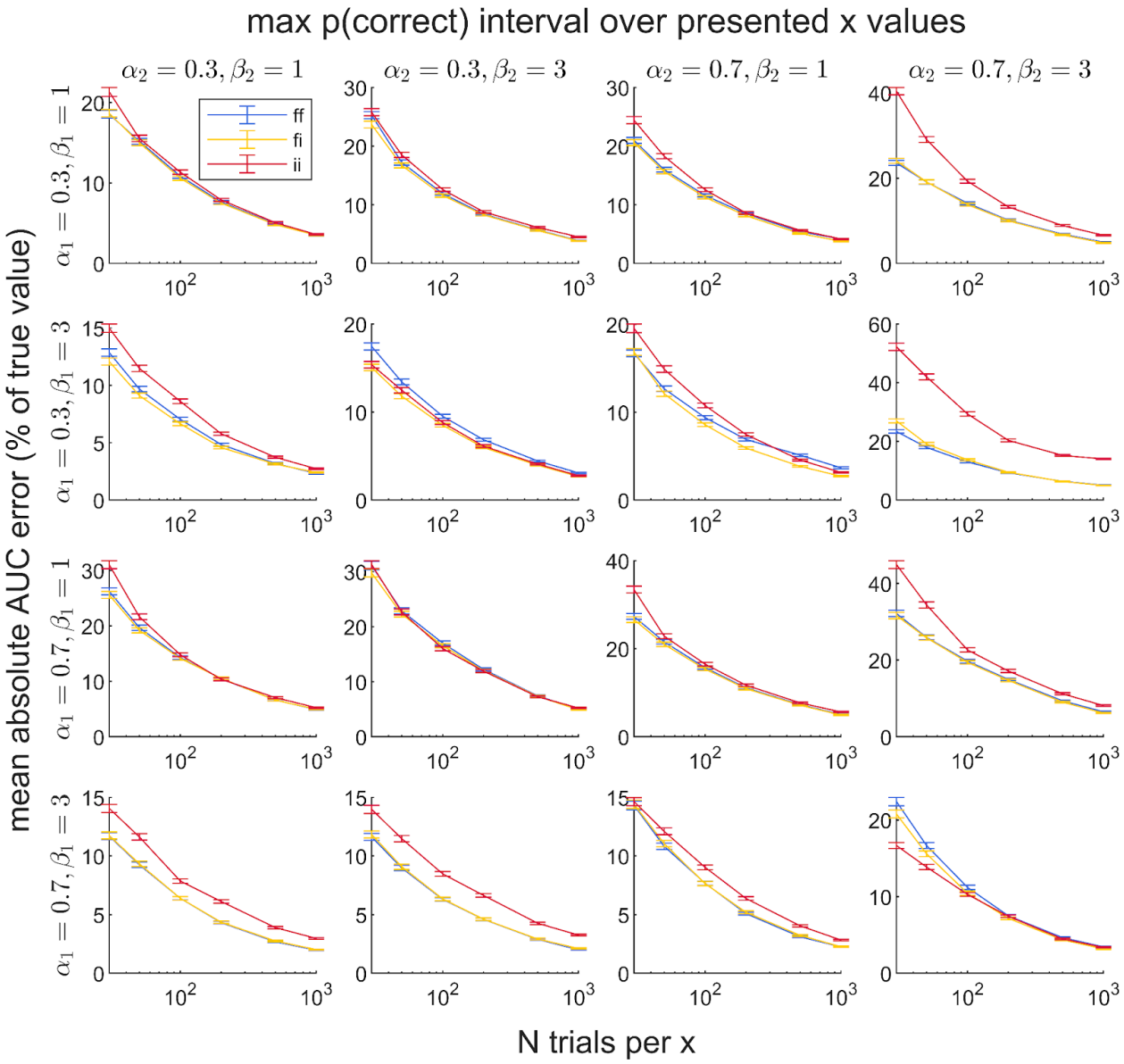

Figure S6. Mean absolute error in AUC estimation for the max p(correct) interval, over each permutation of parameter settings for F_1 and F_2, as a function of fitting method and number of trials. ff ~ fitted, fi ~ partial interpolation, ii ~ full interpolation. The $N_{\text{trials per x}}$ *x*-axis is displayed on a $\log_{10}$ scale. Error bars show standard error of the mean across simulations.

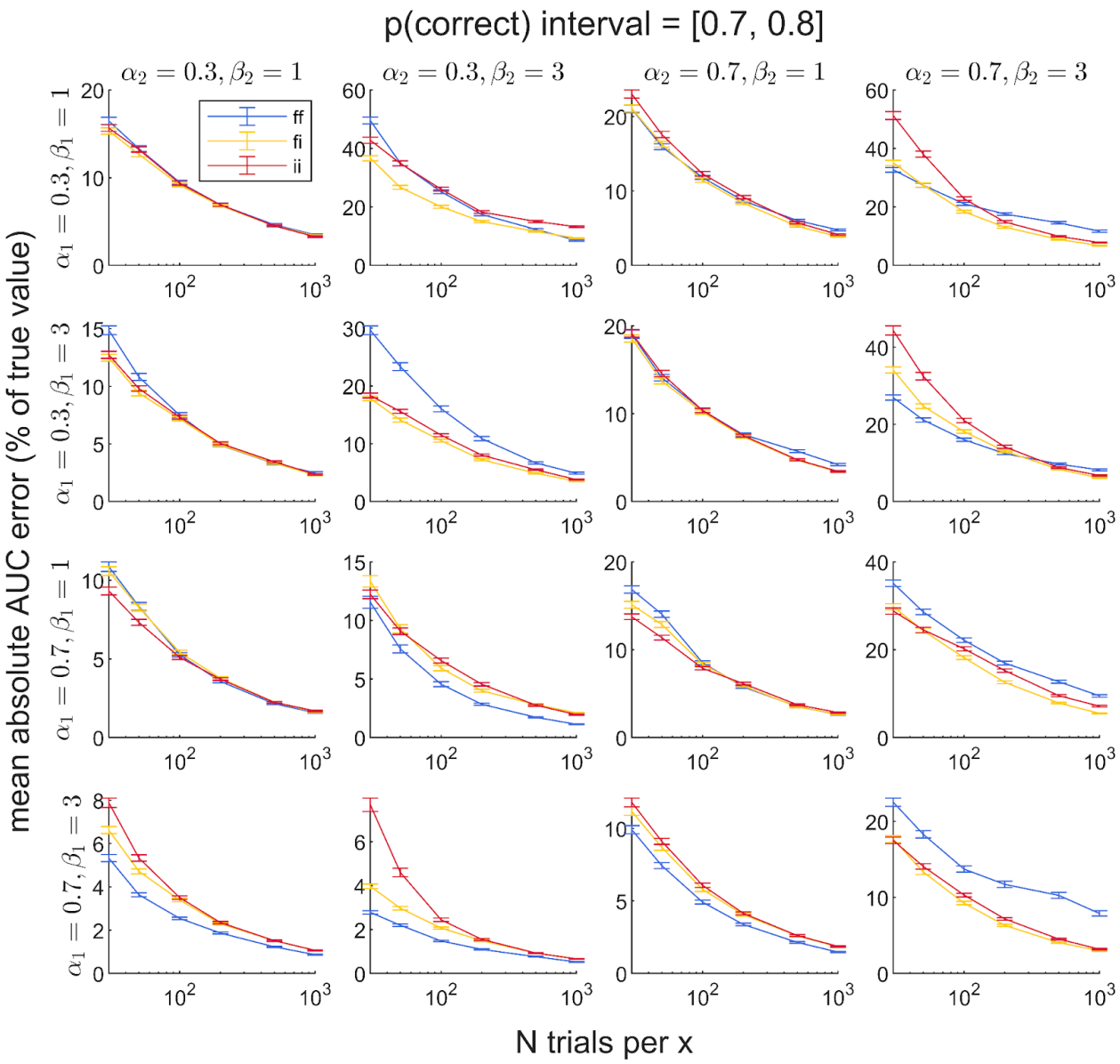

Figure S7. Mean absolute error in AUC estimation for the [0.7, 0.8] p(correct) interval, over each permutation of parameter settings for F_1 and F_2, as a function of fitting method and number of trials. ff ~ fitted, fi ~ partial interpolation, ii ~ full interpolation. The $N_{\text{trials per x}}$ *x*-axis is displayed on a $\log_{10}$ scale. Error bars show standard error of the mean across simulations.

Detailed methods for empirical case study

Participants

27 University of California Riverside students (19 female, 8 male, 26 right-handed, mean age = 20.6 (SD = 3.1)) provided written informed consent to participate in the main study. All

participants had normal or corrected-to-normal vision and normal or corrected-to-normal hearing, and were compensated at a rate of $10/hour for their participation. All study procedures were approved by the University of California Riverside Institutional Review Board.

Prior to the main group-level analysis, data from individual participants were inspected for quality. Data from six participants were excluded from the main analysis due to having performance at or near chance levels across all motion coherence levels (n=3), having completely flat (n=1) or excessively noisy (n=1) confidence vs d' curves, and using a single confidence rating on almost all trials (n=1). Therefore, 21 participants were included in the main analyses reported here.

Stimuli & equipment

All stimuli were presented on a CRT monitor (NEC MultiSync FE2111SB-BK, width 39.6 cm, height 29.7 cm) with refresh rate 75 Hz. A random dot kinematogram (RDK) filling the entire screen (width x height = 43.2 x 33.1 degrees of visual angle (deg)) was presented continuously throughout every block of trials. Dots were black on a white background, with dot size = 0.1 deg, speed = 6 deg/sec, and lifetime = 67 ms (5 frames). When a dot's lifetime expired, it was removed from the screen and replaced with a new dot having a full lifetime and randomly determined location and motion direction. At the start of each block, dots were initialized with uniformly distributed "age," such that on every frame refresh of the screen, one-fifth of the dots expired and were respawned. Dots that moved outside the bounds of the screen continued their motion trajectory from the opposite side of the screen.

Dot Density took on one of three possible values (Low = 1 dot/deg^2, Medium = 3 dots/deg^2, High = 9 dots/deg^2), and was varied either across blocks (Trial Structure: Blocked) or across trials within a block (Trial Structure: Interleaved). When Dot Density decreased from trial N to trial N+1, a randomly selected portion of the dots were deleted in order to achieve the appropriate density. When Dot Density increased, an appropriate number of new dots were spawned with uniformly distributed age and randomly selected location and motion direction.

A fixation cross (width = 0.35 deg) was presented in the center of the screen. Color of the fixation cross changed depending on trial state (see below). Participants were instructed to maintain fixation on the fixation cross throughout each block. To prevent dots from visually interfering with the fixation cross, any dots whose locations fell inside a small circular region in the center of the screen (diameter = 2 deg) were not displayed.

The critical stimulus event occurring on every trial was the occurrence of 533 ms of coherent downward motion in a circular region of the screen (diameter = 8 deg) whose center was located 7 deg to the left or right of fixation, which we will call the "region of coherence." Motion coherence was drawn from one of seven possible values spaced evenly between 10% and 80%, i.e. [10, 21.67, 33.33, 45, 56.67, 68.33, 80]%.

Coherent motion was created by assigning downward motion to all dots spawned with initial locations falling within the region of coherence with probability p(motion coherence) for a period

lasting 493 ms (37 frames). Thus, onset and offset of motion coherence was temporally smoothed due to being yoked to dot respawning, which occurred for one-fifth of the dots on every frame. In total, motion coherence linearly ramped up during the first 53 ms (4 frames) of motion coherence, remained at full motion coherence for the next 427 ms (32 frames), and then linearly ramped down during the final 53 ms (4 frames). Additionally, since motion direction for every dot was constant throughout its lifetime, there were no sharp perceptual edges around the perimeter of the region of coherence due to abrupt changes in dot motion direction as dots entered and exited the region.

Procedure

Participants sat approximately 50 cm from the screen with their chins in a chinrest. Each trial began with presentation of full-field random dot motion for a pre-stimulus period lasting 1 - 3 s. Pre-stimulus duration was drawn randomly from an exponential distribution on each trial such that the hazard rate was roughly held constant; this meant that during the pre-stimulus period, the amount of time elapsed so far was made to be uninformative about whether the target stimulus was about to occur. During this period the fixation cross was red in order to cue the subject to be ready to detect impending coherent motion. Subsequently, the fixation cross turned black and coherent downward motion appeared in one of the two circular regions of coherence (533 ms). The region of coherence was equally likely to appear on either the left or right side of fixation.

After stimulus offset, participants were given three seconds to report the side in which they saw the downward movement (by pressing the 1 or 2 key) and how confident they were in their judgment on a scale of 1 to 4 (using the 7 8 9 0 keys). On trials where participants could not clearly make out the location of coherent motion, they were encouraged to enter a response anyway by making a random guess. To provide feedback on registry of keyboard input, the fixation cross turned gray after entry of the left / right decision and disappeared after entry of confidence. The full 3 s of the response period played out even on trials where participants entered their perceptual decision and confidence rating prior to the expiration of the 3 s time limit. A schematic of trial structure is shown in **Figure 4A** in the main text.

Blocked versus Interleaved Trial Structure design

Participants underwent two trial-order conditions in which Dot Density was either presented pseudorandomly across trials in an *Interleaved* design, or was *Blocked* by Dot Density. In the Interleaved type trials, the density level on each trial was pseudorandomly drawn from any of the three density levels (Low, Medium, or High); in the Blocked type trials, all trials within a block had the same density. In both Trial Structure conditions, within each block of trials all coherence levels were presented in pseudorandom order.

The order of the Blocked versus Interleaved Trial Structure conditions was counterbalanced across two days of testing, such that half of participants underwent the Blocked condition on Day 1 and the Interleaved condition on Day 2, and the other half underwent the Interleaved condition first. Trials in both the Interleaved and Blocked conditions were presented across nine

blocks of trials per day with 84 trials in each block (12 trials per coherence level in each block). In the Blocked trials, Dot Density was pseudorandomly assigned to block number, subject to the constraints that (1) blocks 1-3, 4-6, and 7-9 contained one each of the Low, Medium, and High Dot Density conditions, and (2) density could not be identical across consecutive blocks.

Overall, participants completed 756 trials total in each of the Blocked and Interleaved Trial Structure conditions, with 36 trials for each combination of Trial Structure (Blocked, Interleaved), Dot Density (Low, Medium, High), and motion coherence (7 levels in total, spaced evenly between 10% and 80% coherence). Each day of testing lasted about an hour and 15 minutes, such that participants underwent about 2.5 hours of testing in total. Day 2 occurred between 1 - 3 days after Day 1. A schematic of Trial Structure (Blocked, Interleaved) is shown in **Figure 4B** in the main text.

Prior to testing on each day, participants performed at least one block of practice trials (and possibly more depending on the discretion of the experimenter, who monitored participant performance during practice to ensure adequate understanding and performance of the task). During practice, participants engaged in the same task as the main task, but also received trial-by-trial auditory feedback regarding the correctness of their responses (high tone for correct, low tone for incorrect). Practice blocks contained 12 trials in which the three levels of Dot Density were pseudorandomly interleaved (even on Blocked condition days), with motion coherence set to 100%. During the entirety of the first 6 trials of a practice block, red circles were shown around the edges of the left and right regions of coherence in order to familiarize the participant with what regions of the screen could potentially contain coherent motion. The practice was designed to allow participants to become comfortable with the task and response options, and to ensure they understood the task and key mappings for choices and confidence ratings.

All behavioral procedures were programmed in PsychToolbox and implemented on a MacBook Pro with OSX Version 10.9.5 running Matlab r2013b.

Detailed fitting of P_1 (*d'*)

For fitting *d'* as a function of *x* (in RPF analysis, this would be for fitting P_1) we follow the approach described above to define *d'* as a function of stimulus strength *x* via the scaled Weibull distribution. To fit *d'* to the present empirical dataset, we set constraints to be $\gamma_n = 0$ (since at chance performance, d' = 0) and ω_n = maximum possible *d'* value achievable. This maximum achievable value is controlled by the number of trials present in the dataset at each stimulus level *x*, combined with choices about how to avoid hit and false alarm rates being 1 or 0, respectively. Specifically, we must make decisions about what is called 'padding' to make the maximum possible hit rate (HR) less than 1, with the amount less than 1 depending on the number of trials in the condition of interest. Similarly, we want the minimum possible cell-padded false alarm rate (FAR) to be greater than 0. The *d'* that is maximum for a given cell padding is defined as z(max HR possible with cell padding) - z(min FAR possible with cell padding).

We accomplish this goal by 'padding' the number of responses in a given response category such that no response category contains zero responses. Response categories are defined as the combination of a type 1 response (e.g. left or right) and confidence rating (here, a rating from 1-4). Thus, for each possible stimulus presented (here: coherent motion *presented* on the left or right side of the screen), we count the number of "reported right" and "reported left" responses, separated by the confidence level that was also reported.

Algorithmically, in our RPF toolbox (https://github.com/CNClaboratory/RPF) *d'* is calculated through reliance on the scripts developed by Maniscalco and Lau (Maniscalco and Lau, 2014, 2012), which separates response data into two arrays containing these response categories. Concretely, these two matrices `nR_S1` and `nR_S2` are vectors containing the total number of responses in each response category, conditional on presentation of S1 (e.g., 'stimulus on the left') and S2 (e.g., 'stimulus on the right'). The following description is copied from the relevant section of the RPF toolbox for clarity:

```
% e.g. if nR_S1 = [100 50 20 10 5 1], then when stimulus S1 was
% presented, the subject had the following response counts:
% responded S1, rating=3 : 100 times
% responded S1, rating=2 : 50 times
% responded S1, rating=1 : 20 times
% responded S2, rating=1 : 10 times
% responded S2, rating=2 : 5 times
% responded S2, rating=3 : 1 time
%
% The ordering of response / rating counts for S2 should be the same
as it is for S1. e.g. if nR_S2 = [3 7 8 12 27 89], then when stimulus
S2 was
% presented, the subject had the following response counts:
% responded S1, rating=3 : 3 times
% responded S1, rating=2 : 7 times
% responded S1, rating=1 : 8 times
% responded S2, rating=1 : 12 times
% responded S2, rating=2 : 27 times
% responded S2, rating=3 : 89 times
```

Here, each response count cell in `nR_S1` and `nR_S2` for each level of condition and *x* is padded with a value 1/(2*nRatings) – i.e., this value is added to all response categories – where nRatings refers to the number of available confidence ratings in the experiment (here, 4). For example, using this number of ratings, `nR_S1 = [100 50 20 10 5 1]` becomes `nR_S1 = [100.125 50.125 20.125 10.125 5.125 1.125]`. (Interested readers can also refer to our toolbox README, specifically the RPF_guide('info') section entitled "Fitting *d'* and meta-*d'*" and RPF_guide('padInfo'), for more detailed information).

With this approach, the present data the max cell padded *d'* thus achieved is 3.8759, so we constrain ω_n to be 3.8759 for all conditions. Thus only α_n and β_n are free parameters fitted to the data, which we fit separately for each condition.

Detailed fitting of P_2 (mean confidence rating and meta-*d'*)

To fit mean confidence as a function of *x* (for P_2), we again used the scaled Weibull distribution for mean confidence as described above to fit mean confidence as a function of stimulus strength *x*. In these fits, no constraints were placed on any of the four psychometric function parameters. To fit metacognitive sensitivity (meta-*d'*) as a function of stimulus strength *x* (also for P_2), we again used the custom likelihood functions described above for the meta-*d'* scaled Weibull. For each condition in this dataset, we constrained γ_n = 0 and ω_n = 3.8759 (as with the fit for *d'*).

Notes about group fitting for plotting

In the main text, for illustrative purposes **Figures 5** and **6** show MLE fits of the RPF to the group data concatenated across all subjects into one single large dataset containing all trials for all individual subjects. This concatenation required an assumption about cell padding for the purposes of avoiding HR = 1 and FAR = 0 (as described above) to avoid underestimation of the effect of cell padding choices on the group fit relative to the effect on single-subject fits. Specifically, because we have 21 subjects, the group data has 21 times the amount of trials of any individual subject's dataset. We therefore multiply the cell padding factor in the group fits by 21, such that the cell padding for the group fit would be the same fraction of total trial counts as it was for each individual subject. This plotting approach therefore gives a better representation of the group average over each individual subject analysis. We remind the reader that all statistical measures were derived from single-subject fits to each individual condition for each individual subject, so these choices affect the visual presentation of the group data for illustrative purposes only and have no effect on the statistical analyses and conclusions presented in the main text.