Interval bias in 2AFC detection tasks: sorting out the artifacts

Miguel A. García-Pérez · Rocío Alcalá-Quintana

Published online: 7 July 2011
© Psychonomic Society, Inc. 2011

Abstract Proportion correct in two-alternative forced choice (2AFC) detection tasks often varies when the stimulus is presented in the first or in the second interval. Reanalysis of published data reveals that these order effects (or *interval bias*) are strong and prevalent, refuting the *standard difference model* of signal detection theory. Order effects are commonly regarded as evidence that observers use an off-center criterion under the *difference model with bias*. We consider an alternative *difference model with indecision* whereby observers are occasionally undecided and guess with some bias toward one of the response options. Whether or not the data show order effects, the two models fit 2AFC data indistinguishably, but they yield meaningfully different estimates of sensory parameters. Under indeterminacy as to which model governs 2AFC performance, parameter estimates are suspect and potentially misleading. The indeterminacy can be circumvented by modifying the response format so that observers can express indecision when needed. Reanalysis of published data collected in this way lends support to the indecision model. We illustrate alternative approaches to fitting psychometric functions under the indecision model and discuss designs for 2AFC experiments that improve the accuracy of parameter estimates, whether or not order effects are apparent in the data.

Keywords Two-alternative forced choice · Detection · Psychophysical function · Psychometric function · Interval bias · Order effects

Many empirical studies on detection processes require estimating the psychometric function, a function describing how performance improves with stimulus magnitude. Two-alternative forced choice (2AFC) tasks are the most common method used for obtaining the estimates. Each trial in these tasks consists of two presentation intervals (in temporal 2AFC) or two spatial locations (in spatial 2AFC), one of which displays a blank (i.e., a stimulus of zero magnitude), whereas the other displays the test stimulus of interest.¹ The observer is asked to indicate which interval or location displayed the test, and the response is scored as *correct* when the reported interval is that in which the test was actually displayed. An experimental session typically includes a number of test magnitudes or levels and a sufficiently large number of trials at each level so that

Electronic supplementary material The online version of this article (doi:10.3758/s13414-011-0167-x) contains supplementary material, which is available to authorized users.

M. A. García-Pérez (✉) · R. Alcalá-Quintana
Departamento de Metodología, Facultad de Psicología,
Universidad Complutense,
Campus de Somosaguas,
28223 Madrid, Spain
e-mail: miguel@psi.ucm.es

¹ Some authors (e.g., Jäkel & Wichmann, 2006) use the acronym 2AFC to refer only to what we call *spatial* 2AFC and use the alternative acronym 2IFC (for two-interval forced choice) to refer to what we call *temporal* 2AFC. On this issue, we line up with Green and Swets (1966, pp. 107–108), who always used the term *two-alternative forced-choice task* and the acronym *2AFC* and define the forced choice task as one in which “two or more *stimulus alternatives* are presented The alternatives may be defined by regions of space but, in the more common form of the task, they are defined by intervals of time” (our italics). One further reason to use a single acronym regardless of how the stimulus alternatives are presented is that, on some occasions, the two alternatives are defined both by regions of space and by intervals of time (e.g., Armstrong & Marks, 1997). We find it more convenient to refer to the latter as *spatiotemporal* 2AFC than to seek yet a third acronym for this situation.

proportion correct can be estimated reasonably accurately. The order of presentation of test and blank on each trial is usually decided at random with equiprobability.

Data analysis proceeds by fitting a sigmoidal curve to proportion correct as a function of test level. Data from trials on which the test was presented first vs. second (or left vs. right in spatial 2AFC) are usually pooled together on the assumption that presentation order does not affect performance. Justification for this practice comes from the *standard difference model* in signal detection theory (SDT), which is briefly described next.

In SDT, sensory effects are conceived of as random variables whose mean increases with stimulus level and whose variance is, without loss of generality, independent of stimulus level. For convenience, sensory effects S are assumed to be normally distributed with density

$$f(s; m, v) = \frac{1}{\sqrt{2\pi v}} \exp \left[-\frac{(s - m)^2}{2v} \right], \quad (1)$$

where m is the mean and v is the variance. The relation of mean sensory effect m to stimulus level x is given by the monotone increasing, nonnegative psychophysical function μ , which, for reasons that will become apparent below, we will assume is given by

$$\mu(x) = -\tau \Phi^{-1} \left(\frac{1}{2} - \frac{1}{2} \Phi \left(\frac{x - \theta}{\beta} \right) \right), \quad (2)$$

where Φ is the unit-normal distribution function and $\tau > 0$, $\beta > 0$, and $\theta \in \mathbb{R}$ determine the location and steepness of μ . If the stimulus level x is log contrast, $x \leq 0$ and $\theta \leq 0$ (see Fig. 1a). In the sequel, we will assume a gain factor of $\tau = \sqrt{2}$, except where otherwise indicated.

In a 2AFC detection task, the blank draws a sensory effect from a distribution given by Eq. 1 with $m = 0$, whereas a test with level x draws an effect with a nonzero mean $m = \mu(x)$. In either case, the variance is $v = 1$ (see Fig. 1b). The standard difference model posits that the observer chooses the interval that draws the largest sensory effect. This rule subsumes a decision variable D , representing the difference, say, between the sensory effect S_2 elicited in the second interval and the sensory effect S_1 elicited in the first interval, so that the observer responds “interval 1” when $D < 0$ and “interval 2” when $D > 0$. If the test was displayed in the first interval, D has mean $-\mu(x)$ and variance 2 and the probability of a correct response is given by $\text{Prob}(D < 0)$, whereas if the test was displayed in the second interval, D has mean $\mu(x)$ with the same variance and the probability of a correct response is $\text{Prob}(D > 0)$ (see Fig. 1c). Mathematically, the

(a) Psychophysical function

(b) Distributions of sensory effects

(c) Decision space; test first, test second

(d) Psychometric functions

Fig. 1 The standard difference model in signal detection theory. **a** Psychophysical function given by Eq. 2, with parameters shown in the inset. **b** Normal distributions of sensory effects with unit variance. The mean sensory effect of the blank is 0; the mean sensory effect of the test is 0.8 given its magnitude ($x = -1.018$), as indicated by the horizontal line in panel a. **c** Decision space when the test is presented in the first interval (red) and in the second interval (blue). For the particular stimulus magnitude used in this illustration, the probability of a correct response is .714 whether the test is presented first or second. **d** Psychometric function when the test is presented in the first interval (dashed red curve) and in the second interval (dashed blue curve), each obtained as described in panel c by varying the magnitude of the test stimulus and, hence, the mean of the decision variable

probability Ψ_1 of a correct response when the test is presented first, as a function of test level, is thus

$$\Psi_1(x) = \int_{-\infty}^0 f(d; -\mu(x), 2)dd = \Phi\left(\frac{\mu(x)}{\sqrt{2}}\right) = \frac{1}{2} + \frac{1}{2}\Phi\left(\frac{x - \theta}{\beta}\right), \quad (3a)$$

that is, a scaled cumulative Gaussian with location θ and standard deviation β . If the test is presented second, the probability Ψ_2 of a correct response is

$$\Psi_2(x) = \int_0^{\infty} f(d; \mu(x), 2)dd = \Phi\left(\frac{\mu(x)}{\sqrt{2}}\right) = \Psi_1(x). \quad (3b)$$

Thus, owing to its symmetry, the standard difference model predicts the same psychometric function whether the test is presented first or second (see Fig. 1d). It is now obvious that the form of μ in Eq. 2 with $\tau = \sqrt{2}$ was chosen so as to render psychometric functions with the form of a scaled normal ogive. Alternative forms for μ or for the distribution of sensory effects yield psychometric functions with other mathematical expressions (see García-Pérez & Alcalá-Quintana, 2007), but in all cases, $\Psi_1 = \Psi_2$ under this model, which is the main issue here.

In contrast, when proportion correct has been computed separately for each presentation order (e.g., Jäkel & Wichmann, 2006; Yeshurun, Carrasco, & Maloney, 2008), the data have shown remarkable differences. These order effects refute the standard difference model, and they are more widely known as *interval bias* in visual psychophysics. Order effects can also be dug out of published data that were pooled across presentation orders before fitting psychometric functions. For instance, García-Pérez (2000, his Fig. 8) reported psychometric functions for 3 observers in a 2AFC contrast detection task. The data and fitted functions, as reported in the original article, are shown in black in Fig. 2, whereas functions fitted separately to data for each presentation order are shown in red and in blue.

Proportion correct differs greatly across presentation orders at all stimulus levels, except when performance reaches 100% correct. Also, the magnitude and sign of these order effects vary across observers. Reanalysis of detection data reported in many other studies yields similar results, and this evidence is presented in Section A of the [Supplementary Material](#), which also describes how psychometric functions were fitted in our reanalysis.

As the evidence just mentioned reveals, order effects are prevalent in 2AFC detection data and yet, these data are typically pooled across presentation orders and used to fit psychometric functions without consideration of the contaminating influence of order effects on estimates of sensory parameters. The only case in which order effects are sometimes considered is in experiments aimed at estimating d' for some isolated stimulus level, where a correction for bias is applied that presumably renders uncontaminated estimates (Klein, 2001). This correction arises from the so-called *difference model with bias*, which posits that observers set an off-center, nonzero boundary in the decision space. It is also true that, sometimes, differences in performance across presentation orders are not observed, which is often taken to justify conventional estimation approaches based on data pooled across presentation orders. The main goal of this article is to develop and test an alternative SDT model of performance in 2AFC tasks that posits that observers are sometimes undecided due to insufficient sensory evidence and, then, give a response that is affected by a (nondecisional) key bias. As will be discussed below, this notion of indecision can be traced back to Fechner (1860/1966), but it has not been fully developed formally thus far. The difference model with bias and this alternative *difference model with indecision* are both compatible with the presence or absence of order effects, but the two models yield different estimates of sensory parameters even when the data do not show order effects. This situation calls for research aimed at

Fig. 2 Order effects in the data of García-Pérez (2000). Each panel pertains to a different observer and shows the psychometric function for data aggregated across presentation orders (in black, as reported in Fig. 8 of García-Pérez, 2000) and separated for the subsets of trials in

which the test was presented first (red) or second (blue). Order effects occur at all stimulus levels and vary in magnitude (absolute difference between proportion correct when the test is first and second) and sign (sign of the difference) across observers

determining which model actually governs 2AFC performance and, hence, how the data should be analyzed to obtain adequate estimates of sensory parameters.

The plan of the article is as follows. The indecision model is presented in a broad context including three other well-known amendments to the standard difference model, all of which are compatible with the presence or absence of order effects of some sort. Of these models, only two predict order effects with the empirical characteristics illustrated in Fig. 2 and Section A of the [Supplementary Material](#). These two models are indeed the difference model with bias and the difference model with indecision. The models cannot be tested against each other using conventional 2AFC data, because they differ only at the unobservable decision stage, but they can be tested against data collected with a response format that allows observers to express indecision if needed. Reanalysis of published data from experiments of this type shows that only the indecision model accounts for the data adequately. We then illustrate the magnitude of the misestimation of sensory parameters that may occur when 2AFC data (whether showing order effects or not) are analyzed with conventional methods. We also describe and illustrate several approaches to fitting conventional 2AFC data, as well as data from experiments in which expressing indecision was allowed. Finally, we give recommendations for the design of 2AFC experiments.

Amendments to the standard difference model

Three major and well-known amendments to the standard difference model have been proposed directly or indirectly to account for order effects. These amendments are described next, and then we will present a fourth amendment that also accounts for order effects.

The difference model with bias

The issue of order effects (then dubbed *interval bias*) was probably first addressed within SDT by Green and Swets (1966, pp. 408–410), and they regarded it as potentially indicating off-center decision criteria. As compared with the standard difference model in Fig. 1, the difference model with bias assumes that the boundary for “interval 1” and “interval 2” responses does not occur at $D = 0$ but at $D = c$ for some $c \neq 0$. Thus, this model restores the criterion parameter that was an integral part of the analysis of yes–no data and that the standard difference model for 2AFC tasks eliminated by assuming a symmetric decision rule (i.e., $c = 0$). We will refer to the use of an off-center criterion as *decisional bias*. This model is illustrated in Fig. 3 for the case where $c > 0$, which breaks the symmetry of the standard difference model and renders functions Ψ_1 and Ψ_2 that differ

Fig. 3 The difference model with bias. With respect to the standard difference model illustrated in Fig. 1, the only change is that the decision boundary is not at $D = 0$ but at some nonzero value c , which is 0.5 in this illustration

throughout their domain. When $c > 0$, as in Fig. 3, $\Psi_1 > \Psi_2$; if $c < 0$, $\Psi_1 < \Psi_2$ instead. Thus, this model predicts psychometric functions with the characteristics of empirical data, and these functions are given by

$$\Psi_1(x) = \int_{-\infty}^c f(d; -\mu(x), 2)dd = \Phi\left(\frac{c + \mu(x)}{\sqrt{2}}\right) \quad (4a)$$

and

$$\Psi_2(x) = \int_c^{\infty} f(d; \mu(x), 2)dd = \Phi\left(\frac{-c + \mu(x)}{\sqrt{2}}\right). \quad (4b)$$

The difference model with different sensitivities

Yeshurun et al. (2008) speculated that the differences that they found in proportion correct across presentation orders might reveal differences in sensitivity across intervals. They also designed an experiment whose aim was to test for these differences, and they concluded from their results that sensitivity is higher during the second interval. This conclusion has been disputed because of methodological and statistical flaws and also because of the contradicting results of their two experiments (see García-Pérez & Alcalá-Quintana, *in press*): The first experiment revealed that proportion correct was lower in the second interval, whereas the second experiment revealed that sensitivity was actually higher during the second interval, despite the fact that the same observers participated in both experiments. Which-

ever the actual situation is regarding differences in sensitivity across intervals, it is worth looking at their implications on order effects.

In SDT, different sensitivities imply a different d' in each interval and, hence, different psychophysical functions. Let μ_1 and μ_2 be the psychophysical functions that apply during the first and second intervals, respectively. Under the mathematical form of Eq. 2, these functions may differ as to τ , β , or θ , and the psychometric functions are given by

$$\Psi_1(x) = \int_{-\infty}^0 f(d; -\mu_1(x), 2)dd = \Phi\left(\frac{\mu_1(x)}{\sqrt{2}}\right) \quad (5a)$$

and

$$\Psi_2(x) = \int_0^{\infty} f(d; \mu_2(x), 2)dd = \Phi\left(\frac{\mu_2(x)}{\sqrt{2}}\right). \quad (5b)$$

Suppose that μ_1 and μ_2 differ in parameter τ , as in Fig. 4a. Mean sensory effects thus vary with presentation order (see Fig. 4b), so that symmetry no longer occurs at the decision stage (see Fig. 4c), producing psychometric functions that differ in steepness but not in lower asymptote across presentation orders (see Fig. 4d). If μ_1 and μ_2 differed instead as to β , Ψ_1 and Ψ_2 would differ in steepness and would cross at $x = \theta$, but they would share the lower asymptote. Finally, if μ_1 and μ_2 differed as to θ , Ψ_1 and Ψ_2 would be laterally displaced, but they would have the same steepness and lower asymptote. Across empirical evidence presented in Fig. 2 and Section A of the [Supplementary Material](#), none of the data sets showing clear order effects appear to involve a common lower asymptote for Ψ_1 and Ψ_2 , which rules out differences in sensitivity as a general explanation of order effects.

The difference model with sensation weighting

Green and Swets (1966) noted that interval bias might also reveal “that the measures of sensory excitation in the various intervals are subjected to different multipliers before the largest is chosen” (p. 408). Mathematically, this implies that the decision variable is computed as $D = w_2S_2 - w_1S_1$ with $w_2 \neq w_1$ (see Hellström, 1979). Psychometric functions under this model are given by

$$\Psi_1(x) = \int_{-\infty}^0 f(d; -w_1\mu(x), w_1^2 + w_2^2)dd = \Phi\left(\frac{w_1\mu(x)}{\sqrt{w_1^2 + w_2^2}}\right) \quad (6a)$$

and

$$\Psi_2(x) = \int_0^{\infty} f(d; w_2\mu(x), w_1^2 + w_2^2)dd = \Phi\left(\frac{w_2\mu(x)}{\sqrt{w_1^2 + w_2^2}}\right). \quad (6b)$$

Fig. 4 The difference model with different sensitivities. With respect to the standard difference model illustrated in Fig. 1, the only change is that the psychophysical function differs by a multiplicative factor for stimuli presented in the first or the second interval. In this illustration, parameter τ in Eq. 2 equals $\sqrt{2}$ in the first interval and $1.7\sqrt{2}$ in the second

As can be seen in Fig. 5a, this model implies that the (absolute) mean of D varies across presentation orders as a result of the different weights w_1 and w_2 . Besides the inconsequential change in the variance of D , the sensation-weighting model is equivalent to the difference model with different sensitivities as determined by changes in parameter τ of the psychophysical functions. Figure 5b shows that sensation weighting also yields psychometric functions that differ in steepness but not in lower asymptote across presentation orders and, then, sensation weighting

Fig. 5 The difference model with sensation weighting. With respect to the standard difference model illustrated in Fig. 1, the only change is that the decision variable is a weighted difference of the sensory effects elicited during the first and second intervals, with weights $w_1 = 1.4$ and $w_2 = 0.8$ in this illustration

can also be ruled out as the cause of order effects with the empirical characteristics documented in this article.

The difference model with indecision

The three models just presented share with the standard difference model the defining characteristic that the observer's response is always determined by unequivocal sensory evidence. On any given trial, either $D < c$ or $D > c$, or, put another way, the observer is never undecided. This characteristic contrasts with the experience reported by the early psychophysicists (including Fechner) to the effect that observers were sometimes undecided because they could not tell which interval had displayed the stimulus (for a review of this literature, see the first few chapters of Link, 1992).

Incorporating indecision into SDT requires restoring Fechner's (1860/1966) *interval of uncertainty*, as proposed by Watson, Kellogg, Kawanishi, and Lucas (1973) and further elaborated by Kaernbach (2001), García-Pérez (2002), and García-Pérez and Alcalá-Quintana (2010a; for an application to 2AFC discrimination tasks, see Alcalá-Quintana & García-Pérez, 2011; García-Pérez & Alcalá-Quintana, 2010b). The difference model with indecision posits that observers experience indecision when S_1 and S_2 are within some neighborhood of each other such that $|D| \leq \delta$. In these cases, observers must guess under 2AFC instructions to respond either "interval 1" or "interval 2" on every trial. Tune (1964) reviewed

overwhelming evidence to the effect that in these situations, observers do not produce responses with equiprobability. Actually, for reasons discussed by Jäkel and Wichmann (2006) and García-Pérez and Alcalá-Quintana (2010a), observers may show a *response bias* toward one of the intervals when undecided. Thus, let ξ_1 be the probability that the observer responds "interval 1" when undecided, so that $1 - \xi_1$ is the probability that the observer responds "interval 2." Note that this response bias is qualitatively different from the decisional bias discussed earlier. The psychometric functions for this model (see Fig. 6) are

$$\begin{aligned}\Psi_1(x) &= \int_{-\infty}^{-\delta} f(d; -\mu(x), 2)dd + \xi_1 \int_{-\delta}^{\delta} f(d; -\mu(x), 2)dd \\ &= \Phi\left(\frac{-\delta + \mu(x)}{\sqrt{2}}\right) + \xi_1 \left[\Phi\left(\frac{\delta + \mu(x)}{\sqrt{2}}\right) - \Phi\left(\frac{-\delta + \mu(x)}{\sqrt{2}}\right) \right]\end{aligned}\quad (7a)$$

and

$$\begin{aligned}\Psi_2(x) &= \int_{\delta}^{\infty} f(d; \mu(x), 2)dd + (1 - \xi_1) \int_{-\delta}^{\delta} f(d; \mu(x), 2)dd \\ &= \Phi\left(\frac{-\delta + \mu(x)}{\sqrt{2}}\right) + (1 - \xi_1) \left[\Phi\left(\frac{\delta + \mu(x)}{\sqrt{2}}\right) - \Phi\left(\frac{-\delta + \mu(x)}{\sqrt{2}}\right) \right].\end{aligned}\quad (7b)$$

Fig. 6 The difference model with indecision. With respect to the standard difference model illustrated in Fig. 1, the only change is that there is an *interval of uncertainty* (Fechner, 1860/1966) around the null value of the decision variable, whose half-width δ is 1 in this illustration. When undecided, observers must guess under conventional 2AFC instructions and, then, they give "interval 1" or "interval 2" responses with probabilities that are not necessarily equal ($\xi_1 = .8$), as indicated above the horizontal arrows in panel a. At the stimulus level in this illustration, the probability of a correct response is thus $.444 + .8 \times .455 = .808$ when the test is presented first, whereas it is $.444 + .2 \times .455 = .535$ when the test is presented second

Note that $\xi_1 = .5$ renders $\Psi_1 = \Psi_2$ so that the absence of response bias prevents order effects. Also, $\xi_1 = 0$ yields the same Ψ_1 and Ψ_2 as the difference model with bias when $c = -\delta$, whereas $\xi_1 = 1$ yields the same functions as the difference model with bias when $c = \delta$; when $0 < \xi_1 < 1$, the two models yield different pairs of psychometric functions that are difficult to tell apart, because both share the characteristic that Ψ_1 and Ψ_2 do not cross and differ in lower asymptote.

In sum, of the four alternatives to the (implausible) standard difference model that have been described in this section, only the difference model with bias and the difference model with indecision are compatible with the empirical characteristics of order effects. The question naturally arises as to which of these two models governs performance on a 2AFC task, but this question cannot be addressed empirically with conventional 2AFC data, because the models differ only at the unobservable decision stage and also because, with adequate parameters, both models can fit any 2AFC data virtually identically.² Yet anyone who has participated in a 2AFC experiment has experienced indecision on some trials. On this issue, Klein (2001) stated that “with feedback, lots of practice, and lots of experience being a subject, I was able to reduce this interval bias and equalize the number of times I responded with each interval. Naïve subjects may have a more difficult time reducing the bias” (p. 1425). This statement describes an active participation of the observer in choosing what to respond on some trials, as if there was a choice that the difference model with bias does not actually allow: According to the model, all trials render either $D > c$ or $D < c$, and, hence, the observer’s response is entirely determined by sensory evidence. Changing a response based on sensory evidence by the opposite response in the sole interest of equalization would be a serious infringement of the instructions of the experiment and will represent experimental contamination. We actually see Klein’s (2001) statement as a tacit endorsement of the indecision model, implying that a conscientious observer should try to equalize the number of “interval 1” and “interval 2” responses so as to respond with $\xi_1 = .5$ when undecided. The rest of the article develops the indecision model and illustrates some of its additional advantages, which are best exploited by modifying the 2AFC task to allow observers to report indecision. Unlike the conventional

2AFC format, this variant of the 2AFC task allows testing the different predictions that the model with bias and the model with indecision make regarding undecided responses and, hence, allows determining which model should be used to estimate sensory parameters.

A general indecision model and the unforced choice task

Consider an experiment involving a 2AFC detection task in which observers are given the option to respond “I don’t know” when undecided, instead of being forced to respond “interval 1” or “interval 2” on every trial. This response format is identical to that involved in a three-point rating scale, and, hence, there is nothing unconventional about it. Interestingly, the indecision model makes precise predictions for this task. Specifically, the probabilities of a correct response when the test is displayed in the first or second interval are, respectively,

$$\Psi_1(x) = \int_{-\infty}^{-\delta} f(d; -\mu(x), 2) dd = \Phi\left(\frac{-\delta + \mu(x)}{\sqrt{2}}\right) \quad (8a)$$

and

$$\Psi_2(x) = \int_{\delta}^{\infty} f(d; \mu(x), 2) dd = \Phi\left(\frac{-\delta + \mu(x)}{\sqrt{2}}\right) = \Psi_1(x), \quad (8b)$$

so that no differences in performance should be observed whether the test stimulus is presented first or second. Also, the probability Υ of “I don’t know” responses in each interval are given by

$$\Upsilon_1(x) = \int_{-\delta}^{\delta} f(d; -\mu(x), 2) dd = \Phi\left(\frac{\delta + \mu(x)}{\sqrt{2}}\right) - \Phi\left(\frac{-\delta + \mu(x)}{\sqrt{2}}\right) \quad (9a)$$

and

$$\begin{aligned} \Upsilon_2(x) &= \int_{-\delta}^{\delta} f(d; \mu(x), 2) dd = \Phi\left(\frac{\delta - \mu(x)}{\sqrt{2}}\right) - \Phi\left(\frac{-\delta - \mu(x)}{\sqrt{2}}\right) \\ &= \Upsilon_1(x), \end{aligned} \quad (9b)$$

so that, again, no differences in how the proportion of undecided cases varies with stimulus level should be observed whether the test stimulus is presented first or second. This is illustrated graphically in the top panel of Fig. 7 for the case depicted in Fig. 6. When $\delta = 1$, as in this illustration, the interval of uncertainty is relatively narrow and renders few undecided cases even at low stimulus levels. More typically, the interval of uncertainty will be broader, yielding curves as shown in the bottom panel of Fig. 7. It is interesting to note that the relation of proportion correct to stimulus level (i.e., the shape of Ψ_1 and Ψ_2) differs in the top and bottom panels of Fig. 7, but so does

² The virtually identical fit of the data does not involve a mere replacement of parameters δ and ξ_1 in the difference model with indecision with some value for parameter c in the difference model with bias (or the other way around); it also involves meaningfully different estimates of the sensory parameters θ and β . The fact that fitting one or the other model yields different estimates of the sensory parameters requires establishing which model governs performance and, then, which of the two sets of parameter estimates is authentic.

Fig. 7 Probability of correct responses and indecision as a function of stimulus level under the difference model with indecision and in an experiment in which the response options are “interval 1,” “interval 2,” and “I don’t know.” Curves in the top panel arise from a relatively narrow interval of uncertainty with $\delta = 1$; curves in the bottom panel arise from a broader interval of uncertainty with $\delta = 3$. All remaining parameters are as in Fig. 6, although ξ_1 is not involved here (see Eqs. 8a, 8b, 9a and 9b)

the relation of proportion undecided to stimulus level (i.e., the shape of Υ_1 and Υ_2). Despite these differences in shape contingent on the value of the indecision parameter δ , the underlying sensory parameters θ and β are identical in both cases and can be recovered, although certainly not by a simple approach that relies only on the functions Ψ_1 and Ψ_2 and neglects consideration of the functions Υ_1 and Υ_2 .

These relations allow testing the indecision model against the model with bias, which, instead, predicts differences between Ψ_1 and Ψ_2 accompanied by $\Upsilon_1(x) = \Upsilon_2(x) = 0$. The three-response format has been used in a number of published studies, although the data were not used to test these predictions. Figure 8 shows the results of a reanalysis of data in García-Pérez and Alcalá-Quintana (2007) from a contrast detection task in which psychometric functions were fitted to data pooled across presentation orders and in which undecided responses were converted into “interval 1” and “interval 2” responses with equiprobability. Our reanalysis tallies correct and undecided responses at each stimulus level separately in each interval and reveals for observer 1 a pattern that is consistent with the prediction in Fig. 7, but some deviation from this pattern is clearly apparent for the two other observers: Observer 3 shows a slight lateral displacement of the curves for each presentation order, and observer 2 shows a pattern with

large differences in the lower asymptotes of Ψ_1 and Ψ_2 , but with no differences between Υ_1 and Υ_2 . A description of how the data in Fig. 8 were fitted is given in Section B of the [Supplementary Material](#), which also presents a reanalysis of data from other experiments in which undecided cases were recorded. Next, we show that the departures from model predictions observed in Fig. 8 can be accounted for on the assumption that other factors are involved, such as decisional or response bias (none of which is actually incompatible with the basic postulate of an interval of uncertainty).

Consider a general indecision model in which the interval of uncertainty has half-width δ around an off-center boundary c and in which intervals have psychophysical functions μ_1 and μ_2 . Consider further that, under instructions to report undecided cases, an observer fails to do so with probability κ (which does not vary with presentation order) and, in these cases, produces his/her own guesses with a response bias determined by the probability ξ_1 discussed earlier. Finally, consider that finger errors occur with probability λ , making the observer respond “interval 1” (alternatively, “interval 2”) when the intended response was “interval 2” (alternatively, “interval 1”). Then, the probabilities of a correct response when the test is displayed in the first or second interval are

$$\begin{aligned} \Psi_1(x) = & (1 - \lambda) \int_{-\infty}^{c-\delta} f(d; -\mu_1(x), 2) dd \\ & + \kappa \xi_1 \int_{c-\delta}^{c+\delta} f(d; -\mu_1(x), 2) dd + \lambda \int_{c+\delta}^{\infty} f(d; -\mu_1(x), 2) dd \end{aligned} \quad (10a)$$

and

$$\begin{aligned} \Psi_2(x) = & (1 - \lambda) \int_{c+\delta}^{\infty} f(d; \mu_2(x), 2) dd \\ & + \kappa(1 - \xi_1) \int_{c-\delta}^{c+\delta} f(d; \mu_2(x), 2) dd \\ & + \lambda \int_{-\infty}^{c-\delta} f(d; \mu_2(x), 2) dd. \end{aligned} \quad (10b)$$

Analogously, the probability of an undecided response in each presentation interval is given by

$$\Upsilon_1(x) = (1 - \kappa) \int_{c-\delta}^{c+\delta} f(d; -\mu_1(x), 2) dd \quad (11a)$$

and

$$\Upsilon_2(x) = (1 - \kappa) \int_{c-\delta}^{c+\delta} f(d; \mu_2(x), 2) dd. \quad (11b)$$

For a graphical illustration, consider this model with $\mu_1 = \mu_2$ and $\lambda = 0$ for an observer who fully complies with

Fig. 8 Proportion correct and proportion undecided for each presentation order in the study of García-Pérez and Alcalá-Quintana (2007). These data were reported as conventional psychometric

functions for data aggregated across presentation orders in Fig. 4 of the original article, with each undecided response randomly transformed into a correct or an incorrect response with equiprobability

the instructions to report undecided cases (i.e., $\kappa = 0$) but who has some decisional bias (i.e., $c \neq 0$). The top panel in Fig. 9 shows that decisional bias produces curves that are displaced laterally for each presentation order, as was the case for observer 3 in Fig. 8. Alternatively, consider an observer with no decisional bias (i.e., $c = 0$) but who, at times, produces his/her own guesses (i.e., $\kappa > 0$) with a response bias toward the first interval (i.e., $\xi_1 > .5$). As

is shown in the bottom panel of Fig. 9, in these conditions, Υ_1 and Υ_2 are identical, whereas Ψ_1 and Ψ_2 differ in their lower asymptotes, in agreement with the data from observer 2 in Fig. 8. In sum, the general indecision model captures the characteristics of order effects in Fig. 7.

Implications for the analysis of 2AFC detection data

Detection performance is determined by (1) sensory factors that define the characteristics of the sensory effects S_1 and S_2 and (2) decisional/response factors that define how the observer uses the realizations of S_1 and S_2 to give a response. The standard difference model estimates d' as $\hat{d}' = \sqrt{2}\Phi^{-1}(\hat{p})$, where $\hat{p}$ is the proportion of correct responses pooled across presentation orders, but order effects may seriously contaminate $\hat{p}$ Green and Swets (1966, pp. 108, 408–410) referred to “a procedure for correcting the estimate of [proportion correct] for any interval bias that may persist,” a procedure consisting of estimating d' under the difference model with bias. Indeed, Eqs. 4a and 4b prescribe estimating d' as $\hat{d}' = [\Phi^{-1}(\hat{p}_1) + \Phi^{-1}(\hat{p}_2)]/\sqrt{2}$, which is the correction proposed by Green and Swets. The difference model with bias also estimates c as $\hat{c} = [\Phi^{-1}(\hat{p}_1) - \Phi^{-1}(\hat{p}_2)]/\sqrt{2}$. Of course, this model yields $c = 0$ and the standard estimate of d' when $\hat{p}_1 = \hat{p}_2 = \hat{p}$.

But use of this correction as advocated by Klein (2001) yields incorrect estimates if observers behave according to the indecision model, as is shown next. Consider first the example in Fig. 3, where a biased observer with $c = 0.5$ and $d' = 0.8$ would (without sampling error) obtain $\hat{p}_1 = .821$ and $\hat{p}_2 = .584$. With the difference model with bias, the estimated sensitivity is $\hat{d}' = [\Phi^{-1}(.821) + \Phi^{-1}(.584)]/\sqrt{2} = 0.8$, and the estimated criterion is $\hat{c} = [\Phi^{-1}(.821) - \Phi^{-1}(.584)]/\sqrt{2} = 0.5$, thus recovering the true d' and c . Consider now the example in Fig. 6, where an unbiased

Fig. 9 Predictions of the general indecision model. The top panel shows predictions for an observer with $\delta = 3$ who complies with the directions to report indecision ($\kappa = 0$) but has decisional bias ($c = -0.5$). The bottom panel shows predictions for an observer also with $\delta = 3$ who fails to comply with the instructions to report indecision and, thus, produces his/her own guesses with probability $\kappa = .5$ and who does not have decisional bias ($c = 0$) but has response bias toward the first interval ($\xi_1 = .8$). In both panels, the psychophysical function is as illustrated in Fig. 1a

observer ($c = 0$) with indecision ($\delta = 1$) and response bias ($\xi_1 = .8$), but otherwise with the same $d' = 0.8$, would obtain $\hat{p}_1 = .808$ and $\hat{p}_2 = .535$. The correction from the (inadequate) difference model with bias yields $\hat{d}' = [\Phi^{-1}(.808) + \Phi^{-1}(.535)]/\sqrt{2} = 0.678$ (about 15% lower than the true d') and “recovers” a nonexistent $\hat{c} = [\Phi^{-1}(.808) - \Phi^{-1}(.535)]/\sqrt{2} = 0.55$ alongside. What this example illustrates is not that the classical correction for bias is intrinsically wrong but, rather, that its validity cannot be taken for granted without knowledge of what response process actually generated the data. With only two independent sources of data (the empirical proportions $\hat{p}_1$ and $\hat{p}_2$) and also two parameters to estimate (d' and c), the difference model with bias will always fit the data errorlessly. Its downside is that no degrees of freedom are left to test the model. Then, the model “succeeds” by simply enforcing the interpretation that decisional bias is all that makes $\hat{p}_1$ and $\hat{p}_2$ differ.

The picture is actually a little more complex because the absence of differences in proportion correct across presentation orders does not provide evidence of the validity of the standard difference model. As was mentioned earlier, an undecided observer (i.e., $\delta > 0$) without response bias (i.e., $\xi_1 = .5$) will obtain the same proportion of correct responses in both intervals. To illustrate, consider again the case depicted in Fig. 6, but now with $\xi_1 = .5$, so that $\hat{p}_1 = \hat{p}_2 = .444 + .5 \times .455 = .6715$. Then, under the presumed applicable standard difference model, $\hat{d}' = \sqrt{2}\Phi^{-1}(.6715) = 0.628$, which is about 22% lower than the true d' of 0.8.

In sum, knowledge of the source model is needed to estimate d' adequately. Interestingly, experimental methods have been devised that provide sufficient degrees of freedom to test the general indecision model and allow estimating its parameters (see García-Pérez & Alcalá-Quintana, 2010a). This can help to determine whether $\delta = 0$ and, thus, whether the standard difference model or the difference model with bias holds on a subject-by-subject basis. Similar considerations apply when the goal is to fit a psychometric function to 2AFC detection data, an issue that we address in the next section.

Fitting psychometric functions to 2AFC detection data

This section discusses several approaches to fitting psychometric functions to 2AFC detection data in the presence of order effects. The various approaches differ according to whether undecided cases were or were not recorded explicitly. These approaches are not presented here as competitors but as alternatives to be used

according to what each particular situation demands. This is one of the reasons that we are not comparing the goodness of the fit across approaches³: Actually, the illustrations below show that all approaches produce curves that are indistinguishable by eye and that follow the path of the data equally well. In these conditions, indices of goodness of fit provide only quantitative confirmation of this characteristic. We will nevertheless report p -values for the log-likelihood ratio statistic G^2 in each case.

Fitting psychometric functions when undecided cases are not recorded

If observers are forced to respond either “interval 1” or “interval 2” on each trial, a sensible option is to fit the general indecision model in Eqs. 10a and 10b with $\kappa = 1$ (i.e., observers always produce their own guesses when undecided, affected by a response bias given by parameter ξ_1). This approach involves two data sets (one from trials on which the test was presented first and another one from trials on which it was presented second) and, allowing for decisional bias and potential differences in sensitivity across intervals, seeks to estimate eight parameters (λ , c , δ , ξ_1 , θ_1 , θ_2 , β_1 , and β_2).

Let $X^{(1)} = \{x_1^{(1)}, x_2^{(1)}, \dots, x_{N^{(1)}}^{(1)}\}$ be the set of $N^{(1)}$ distinct test levels across trials on which the test was presented first, and let $X^{(2)} = \{x_1^{(2)}, x_2^{(2)}, \dots, x_{N^{(2)}}^{(2)}\}$ be the set of $N^{(2)}$ distinct test levels across trials on which the test was presented second. Also, let $C_i^{(1)}$ and $W_i^{(1)}$ be the number of correct and wrong responses, respectively, at test level $x_i^{(1)}$ ($1 \leq i \leq N^{(1)}$) and similarly let $C_j^{(2)}$ and $W_j^{(2)}$ be the number of correct and wrong responses, respectively, at test level $x_j^{(2)}$ ($1 \leq j \leq N^{(2)}$). Maximum-

³ However, the most important reason to omit a detailed analysis of the goodness of the fit (which also appears unnecessary in the circumstances just described) is that conventional indices of fit are inadequate in situations (like ours) in which (1) the various approaches imply nonnested models (for a discussion of this issue, see the March 2000 and April 2006 issues of the *Journal of Mathematical Psychology*), (2) the models, in turn, imply distinct data sets (dichotomous correct/incorrect responses vs. trichotomous correct/incorrect/undecided responses, sometimes even from a common data source), and (3) the implied theoretical probabilities are very small or very large at the far left and far right ends of the fitted functions (on the relevance of this issue, see García-Pérez & Núñez-Antón, 2004). In addition, bootstrap methods are inadequate for data collected adaptively because bootstrap samples typically involve different numbers of stimulus levels and, thus, imply different bootstrap distributions.

likelihood estimates of the eight free parameters are thus obtained by maximizing the joint likelihood function

$$\prod_{i=1}^{N^{(1)}} \left[\Psi_1(x_i^{(1)}) \right]^{C_i^{(1)}} \left[1 - \Psi_1(x_i^{(1)}) \right]^{W_i^{(1)}} \times \prod_{j=1}^{N^{(2)}} \left[\Psi_2(x_j^{(2)}) \right]^{C_j^{(2)}} \left[1 - \Psi_2(x_j^{(2)}) \right]^{W_j^{(2)}}, \quad (12)$$

where Ψ_1 and Ψ_2 are given by Eqs. 10a and 10b, respectively, with $\kappa = 1$.

Figure 10a shows the psychometric functions arising from the fit of the general indecision model to data presented in Fig. 2. Table 1 lists the estimated parameters for each observer. It should be stressed that the black curve in each panel is simply the average of the fitted functions Ψ_1 and Ψ_2 , just as each of the black circles is the weighted average of the ordinates of red and blue circles at the same abscissa.⁴ Thus, the black curve has not been fitted to the data indicated by black circles. Put another way, the psychometric function for data pooled across presentation orders is a mere by-product of the functions describing performance under each presentation order.

It should also be noted that the overall fit is just as good as that which was obtained when an independent psychometric function was separately fitted to each subset of data in Fig. 2, although there are some slight but noticeable discrepancies between the fitted curves and the paths of each subset of data in Fig. 10a. The reason for the discrepancy does not lie in the indecision model itself but in the additional assumptions that come with, namely, psychophysical functions given by Eq. 2 with $\tau = \sqrt{2}$ and normally distributed sensory effects given by Eq. 1. Yet the path of each data set in Fig. 10a does not seem to accommodate the characteristics of a scaled normal ogive that is odd symmetric about its inflection point, which is what the extra assumptions bring around. The psychometric functions in Fig. 2 were, instead, Weibull functions given by

$$\Psi_i(x) = 1 - \lambda_i - (1 - \lambda_i - \gamma_i) \exp \left[-10^{\beta_i(x - \theta_i)} \right], \quad (13)$$

⁴ The reason for the weighted averaging is that the number of trials with each presentation order was not the same at each stimulus level so that the aggregated proportion correct is not the simple average of the separate proportions.

⁵ For aggregated data, a third function is necessary that has the same general form but whose parameters are not subscripted and where $\gamma = .5$ is no longer a free parameter.

where $i \in \{1, 2\}$ denotes the interval in which the test was presented⁵ and where γ and λ determine the lower and upper asymptotes. In cases like this, one might consider replacing the mathematical forms of μ_1 and μ_2 or the distributions of sensory effects so as to render Weibull psychometric functions (see García-Pérez & Alcalá-Quintana, 2007), but this is an unnecessary complication given the feasibility of an alternative approach discussed next.

The independent fits in Fig. 2 allowed different psychophysical functions (whichever their form is) through consideration of independent slope and location parameters for each presentation order (β_i and θ_i), but they also incorporated independent parameters γ_1 and γ_2 for the lower asymptotes of the curves and which do not exist as such in the indecision model. But the indecision model imposes constraints upon the latter parameters. In particular, for the lower asymptote parameter defined as $\gamma_i = \lim_{x \rightarrow -\infty} \Psi_i(x)$, it is obvious from Eqs. 10a and 10b with $\kappa = 1$ that $\lim_{x \rightarrow -\infty} \Psi_1(x) + \lim_{x \rightarrow -\infty} \Psi_2(x) = \gamma_1 + \gamma_2 = 1$, because $\lim_{x \rightarrow -\infty} \mu_i(x) = 0$. The constraint $\gamma_1 + \gamma_2 = 1$ actually holds for all the models considered earlier. Also, parameters λ_1 and λ_2 should have the same value, because the probability of a finger error should not vary with presentation order. In sum, this alternative approach amounts to seeking maximum-likelihood estimates by also maximizing Eq. 12, but with Ψ_1 and Ψ_2 given by Eq. 13 (or any other convenient function, for that matter) under the constraints $\gamma_2 = 1 - \gamma_1$ and $\lambda_2 = \lambda_1$.⁶ This leaves only six free parameters (namely, γ_1 , λ_1 , θ_1 , θ_2 , β_1 , and β_2) and yields the results shown in Fig. 10b with parameter estimates given in Table 2. The black curve in each panel of Fig. 10b is again the average of Ψ_1 and Ψ_2 and was not fitted to data represented as black circles. The fit is somewhat better than in Fig. 10a (surely because a Weibull function does more justice to the path of the data) and is similar to what was shown in Fig. 2 by fitting a totally independent four-parameter Weibull function to each data set, although there were 11 free parameters in Fig. 2 and only 6 in Fig. 10b.

Thus far, we have considered that the psychophysical functions μ_1 and μ_2 may differ. One might also impose the constraints $\theta_2 = \theta_1$ and $\beta_2 = \beta_1$ and fit the constrained model with only four free parameters (namely, γ_1 , λ_1 , θ_1 , and β_1), if only to check whether the data can be accounted for without assuming differences in sensitivity across intervals. The result is shown in Fig. 10c, with parameter estimates given in Table 3. The fit appears good again as judged by eye: No meaningful differences can be noted in

⁶ These constraints are analogous to those that hold across presentation orders in 2AFC discrimination tasks (see García-Pérez & Alcalá-Quintana, 2011).

Fig. 10 Fitting the indecision model to the data in Fig. 2. **a** Fit of the indecision model in Eqs. 10a and 10b with $\kappa = 1$ and $c = 0$. **b** Fit of the Weibull functions in Eq. 13 under the constraints $\lambda_2 = \lambda_1$ and $\gamma_2 = 1 - \gamma_1$. **c** Fit of the Weibull functions in Eq. 13 under the same constraints as those in panel b, plus the additional constraints $\theta_2 = \theta_1$ and $\beta_2 = \beta_1$. These additional constraints force the functions that apply under each presentation order to differ only in their lower asymptote and, thus, they cannot accommodate the separate paths of the data from observer 3 at high stimulus levels for each presentation order. In spite of these slight discrepancies between the path of the

data and that of the fitted curves in some cases, none of the three approaches rendered a significant G^2 goodness-of-fit statistic for observer 1 or 3, and all of them rendered significant G^2 for observer 2 ($p = .026, .018$, and $.041$, respectively, for the approaches illustrated in panels a, b, and c). The significant results for observer 2 are undoubtedly caused by inflation due to the extremely high empirical and theoretical probabilities of correct responses when the test was presented second (blue curves), and not by a lack of correspondence between the path of the data and that of the fitted curve

the shapes of the fitted curves in Fig. 10c, as compared with those in Fig. 10a or b, except perhaps for observer 3.

Fitting psychometric functions when undecided cases are recorded separately

Experiments in which undecided cases are recorded separately call for the general indecision model in

Eqs. 10a, 10b, 11a and 11b with nine free parameters (λ , c , δ , κ , ξ_1 , θ_1 , θ_2 , β_1 , and β_2). The notation that was introduced in the preceding section must thus be extended to let $U_i^{(1)}$ be the number of undecided responses when the test level $x_i^{(1)}$ ($1 \leq i \leq N^{(1)}$) was presented in the first interval and, similarly, to let $U_j^{(2)}$ be the number of undecided responses when the test level $x_j^{(2)}$ ($1 \leq j \leq N^{(2)}$) was presented in the second interval. Maximum-likelihood

Table 1 Estimated parameters of the indecision model in Fig. 10a

Observer	λ	c	δ	ξ_1	θ_1	θ_2	β_1	β_2
1	.000	0	2.698	.258	-2.470	-2.566	0.172	0.223
2	.046	0	2.738	.071	-2.177	-2.571	0.129	0.281
3	.005	0	2.717	.558	-2.461	-2.483	0.161	0.203

c was actually set to 0 and not regarded as a free parameter

Table 2 Estimated parameters of the constrained Weibull fit in Fig. 10b

Observer	λ_1	γ_1	θ_1	θ_2	β_1	β_2
1	.005	.255	-2.119	-2.119	2.362	2.031
2	.040	.069	-1.908	-1.872	3.064	5.299
3	.009	.584	-2.127	-2.078	2.849	1.894

estimates of the nine free parameters are thus obtained by maximizing the joint likelihood function

$$\prod_{i=1}^{N^{(1)}} [\Psi_1(x_i^{(1)})]^{C_i^{(1)}} [\Upsilon_1(x_i^{(1)})]^{U_i^{(1)}} [1 - \Psi_1(x_i^{(1)}) - \Upsilon_1(x_i^{(1)})]^{W_i^{(1)}} \times \prod_{j=1}^{N^{(2)}} [\Psi_2(x_j^{(2)})]^{C_j^{(2)}} [\Upsilon_2(x_j^{(2)})]^{U_j^{(2)}} [1 - \Psi_2(x_j^{(2)}) - \Upsilon_2(x_j^{(2)})]^{W_j^{(2)}}, \quad (14)$$

where Ψ_1 , Ψ_2 , Υ_1 , and Υ_2 are given by Eqs. 10a, 10b, 11a, and 11b, respectively.

Figure 11a shows the functions arising from the fit of this model to data presented earlier in Fig. 8, and Table 4 lists the estimated parameters. The fit is similar to what was shown in Fig. 8, but only 9 free parameters are involved here (actually 8, because $c = 0$ was assumed), as compared with 16 parameters needed in Fig. 8. The fit in Fig. 11a implies allowance for different psychophysical functions across intervals. One might also impose the constraints $\theta_2 = \theta_1$ and $\beta_2 = \beta_1$ and then fit the reduced model. The results (which thus involve only 7 parameters) are shown in Fig. 11b, and the estimated parameters are given in Table 5. The fit is equally good by eye, since no differences can be noted between the shape of the fitted curves in Fig. 11b and those in Fig. 11a. A comparison of Tables 4 and 5 reveals that removal of parameters θ_2 and β_2 under the assumption of no differences in sensitivity across intervals is captured by the nonzero value of parameter c without altering the values of the remaining parameters in the model, and therefore, neither approach seems to provide distinctly different estimates of the sensory parameters that are typically sought in the experiments.

Table 3 Estimated parameters of the constrained Weibull fit in Fig. 10c

Observer	λ_1	γ_1	θ_1	β_1
1	.006	.250	-2.121	2.319
2	.048	.070	-1.912	3.167
3	.007	.584	-2.094	2.046

The alternative approach of jointly fitting functions of arbitrary shape to each data set under the constraints imposed by the indecision model is also possible. If empirical data appeared to describe logistic functions, the general form of the model functions would be

$$\Psi_i(x) = \gamma_i + \frac{1 - \lambda_i - \gamma_i}{1 + \exp[-\beta_i(x - \theta_i)]} \quad (15)$$

and

$$\Upsilon_i(x) = \frac{\alpha_i}{1 + \exp[\beta_i(x - \theta_i)]}. \quad (16)$$

Note that the location and slope parameters θ_i and β_i are the same in Ψ_i and Υ_i . This is one of the constraints imposed upon this fit, because Ψ_i and Υ_i have the same origin under the indecision model (namely, the function μ_i that applies in each interval). The remaining constraints affect again the asymptote parameters γ_i , λ_i , and α_i , as follows. First, as before, the finger-error parameters can be assumed equal for both presentation orders, so that $\lambda_2 = \lambda_1$. Next, from Eqs. 11a and 11b, $\alpha_i = \lim_{x \rightarrow -\infty} \Upsilon_i(x)$ and $\alpha_2 = \alpha_1$, again because $\lim_{x \rightarrow -\infty} \mu_i(x) = 0$. Finally, from Eqs. 10a and 10b, $\gamma_i = \lim_{x \rightarrow -\infty} \Psi_i(x)$. It is also easy to see that these parameters satisfy $\alpha_1 + \gamma_1 + \gamma_2 = 1$. Then, fitting Eqs. 15 and 16 under the constraints of the indecision model amounts to estimating only seven free parameters (λ_1 , γ_1 , γ_2 , θ_1 , θ_2 , β_1 , and β_2). The resultant fit is equally good (i.e., no difference between the shapes of the fitted curves in this case and the shapes of the curves arising from the previous approaches), as shown in Fig. 11c with estimated parameters given in Table 6. However, imposing the constraints $\theta_2 = \theta_1$ and $\beta_2 = \beta_1$ does not seem reasonable for these particular data. As can be seen in Fig. 11d, the fitted curves for both presentation orders are identical as a result of the extra constraint. In contrast, the three other approaches yielded curves that are slightly laterally displaced for each presentation order, just as the data seem to require (see the panels for observer 3 in Figs. 11a–c).

Finally, a third approach to dealing with undecided cases is to apply Fechner's (1860/1966) method of considering them "half right and half wrong." This strategy can be referred back to the general indecision model in Eqs. 10a, 10b, 11a and 11b, since it implies an "ideal equalizer" who, at each stimulus level and presentation order, would produce the same number of "interval 1" and "interval 2" responses when undecided (i.e., $\kappa = 1$ and $\xi_1 = .5$). Yet this strategy is better regarded as a two-step process in which observers behave with $\kappa = 0$ (i.e., they always report indecision) and an external mechanism evenly transforms undecided cases into correct and incorrect responses. The reason is that actual observers may occasionally fail to

Fig. 11 Fitting the indecision model to the data in Fig. 8. **a** Fit of the indecision model in Eqs. 10a, 10b, 11a and 11b, with $c = 0$. **b** Fit of the indecision model in Eqs. 10a, 10b, 11a and 11b under the constraints $\theta_2 = \theta_1$ and $\beta_2 = \beta_1$. **c** Fit of the logistic functions in Eqs. 15 and 16 under the constraints $\alpha_2 = \alpha_1$ and $\gamma_2 = 1 - \alpha_1 - \gamma_1$. **d** Fit of the logistic functions in Eqs. 15 and 16 under the same constraints as those in panel c, plus the additional constraints $\theta_2 = \theta_1$

and $\beta_2 = \beta_1$. These additional constraints force the functions that apply under each presentation order to have the same location and slope, and, thus, they cannot accommodate the separate paths of the data from observer 3 for each presentation order, which are better described under the approaches illustrated in panels a–c. Nevertheless, none of the four approaches to fitting the data rendered a significant G^2 statistic for any of the observers

Table 4 Estimated parameters of the indecision model in Fig. 11a

Observer	λ	c	δ	κ	ξ_1	θ_1	θ_2	β_1	β_2
1	.001	0	5.215	.043	.282	−2.511	−2.471	0.112	0.107
2	.000	0	5.110	.455	.871	−2.408	−2.471	0.091	0.106
3	.017	0	4.586	.042	.086	−2.635	−2.695	0.149	0.138

c was actually set to 0 and not regarded as a free parameter

Table 5 Estimated parameters of the indecision model in Fig. 11b

Observer	λ	c	δ	κ	ξ_1	$\theta_1 = \theta_2$	$\beta_1 = \beta_2$
1	.001	0.127	5.223	.043	.290	−2.492	0.110
2	.000	−0.030	5.105	.452	.863	−2.452	0.102
3	.017	−0.442	4.377	.041	.093	−2.645	0.144

report indecision (i.e., behave with $\kappa > 0$) and produce “interval 1” or “interval 2” with a response bias given by their ξ_1 . In these conditions,

$$\Psi_1(x) = (1 - \lambda) \int_{-\infty}^{c-\delta} f(d; -\mu_1(x), 2) dd + \eta_1 \int_{c-\delta}^{c+\delta} f(d; -\mu_1(x), 2) dd + \lambda \int_{c+\delta}^{\infty} f(d; -\mu_1(x), 2) dd \quad (17a)$$

with $\eta_1 = \kappa \xi_1 + (1 - \kappa)/2$ and

$$\Psi_2(x) = (1 - \lambda) \int_{c+\delta}^{\infty} f(d; \mu_2(x), 2) dd + (1 - \eta_1) \int_{c-\delta}^{c+\delta} f(d; \mu_2(x), 2) dd + \lambda \int_{-\infty}^{c-\delta} f(d; \mu_2(x), 2) dd. \quad (17b)$$

Naturally, $\Upsilon_1(x) = \Upsilon_2(x) = 0$ here because undecided cases reported by the observer are evenly transformed into correct and incorrect responses by the external mechanism, yielding the term $(1 - \kappa)/2$ of parameter η_1 . Note also that $\xi_1 = .5$ (i.e., an observer without response bias) always renders $\Psi_1 = \Psi_2$ when $\mu_1 = \mu_2$ and $c = 0$, which was perhaps Fechner’s intuition.

In practice, and with the notation used thus far, fitting the model of Eqs. 17a and 17b under this strategy amounts to adjusting the counts of correct and wrong responses by defining $\tilde{C}_i^{(1)} = C_i^{(1)} + U_i^{(1)}/2$ and $\tilde{W}_j^{(1)} = W_j^{(1)} + U_i^{(1)}/2$ (and analogously defining $\tilde{C}_j^{(2)}$ and $\tilde{W}_j^{(2)}$) and then using $\tilde{C}_i^{(1)}, \tilde{W}_i^{(1)}, \tilde{C}_j^{(2)}$, and $\tilde{W}_j^{(2)}$ in Eq. 12 with Ψ_1 and Ψ_2 given by Eqs. 17a and 17b. This strategy transforms the data points across the panels of Fig. 11 into the adjusted proportions that are shown in Fig. 12a along with the fitted curves. The fit is also good and yields the estimated

parameters in Table 7 under the assumption that μ_1 and μ_2 differ, although Table 7 shows that the estimates of θ_1 and θ_2 (and also the estimates of β_1 and β_2) are very similar for each observer. Imposing the additional constraints $\theta_2 = \theta_1$ and $\beta_2 = \beta_1$ to check for the need to assume different psychophysical functions across intervals renders an almost equally good fit (see Fig. 12b and estimated parameters in Table 8). A comparison of Tables 7 and 8 again reveals that estimates of λ , δ , η_1 , θ_1 , and β_1 are virtually unchanged whether θ_2 and β_2 are free parameters (Table 7) or are constrained to have the same values as θ_1 and β_1 (Table 8) and that only parameter c needs to be altered to account for laterally shifted psychometric functions (see observer 3) under the more tenable assumption of equal sensitivities in each interval.

Should a single psychometric function be fitted to data aggregated across presentation orders?

None of the approaches just discussed fitted a psychometric function to data pooled across presentation orders. As was discussed earlier, such a psychometric function is only a by-product that takes the form of the average of the psychometric functions that hold for each presentation order. If the latter two differ for one reason or another, the parameters of a psychometric function fitted to pooled data may be misleading as illustrated next.

Consider an experiment collecting 2AFC data with the method of constant stimuli and under conventional instructions to respond either “interval 1” or “interval 2” on every trial. Consider also an observer who behaves according to the difference model with indecision with $\kappa = 1$ (as forced by the instructions), $\lambda = 0$ (i.e., no finger errors), $c = 0$ (i.e., no decisional bias), an interval of uncertainty of half-width $\delta = 2$, a moderately large response bias toward responding “interval 1” when undecided (i.e., $\xi_1 = .8$), and a single psychophysical function μ such as that in Fig. 1a (i.e., no

Table 6 Estimated parameters of the constrained logistic fit in Fig. 11c

Observer	λ	γ_1	γ_2	θ_1	θ_2	β_1	β_2
1	.001	.012	.029	−2.118	−2.098	14.797	14.764
2	.000	.399	.058	−2.093	−2.110	18.812	15.215
3	.017	.004	.039	−2.184	−2.279	10.783	12.038

Fig. 12 Fitting the indecision model to data in Fig. 8 transformed so as to count undecided cases as half right and half wrong. **a** Fit of the indecision model in Eqs. 17a and 17b with $c = 0$. **b** Fit of the

indecision model in Eqs. 17a and 17b under the constraints $\theta_2 = \theta_1$ and $\beta_2 = \beta_1$. Neither approach rendered a significant G^2 statistic for any of the observers

differences in sensitivity across intervals). Randomization of presentation order equalizes the overall number of trials on which the test is presented first or second, but it does not ensure equal numbers of trials with each order of presentation at each test level. Differences in these numbers may distort the pattern described by pooled data even when the data for each presentation order are noiseless, as can be seen in Fig. 13. Obviously, the method of constant stimuli could easily be set up so as to administer the same number of trials with each presentation order at each test level. Although this will solve the problem illustrated in Fig. 13, it will not prevent the psychometric function for pooled data from being potentially misleading for a reason discussed next—namely, that the breadth of the interval of uncertainty distorts the psychometric function for pooled data.

Figure 14a shows the decision space under the indecision model in the conditions of Fig. 6, except that $\delta = 2$; Fig. 14b is analogous, except that $\delta = 0.75$. Figure 14c shows, for an observer with response bias ($\xi_1 = .8$), the resultant psychometric functions in each case (dashed curves for $\delta = 2$; continuous curves for $\delta = 0.75$) separately for each presentation order (red and blue curves) and for pooled data (black curves). The sensory parameters θ and β

coming from the psychophysical function μ are the same in all cases, and it is only δ that varies. This change in δ produces psychometric functions that differ in their lower asymptote across presentation orders (red and blue curves), but, at the overall level of pooled data (black curves), the change in δ renders psychometric functions that differ in slope and location. These changes are not caused by the response bias of the observer. This is easily seen by noting that lack of response bias (i.e., $\xi_1 = .5$) renders psychometric functions that are identical for both presentation orders (and, thus, also for pooled data) when $\mu_1 = \mu_2$ and $c = 0$ (see Eqs. 7a and 7b), and yet, Fig. 14d shows that changes in δ still affect the slope and location of the psychometric function for pooled data. Note that this problem will not arise when the three-response format is used, which would result in distinctly different and identifiable curves when only δ varies, as illustrated in Fig. 7.

In contrast, conventional psychometric functions fitted to pooled data can only account for the differences shown in Fig. 14d through different parameters θ and β , and the interpretation would then be that sensitivity has increased when only the breadth of the interval of uncertainty has shrunk. This brings to mind the case of perceptual learning, since recent evidence (see Wenger,

Table 7 Estimated parameters of the indecision model in Fig. 12a

Observer	λ	c	δ	η_1	θ_1	θ_2	β_1	β_2
1	.004	0	3.418	.489	-2.344	-2.316	0.105	0.102
2	.004	0	3.210	.650	-2.301	-2.274	0.095	0.088
3	.000	0	3.791	.485	-2.520	-2.644	0.140	0.157

c was actually set to 0 and not regarded as a free parameter

Table 8 Estimated parameters of the indecision model in Fig. 12b

Observer	λ	c	δ	η_1	$\theta_1 = \theta_2$	$\beta_1 = \beta_2$
1	.004	0.132	3.472	.490	-2.335	0.104
2	.004	0.052	2.940	.654	-2.262	0.089
3	.000	-0.323	3.659	.480	-2.565	0.148

Copeland, Bittner, & Thomas, 2008; Wenger & Rasche, 2006) indicates that it has a strong component from variations in the response criterion. In the words of Law and Gold (2010), “at least some forms of perceptual learning do not appear to affect the response properties of neurons that represent the sensory evidence. Instead, improved perceptual performance results from changes in how the sensory evidence is selected and weighted to form the decision” (p. 226). It is likely that changes in the slope and location of psychometric functions for pooled data reflect variations in decisional/response factors, and not only (or not at all) in sensory parameters. Fitting psychometric functions through the indecision model under any of the approaches illustrated above would allow determining the extent to which variations in δ are responsible for (at least part of) the observed variations in performance, something that conventional analyses of pooled data cannot provide.

It is true that in some experiments, differences have been sought across stimulus conditions under circumstances in which variations in decisional factors may not be involved. But all things considered, we find no justification for fitting

Fig. 13 Noisy aspect of data pooled across presentation orders (black circles) despite the fact that data for each presentation order (red and blue circles) follow closely the path of the applicable psychometric functions (red and blue curves). Data are artificial and were generated under the indecision model in Eqs. 10a and 10b with $\lambda = 0$, $c = 0$, $\delta = 2$, $\kappa = 1$, $\xi_1 = .8$, and psychophysical functions $\mu_1 = \mu_2$ with parameters given in Fig. 1a. Black circles are always in between the red and blue circles at the same abscissa but pulled toward the color circle pertaining to the interval (first or second) in which the test was displayed more often (indicated by the numerals at the top)

(a) Decision space, $\delta = 2$; test first, test second **(b)** Decision space, $\delta = 0.75$; test first, test second **(c)** Psychometric functions when $\xi_1 = 0.8$ **(d)** Psychometric functions when $\xi_1 = 0.5$

Fig. 14 Effect of the width of the interval of uncertainty on the location and slope of the psychometric function for pooled data across presentation orders. **a** Decision space for parameter values as in Fig. 6, except that $\delta = 2$ here. **b** Decision space for identical conditions except that $\delta = 0.75$. **c** Psychometric functions for an observer with response bias ($\xi_1 = .8$) when the test is presented first (red curves), when presented second (blue curves), and for pooled data (black curves) when $\delta = 2$ (dashed curves) and when $\delta = 0.75$ (continuous curves). **d** Psychometric functions for an observer without response bias ($\xi_1 = .5$), which do not differ according to presentation order (and, thus, for pooled data) but vary whether $\delta = 2$ (dashed curve) or $\delta = 0.75$ (continuous curve). Narrowing the interval of uncertainty results in a psychometric function for pooled data that is steeper and shifted leftward. What is nothing but a change in decisional factors may thus be mistaken for a change in sensitivity

psychometric functions to pooled data when the alternative approaches discussed earlier are just as easy to implement, serving the extra goal of testing for differences in sensitivity across intervals (when parameters θ_1 and θ_2 or parameters β_1 and β_2 differ significantly). This will also be useful for determining whether changes in performance across conditions (e.g., improved performance as training progresses in studies of perceptual learning, or higher sensitivity in the presence of flankers in studies of lateral interactions) must be attributed to changes in the parameters describing the sensory process or, rather, to changes in decisional/response factors.

Recommendations for the design of 2AFC detection experiments

For years, psychophysicists (including the authors) have trusted that 2AFC tasks allow the influence of nonsensory factors to be eliminated or easily compensated for so that pure sensitivity measures can be obtained. Yet evidence of order effects indicates that decisional/response factors remain an integral part of the observed performance. Their contaminating influence can be eliminated only with recourse to a realistic model of the process by which the observable responses are produced. We have shown that the difference model with indecision is a suitable candidate, and we have also shown that recording undecided cases separately allows testing and fitting the model. In strict sense, the 2AFC task with an “I don’t know” option (which was dubbed *unforced choice task* by Kaernbach, 2001) is equivalent to a three-point rating scale, a format that has often been used under this alternative name and whose use has been advocated (Klein, 2007). It is interesting to note in this respect that Fechner (1860/1966) discussed both options (the conventional 2AFC format and the three-response format) upon describing how the method of right and wrong cases was applied in his weight-lifting (discrimination) experiments:

There are two ways in which the method may be carried out. According to the first procedure one decides which weight is heavier or lighter only after much lifting back and forth of the weighted containers. According to the second, one commits oneself firmly after each container is lifted once for comparison, or in case of doubt counts the undecided case as half right and half wrong. (p. 78)

Clearly, the first approach implies repeating each trial as many times as needed until instances of S_1 and S_2 are obtained that allow for an “interval 1” or an “interval 2” response; the second approach implies the three-response format followed by adjustment of the counts of correct and

wrong responses. Fechner’s (1860/1966) assessment of the two methods was also unhesitating:

In the beginning I always used the first procedure, but later I discarded all experiments done in that way and used the second course exclusively, having become convinced of its much greater advantages. It is not only that this course leads to greater uniformity than the first, but it is the only one which can also yield a basis for elimination and precise determination of the influences of time and position, which cause constant errors. Only with the second approach can these influences be pitted against one another properly. (p. 78)

Fechner’s observations seem to have passed unnoticed and, to the authors’ knowledge, the advantages that an explicit model for the three-response format offers for separating sensory and decisional/response factors have not been acknowledged or exploited in SDT analyses beyond what ROC curves allow in the context of rating scale tasks.

Besides using the three-response format and fitting psychometric functions through one of the approaches illustrated in Figs. 10, 11 and 12, other suggestions for the design of 2AFC experiments emerge from our discussion of order effects. The ensuing recommendations are itemized next.

Inclusion of catch trials

By analogy with yes–no tasks, a catch trial in a 2AFC task is one in which both intervals display a blank. For lack of stimuli, the psychophysical function is not involved in these trials, and responses on them become direct indicators of the decision process used by the observer. Hence, they provide invaluable evidence for the estimation of parameters c and δ . Specifically, the proportions of “interval 1” and “interval 2” responses should be identical if $c = 0$, which provides for a simple test of decisional bias (see García-Pérez & Alcalá-Quintana, 2010a). Also, the overall proportion of undecided responses in catch trials is an indicator of the half-width δ of the interval of uncertainty. Of course, the picture is slightly corrupted if there are finger errors (i.e., when $\lambda > 0$) or when the observer fails to report all undecided cases (i.e., when $\kappa > 0$), but catch trials nevertheless represent a valuable source of data for an accurate estimation of the parameters of the indecision model. Hence, inclusion of a relatively large number of catch trials is advisable.

Use of adaptive psychophysical procedures

This article has focused on cases in which psychometric functions are fitted to data. In contrast, a nonnegligible number of psychophysical experiments are carried out

using adaptive methods (mostly up–down or Bayesian staircases) that seek to estimate detection thresholds defined as some percent point on the psychometric function. Therefore, it is worth saying a word about the implications of order effects for the adaptive estimation of percent points on a psychometric function.

Several identical staircases for the same experimental condition are usually interwoven along a session (or in repeat sessions), and presentation order is also randomly decided on each trial. But the properties of adaptive estimation methods have been documented only for the case in which the psychometric function is unique, something that is not true when order effects are present. Thus, stimulus levels are not optimally placed when they are selected on the basis of the observer's response on a trial in which the test was first (and, then, performance was governed by Ψ_1) and the next trial along that staircase will display the test second (so that performance on that trial will be governed by Ψ_2). It is unlikely that in these conditions, adaptive staircases will provide estimates of the same quality as when the psychometric function is unique. A simple way around this difficulty might seem to be to set up the interwoven staircases for a given condition so that half of them always display the test first and the other half always display the test second. Each staircase will thus track the (fixed) location of the chosen percent point on a single psychometric function, and the interweaving along the session will yield a sequence of trials on which order of presentation is effectively randomized. However, such design raises two issues, which we discuss next.

The first issue is that each type of staircase will provide a meaningfully different estimate. As Fig. 2 shows, Ψ_1 and Ψ_2 generally differ and will thus have their, say, 75% point at different locations. This causes some problems for an experimenter who seeks to find out just where the 75% point is: Its location depends on presentation order. The second issue is that the 75% point may simply not exist for one of the presentation orders. This is certainly true when order effects are as strong as those for observers 1 and 2 in Fig. 2. In these circumstances, the staircase will not converge (although it will eventually stop), and it will fail to provide a meaningful estimate.

Of course, neither of these problems implies that adaptive methods should be abandoned, but they both argue against their use for estimating percent points on a psychometric function. In the presence of order effects, a useful picture of performance on the task is provided only by the psychometric functions for each presentation order. This does not mean that experimenters should turn back to the inefficient method of constant stimuli, because adaptive methods can be used for the efficient estimation of psychometric functions, which solves the

problems discussed above. Watt and Andrews (1981) were perhaps the first to refer to an “adaptive method of constant stimuli” that deploys trials at stimulus levels where the observer's responses are most informative. García-Pérez and Alcalá-Quintana (2005) have shown that up–down staircases can be optimally set up to allow accurate estimation of psychometric functions using about the same number of trials that will be needed to obtain only an estimate of threshold as the average of reversals across a few replicated staircases. Some of the data sets in the [Supplementary Material](#) were collected in this manner with the actual purpose of fitting psychometric functions to them.

Summary and conclusion

Reanalyses of 2AFC data from published studies show that performance typically differs greatly with presentation order, something that refutes the standard difference model of SDT. The characteristics of these order effects are inconsistent with differences in sensitivity across intervals and with sensation weighting, but they are compatible with either decisional bias or response bias when undecided. Analysis of data sets in which observers were allowed to express indecision suggests that response bias when undecided is the major cause of order effects.

The indecision model incorporates Fechner's (1860/1966) interval of uncertainty, and we have discussed how 2AFC detection data should be used under this model to fit psychometric functions and estimate sensory and non-sensory parameters. The indecision model allows for differences in sensory parameters across intervals, and our illustrations have shown how the model can be fitted so as to actually check for these differences. This particular issue must be considered in greater detail, but a thorough treatment is beyond the scope of this article. In any case, the principle of parsimony suggests that differences in sensitivity across the two temporal intervals of a 2AFC trial should not be claimed unless the data cannot be accounted for otherwise, particularly because there is no a priori reason that these differences should exist.⁷

⁷ The same cannot be said of spatial 2AFC detection tasks, where differences in sensitivity across spatial locations could easily be understood as true differences caused by regional variations in sensitivity across the visual field (García-Pérez & Sierra-Vázquez, 1996) or as artifactual differences caused by the nonuniform rendition of luminance across locations on a monitor (García-Pérez & Peli, 2001). Also, differences in sensitivity across the intervals of a temporal 2AFC discrimination task are often the result of desensitization caused by temporary adaptation or aftereffects produced by the high-magnitude stimulation that is inherent to discrimination trials (Alcalá-Quintana & García-Pérez, 2011).

Most psychophysical research aims at determining sensory differences across conditions, but these are hard to establish if there are differences in performance across presentation orders within a given condition. These sensory differences across conditions are also hard to establish if the various conditions affect decisional/response factors, and not only sensory parameters, because a conventional analysis of 2AFC data will inescapably attribute any observed differences in performance to differences in sensory parameters. Generally speaking, order effects are only the observable manifestation of a potentially serious contamination of 2AFC data, and, hence, the best way around it is to design 2AFC experiments so that the source of order effects can be elucidated and its effects on parameter estimates removed. The indecision model and the three-response format stand out as useful tools for accomplishing this goal.

We must finally stress that the order effects reported here are also likely to occur in experiments for which the label *2AFC detection task* is not appropriate but for which the standard difference model and its amendments certainly apply. Besides 2AFC discrimination tasks for which order effects have already been described (see Alcalá-Quintana & García-Pérez, 2011), these other experiments involve single-presentation identification tasks in which each trial displays a single stimulus and the observer is asked to report, for example, whether the stimulus was horizontally or vertically oriented (Peli, Arend, Young, & Goldstein, 1993) or whether it drifted to the right or to the left (García-Pérez & Peli, 1999). Unfortunately, we have been unable to locate published data sets for which presentation order (loosely speaking) was recorded in the experiments. But the three-response format and the indecision model are formally and technically appropriate in these cases, where they are also potentially useful to separate sensory and decisional/response factors.

Acknowledgments This research was supported by Grant PSI2009-08800 from Ministerio de Ciencia e Innovación (Spain). We thank Craig Abbey, Rob Giorgi, Robbe Goris, Ronald Hübner, Frank Jäkel, Bart Machilsen, Lotte Meteyard, Dov Sagi, Johan Wagemans, and Russell Woods for kindly making their data available for our reanalyses. We also thank the anonymous reviewers for their comments on earlier drafts of the manuscript.

References

- Alcalá-Quintana, R., & García-Pérez, M. A. (2011). A model for the time-order error in contrast discrimination. *Quarterly Journal of Experimental Psychology*, 64, 1221–1248. doi:10.1080/17470218.2010.540018
- Armstrong, L., & Marks, L. E. (1997). Differential effects of stimulus context on perceived length: Implications for the horizontal-vertical illusion. *Perception & Psychophysics*, 59, 1200–1213.
- Fechner, G. T. (1966). *Elements of psychophysics*. New York: Holt. (Original work published 1860).
- García-Pérez, M. A. (2000). Optimal setups for forced-choice staircases with fixed step sizes. *Spatial Vision*, 13, 431–448. doi:10.1163/156856800741306
- García-Pérez, M. A. (2002). Properties of some variants of adaptive staircases with fixed step sizes. *Spatial Vision*, 15, 303–321. doi:10.1163/15685680260174056
- García-Pérez, M. A., & Alcalá-Quintana, R. (2005). Sampling plans for fitting the psychometric function. *Spanish Journal of Psychology*, 8, 256–289 Retrieved from http://www.ucm.es/info/Psi/docs/journal/v8_n2_2005/art256.pdf
- García-Pérez, M. A., & Alcalá-Quintana, R. (2007). The transducer model for contrast detection and discrimination: Formal relations, implications, and an empirical test. *Spatial Vision*, 20, 5–43. doi:10.1163/156856807779369724
- García-Pérez, M. A., & Alcalá-Quintana, R. (2010a). The difference model with guessing explains interval bias in two-alternative forced-choice detection procedures. *Journal of Sensory Studies*, 25, 876–898. doi:10.1111/j.1745-459X.2010.00310.x
- García-Pérez, M. A., & Alcalá-Quintana, R. (2010b). Reminder and 2AFC tasks provide similar estimates of the difference limen: A reanalysis of data from Lapid, Ulrich, and Rammsayer (2008) and a discussion of Ulrich and Vorberg (2009). *Attention, Perception, & Psychophysics*, 72, 1155–1178. doi:10.3758/APP.72.4.1155
- García-Pérez, M. A., & Alcalá-Quintana, R. (2011). Improving the estimation of psychometric functions in 2AFC discrimination tasks. *Frontiers in Psychology: Quantitative Psychology and Measurement*, 2(96). doi:10.3389/fpsyg.2011.00096
- García-Pérez, M. A., & Alcalá-Quintana, R. (in press). Testing equivalence with repeated measures: Tests of the difference model of two-alternative forced-choice performance. *Spanish Journal of Psychology*. doi:10.5209/rev_SJOP.2011.v14.n2.48
- García-Pérez, M. A., & Núñez-Antón, V. (2004). Small-sample comparisons for goodness-of-fit statistics in one-way multinomials with composite hypotheses. *Journal of Applied Statistics*, 31, 161–181. doi:10.1080/0266476032000148849
- García-Pérez, M. A., & Peli, E. (1999). Imputation of direction of motion in one dimension. *Journal of the Optical Society of America A*, 16, 1531–1540. doi:10.1364/JOSAA.16.001531
- García-Pérez, M. A., & Peli, E. (2001). Luminance artifacts of cathode-ray tube displays for vision research. *Spatial Vision*, 14, 201–215. doi:10.1163/156856801300202931
- García-Pérez, M. A., & Sierra-Vázquez, V. (1996). Do channels shift their tuning towards lower spatial frequencies in the periphery? *Vision Research*, 36, 3339–3372. doi:10.1016/0042-6989(96)00345-8
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
- Hellström, Å. (1979). Time errors and differential sensation weighting. *Journal of Experimental Psychology: Human Perception and Performance*, 5, 460–477. doi:10.1037/0096-1523.5.3.460
- Jäkel, F., & Wichmann, F. A. (2006). Spatial four-alternative forced-choice method is the preferred psychophysical method for naïve observers. *Journal of Vision*, 6, 1307–1322. Retrieved from <http://journalofvision.org/6/11/13/>, doi:10.1167/6.11.13
- Kaernbach, C. (2001). Adaptive threshold estimation with unforced-choice tasks. *Perception & Psychophysics*, 63, 1377–1388.
- Klein, S. A. (2001). Measuring, estimating, and understanding the psychometric function: A commentary. *Perception & Psychophysics*, 63, 1421–1455.
- Klein, S. A. (2007). A local measure for modeling contrast discrimination: Response to Katkov, Tsodyks and Sagi. *Vision Research*, 47, 2912–2917. doi:10.1016/j.visres.2007.04.006
- Law, C.-T., & Gold, J. I. (2010). Shared mechanisms of perceptual learning and decision making. *Topics in Cognitive Science*, 2, 226–238. doi:10.1111/j.1756-8765.2009.01044.x

- Link, S. W. (1992). *The wave theory of difference and similarity*. Hillsdale, NJ: Erlbaum.
- Peli, E., Arend, L. E., Young, G. M., & Goldstein, R. B. (1993). Contrast sensitivity to patch stimuli: Effects of spatial bandwidth and temporal presentation. *Spatial Vision*, 7, 1–14. doi:[10.1163/156856893X00018](https://doi.org/10.1163/156856893X00018)
- Tune, G. S. (1964). Response preferences: A review of some relevant literature. *Psychological Bulletin*, 61, 286–302. doi:[10.1037/h0048618](https://doi.org/10.1037/h0048618)
- Watson, C. S., Kellogg, S. C., Kawanishi, D. T., & Lucas, P. A. (1973). The *uncertain* response in detection-oriented psychophysics. *Journal of Experimental Psychology*, 99, 180–185. doi:[10.1037/h0034736](https://doi.org/10.1037/h0034736)
- Watt, R. J., & Andrews, D. P. (1981). APE: Adaptive probit estimation of psychometric functions. *Current Psychological Reviews*, 1, 205–214. doi:[10.1007/BF02979265](https://doi.org/10.1007/BF02979265)
- Wenger, M. J., Copeland, A. M., Bittner, J. L., & Thomas, R. D. (2008). Evidence for criterion shifts in visual perceptual learning: Data and implications. *Perception & Psychophysics*, 70, 1248–1273. doi:[10.3758/PP.70.7.1248](https://doi.org/10.3758/PP.70.7.1248)
- Wenger, M. J., & Rasche, C. (2006). Perceptual learning in contrast detection: Presence and cost of shifts in response criteria. *Psychonomic Bulletin & Review*, 13, 656–661.
- Yeshurun, Y., Carrasco, M., & Maloney, L. T. (2008). Bias and sensitivity in two-interval forced choice procedures: Tests of the difference model. *Vision Research*, 48, 1837–1851. doi:[10.1016/j.visres.2008.05.008](https://doi.org/10.1016/j.visres.2008.05.008)