Order effects in two-alternative forced-choice tasks invalidate adaptive threshold estimates

Miguel A. García-Pérez¹ · Rocío Alcalá-Quintana¹

Published online: 30 March 2020
© The Psychonomic Society, Inc. 2020

Abstract

Adaptive psychophysical methods are widely used for the quick estimation of percentage points (thresholds) on psychometric functions for two-alternative forced-choice (2AFC) tasks. The use of adaptive methods is supported by numerous simulation studies documenting their performance, which have shown that thresholds can be reasonably estimated with them when their founding assumptions hold. One of these assumptions is that the psychometric function is invariant, but empirical evidence is mounting that human performance in 2AFC tasks needs to be described by two different psychometric functions, one that holds when the test stimulus is presented first in the 2AFC trial and a different one that holds when the test is presented second. The same holds when presentations are instead simultaneous at two spatial locations rather than sequential. We re-evaluated the performance of adaptive methods in the presence of these order effects via simulation studies and an empirical study with human observers. The simulation study showed that thresholds are severely overestimated by adaptive methods in these conditions, and the empirical study corroborated these findings. These results question the validity of threshold estimates obtained with adaptive methods that incorrectly assume the psychometric function to be invariant with presentation order. Alternative ways in which thresholds can be accurately estimated in the presence of order effects are discussed.

Keywords Psychophysics · Psychometric function · Adaptive methods · Threshold · Indecision model

One of the oldest-known and most intriguing regularities of psychophysical performance is what Fechner (1966, p. 75–77) called the “constant error”, by which the outcome of the perceptual comparison of two stimuli varies with their order of presentation. In Fechner’s studies, the size of the constant error reportedly varied across sessions for the same observers and he declared their general occurrence to be “totally unexpected” (p. 76). The cause of constant errors was unknown and early psychophysical studies on heaviness discrimination were always conducted with the standard weight lifted first in each comparison (e.g., Urban, 1908; Fernberger, 1913, 1914a, b, 1916, 1920, 1921). Constant errors arise in sequential presentations (e.g., lifting one weight and then the other) and in simultaneous presentations when these are possible (e.g.,

displaying two visual stimuli side by side) but here we will use the language of sequential presentations without loss of generality (i.e., we will refer to presentation intervals instead of presentation locations). The constant error was later called “time error” (e.g., Stewart, 1914), “stimulus error” (e.g., Fernberger, 1921), “time-order error” (e.g., Stott, 1935; Woodrow, 1935), “interval bias” (e.g., Green & Swets, 1966, pp. 408–411), or, in simultaneous presentations, “space-order error” (e.g., Hellström, 1985).

Constant errors are currently ignored and experimental protocols randomize presentation order across trials to characterize performance by a single outcome measure. In detection tasks, where one of the presentations is a blank in all trials and the other is a test stimulus of selected magnitude in each trial, the outcome measure is the proportion of correct responses (i.e., selecting the test) at each test magnitude. In discrimination tasks, where one of the presentations displays a standard stimulus of a fixed magnitude in all trials while the other displays a test stimulus of selected magnitude in each trial, the outcome measure is the proportion of trials in which the observer reports the test to have higher magnitude than the standard at each test magnitude. (Incidentally, detection involves discrimination from a blank and, thus, henceforth we

Electronic supplementary material The online version of this article (<https://doi.org/10.3758/s13428-020-01384-6>) contains supplementary material, which is available to authorized users.

✉ Miguel A. García-Pérez
miguel@psi.ucm.es

¹ Departamento de Metodología, Facultad de Psicología, Universidad Complutense, Campus de Somosaguas, 28223 Madrid, Spain

will always refer to test and standard stimuli on the understanding that the latter is a blank in detection tasks.) Ignoring presentation order upon computing these proportions produces a single estimate of the desired performance measure (i.e., detection threshold, discrimination threshold, etc.) in which the effects of presentation order are supposedly eliminated (e.g., Mayer, Di Luca, & Ernst, 2014; Nesti, Barnett-Cowan, MacNeilage, & Bühlhoff, 2014).

Data are sometimes collected with the method of constant stimuli to fit a psychometric function (Urban, 1908, 1910) whose location (detection threshold, discrimination threshold, or point of subjective equality) or slope is extracted for subsequent analyses. In other cases, adaptive methods (Leek, 2001; Treutwein, 1995) are used for efficient tracking of selected percentage points on the psychometric function without the burden of measuring and fitting it in full. Both practices appear inadequate when used without consideration of presentation order. The reason is that order effects violate their founding assumption, namely, that the probability of the target response depends only on the level of the test stimulus. Order effects imply that performance is governed by two psychometric functions, one that holds for trials in which the test is presented first and another one that holds for trials in which it is presented second. These two psychometric functions are clearly apparent when data are plotted separately for each presentation order (see, e.g., Alcalá-Quintana & García-Pérez, 2011; Bausenhardt, Dyjas, & Ulrich, 2015; Cai & Eagleman, 2015; Dyjas & Ulrich, 2014; Dyjas, Bausenhardt, & Ulrich, 2012, 2014; Ellinghaus, Gick, Ulrich, & Bausenhardt, 2019; García-Pérez & Alcalá-Quintana, 2011a, b; García-Pérez & Peli, 2014, 2015, 2019; García-Pérez, Giorgi, Woods, & Peli, 2005; Nachmias, 2006; Self, Mookhoek, Tjalma, & Roelfsema, 2015; Ulrich & Vorberg, 2009; von Castell, Hecht, & Oberfeld, 2017). These order effects can be accounted for and their contaminating influence eliminated when separate psychometric functions are fitted for each presentation order. Several competing models have been proposed to account for order effects on alternative theoretical grounds (for a review, see García-Pérez & Alcalá-Quintana, 2019). However, an important question that remains in this context is how order effects affect threshold estimates obtained with adaptive methods and whether this influence can be eliminated.

Many simulation studies have documented adequate performance of adaptive methods, but all of them generated data from a single psychometric function on all trials and presentation order was not even considered. When psychometric functions differ with presentation order, a crucial assumption of these methods is violated, namely, that the probabilities of responses depend only on the level of the test stimulus. The consequences of this violation are hard to work out, but they surely make estimated thresholds devoid of meaning. This seems to imply that two adaptive methods should be

interwoven, one that seeks threshold for test presentations in the first interval and the other for test presentations in the second. Although this has been done for studying order effects (e.g., Ellinghaus, Ulrich, & Bausenhardt, 2018), when the interest lies in how threshold varies across experimental conditions, the duality raises the question of which of the thresholds thus estimated is the true one (if any).

The main goal of this paper is to investigate the implications of order effects on the outcomes of adaptive methods set up conventionally, that is, with randomization of the order of presentation of test and standard stimuli under the assumption that the psychometric function is invariant with presentation order. This issue is addressed with complementary simulation and empirical studies.

In the simulation study, data were generated by the indecision model (García-Pérez & Alcalá-Quintana, 2017), which produces order effects as the combined result of a decisional bias and a response bias towards one of the response options upon guessing when undecided. Conditions covered several forms and amounts of indecision, including a lack thereof that renders the absence of order effects usually assumed in simulations. The main goal of the study was to determine the extent to which threshold estimates vary with the magnitude of order effects relative to the true threshold in an equivalent scenario without order effects. A prevalent adaptive method was used, namely, up-down staircases, as there are reasons to think that other adaptive methods (e.g., Bayesian staircases) will be similarly affected.

In the empirical study, discrimination thresholds for length were measured with the up-down method used in the simulations and, concurrently in the same sessions, the full psychometric functions for each presentation order were also measured via sampling plans optimally configured for this purpose. Psychometric functions thus provide estimates of the sensory, decisional, and response processes responsible for observed performance under adaptive estimation and the comparison should match expectations from the simulation study.

The organization of the paper is as follows. The next section describes the indecision model briefly. The simulation and empirical studies are subsequently described and discussed in consecutive sections. Finally, general conclusions are drawn, and recommendations are given regarding the use of adaptive methods for quick threshold estimation. We anticipate here that the results advise against the use of adaptive methods in their current form when the goal is to estimate a threshold that describes isolated sensory processing.

The indecision model and order effects

The core of the indecision model is that an observer asked to compare the magnitudes of stimuli A and B always makes one of three judgments: A subjectively greater than B, B

subjectively greater than A, or A and B subjectively equal. Consideration of these three judgment categories in psychophysical research dates back at least to Hegelmaier (1852; for an English translation see Laming & Laming, 1992). The ternary format was always used by Fechner (1966) and continued in use throughout the first half of the 20th century although responses were immediately classified as right or wrong relative to the physical reality of the stimuli, including a dichotomization that counted undecided responses as half right and half wrong. Use of the ternary format was discontinued when signal detection theory (SDT) set in in the 1960s, as it denies indecision and posits that observers will always judge one of the magnitudes to be subjectively greater than the other (but see Greenberg, 1965; Olson & Ogilvie, 1972; Saito, 1994; Treisman & Watts, 1966; Vickers, 1975; Watson, Kellogg, Kawanishi, & Lucas, 1973). As a result, psychophysical practice past the mid-20th century has almost exclusively used the binary format without allowance for undecided responses although observers are routinely admonished to guess when undecided (presumably under the implicit beliefs that they guess with equiprobability in such cases and that this is nevertheless inconsequential). The indecision model resolves the contradiction of admitting that observers may be undecided while measuring performance with methods justifiable only if indecision never happens.

The indecision model shares the assumptions of SDT except for a partition of decision space into three regions, even when the response format excludes undecided responses. This decision space makes observers undecided with some probability, which is the occasion for them to guess when this response is excluded. Detailed formal descriptions of the indecision model are available elsewhere and will not be repeated here (see, e.g., Alcalá-Quintana & García-Pérez, 2011; García-Pérez, 2014b; García-Pérez & Alcalá-Quintana, 2010a, b, 2011a, b, 2013, 2017, 2018, 2019; García-Pérez & Peli, 2014, 2015). Nevertheless, a brief presentation in the context of the problem addressed in this paper will be useful.

As in SDT, the perceived magnitude S of a stimulus with physical magnitude x is a random variable with mean $\mu(x)$ and variance $\sigma^2(x)$. We assume without loss of generality that distributions are normal, $\sigma^2(x) = 1$, and μ is the psychophysical function describing average perceived magnitude as a function of physical magnitude. We assume the function

$$\mu(x) = \log\left(1 + 2e^{(x-\alpha)/\beta}\right) \quad (1)$$

(for a discussion and justification of this particular form, see García-Pérez & Alcalá-Quintana, 2017). On a dual-presentation trial in which the test stimulus has magnitude x (variable across trials) and the standard stimulus has magnitude x_s (fixed on all trials), the decision variable D is the difference between the perceived magnitudes S_1 and S_2 of

the stimuli presented first and second, respectively, which are here assumed to be independent from one another as in SDT. Then, $D = S_2 - S_1$ is a normal random variable with variance 2 and mean $\mu(x_s) - \mu(x)$ when the test is presented first or $\mu(x) - \mu(x_s)$ when it is presented second. With a ternary response format, the observer judges that the first stimulus is subjectively higher when D is large and negative, that the second stimulus is subjectively higher when D is large and positive, and that both stimuli are subjectively equal when D is within some vicinity of 0. The decision space is thus partitioned into three regions with boundaries δ_1 and δ_2 (with $\delta_1 < \delta_2$). Note that the usual SDT model assumes $\delta_2 = \delta_1 = 0$ instead, leaving no room for indecision. Thus, the probabilities $p_{F,m}$, $p_{U,m}$, and $p_{S,m}$ of a “first higher” (F), “undecided” (U), or “second higher” (S) judgment when the test is presented in interval $m \in \{1, 2\}$ are given by

$$p_{F,1}(x) = \Phi\left(\frac{\delta_1 - \mu(x_s) + \mu(x)}{\sqrt{2}}\right) \quad (2a)$$

$$p_{U,1}(x) = \Phi\left(\frac{\delta_2 - \mu(x_s) + \mu(x)}{\sqrt{2}}\right) - \Phi\left(\frac{\delta_1 - \mu(x_s) + \mu(x)}{\sqrt{2}}\right) \quad (2b)$$

$$p_{S,1}(x) = 1 - \Phi\left(\frac{\delta_2 - \mu(x_s) + \mu(x)}{\sqrt{2}}\right) \quad (2c)$$

$$p_{F,2}(x) = \Phi\left(\frac{\delta_1 - \mu(x) + \mu(x_s)}{\sqrt{2}}\right) \quad (2d)$$

$$p_{U,2}(x) = \Phi\left(\frac{\delta_2 - \mu(x) + \mu(x_s)}{\sqrt{2}}\right) - \Phi\left(\frac{\delta_1 - \mu(x) + \mu(x_s)}{\sqrt{2}}\right) \quad (2e)$$

$$p_{S,2}(x) = 1 - \Phi\left(\frac{\delta_2 - \mu(x) + \mu(x_s)}{\sqrt{2}}\right) \quad (2f)$$

where Φ is the unit-normal cumulative distribution function. Judgments may be misreported due to response errors (see García-Pérez & Alcalá-Quintana, 2012a, 2012b, 2017) but otherwise they translate directly into responses and, then, Eqs. 2 are also the psychometric functions $\Psi_{F,m}$, $\Psi_{U,m}$, and $\Psi_{S,m}$ for F, U, and S responses when the test is presented in interval m . Note that $\delta_2 = -\delta_1$ makes $p_{F,1}(x) = p_{S,2}(x)$, $p_{F,2}(x) = p_{S,1}(x)$, and $p_{U,1}(x) = p_{U,2}(x)$ so that indecision does not produce order effects, as the psychometric functions are then identical for both presentation orders. A particular case arises when $\delta_2 = \delta_1 = 0$, whereby lack of indecision additionally makes $p_{U,1}(x) = p_{U,2}(x) = 0$ as in the conventional model.

When U responses are not allowed under the binary response format, observers are forced to misreport U judgments as F or S responses at random. Let ξ capture response bias as the probability with which an observer gives F responses upon U judgments. Without response errors, the psychometric functions for F and S responses under this response format become

$$\Psi_{F,1}^*(x) = p_{F,1}(x) + \xi p_{U,1}(x) \quad (3a)$$

$$\Psi_{S,1}^*(x) = p_{S,1}(x) + (1-\xi) p_{U,1}(x) \quad (3b)$$

$$\Psi_{F,2}^*(x) = p_{F,2}(x) + \xi p_{U,2}(x) \quad (3c)$$

$$\Psi_{S,2}^*(x) = p_{S,2}(x) + (1-\xi) p_{U,2}(x). \quad (3d)$$

Thus, the psychometric functions for F and S responses under binary responding reflect a mixture of authentic judgments and misreports (guesses). Note that absence of indecision via $\delta_2 = \delta_1 = 0$ degenerates to the conventional model and makes $\Psi_{F,1}^*(x) = \Psi_{S,2}^*(x)$ so that the probability of a correct response is invariant with presentation order in these conditions, but not in general when $\delta_2 = \delta_1 \neq 0$. Incidentally, Fechner's (1966) strategy of counting undecided responses as half right and half wrong is formally equivalent to ensuring that observers guess with $\xi = 0.5$ when undecided.

Figures 1a and b illustrate these psychometric functions for detection and discrimination tasks, respectively, under the ternary response format and the binary format with $\xi = 0.25$. The illustration uses the psychophysical function of Eq. 1 with $\alpha = -1.05$ and $\beta = 0.08$ and a partition of decision space where $\delta_1 = -1.8$ and $\delta_2 = 1.2$. In fact, δ_1 and δ_2 need not be symmetrically placed with respect to $D = 0$ and empirical evidence overwhelmingly indicates that $\delta_2 \neq -\delta_1$ beyond estimation error virtually in all cases (García-Pérez & Peli, 2014, 2015, 2019; García-Pérez & Alcalá-Quintana, 2013, 2017; Self et al., 2015). The region $[\delta_1, \delta_2]$ is the perceptual analogue to what Fechner (1966, p. 63) called the *interval of uncertainty* and its asymmetric location is by itself responsible for the order effects seen in Fig. 1 in the form of psychometric functions that differ across presentation orders even when data are collected with the ternary response format. Psychometric functions under the binary response format also display order effects caused by an asymmetric

Fig. 1 Order effects in detection (a) and discrimination (b) tasks in the context of visual contrast perception. The upper panel in each part plots the psychophysical function in Eq. 1 with $\alpha = -1.05$ and $\beta = 0.08$. A solid vertical line indicates the level of the test stimulus ($x = -1.021$ for detection and $x = -0.607$ for discrimination) used in the panels immediately underneath. Dashed lines in b indicate the level of the standard stimulus ($x_s = -0.7$) in the discrimination task. The left column at the bottom in each part shows the distribution of D when a test stimulus with the level indicated by the solid lines in the upper row is presented first (top panel) or second (bottom panel), also illustrating the partition arising from $\delta_1 = -1.8$ and $\delta_2 = 1.2$ with identification of the judgment (F, U, or S) associated with each region. The colored regions under each distribution indicate the probability of the corresponding judgment at this particular test level, which are also the probabilities of the corresponding responses in the absence of response errors. Probabilities at other test levels are analogously obtained from distributions whose

location is displaced as determined by the psychophysical function. The right column in each part shows psychometric functions under each presentation order for each of the possible responses under the ternary format (Eqs. 2; top panel) and for correct responses in the detection task or “test higher” responses in the discrimination task (i.e., F when the test is first and S when the test is second) under the binary response format with $\xi = 0.25$ (Eqs. 3a and 3d; bottom panel). The color code is identical to that used in the left column. The solid vertical line in each panel indicates the test level for which the distributions of D were illustrated in the left column and its crossings with the psychometric functions indicate the area of the corresponding region in that illustration. The dashed vertical line in panels for the discrimination task indicates the standard level. Note that, in all cases, the probability of any given response differs when the test is presented first or second. The solid horizontal line in panels for the binary format indicate the percentage level associated with a definition of threshold (see the main text for discussion)

location of the interval of perceptual uncertainty but these are further enhanced if there is a bias towards one of the response options upon guessing when undecided (i.e., if $\xi \neq 0.5$).

The shape of the psychometric functions changes for different values of the non-sensory parameters δ_1 , δ_2 , and ξ , even when the psychophysical function remains unchanged. With the binary format, the distance between the lower asymptotes in psychometric functions for detection or the lateral shift of psychometric functions for discrimination across presentation orders varies with δ_1 , δ_2 , and ξ . These influences do not cancel out (nor are they eliminated) via aggregation over presentation orders. In consequence, the location of a given percentage point on a psychometric function is not only determined by the sensory processes of interest but also by irrelevant influences from decisional and response processes.

The choice of test levels in Fig. 1 serves an additional purpose here. The design of adaptive methods for threshold estimation with a binary response format rests on the assumptions of lack of order effects and lack of response bias, which implies $\delta_2 = \delta_1 = 0$ in the indecision model (i.e., no guesses and a centered criterion). These assumptions underlie the definition of threshold as the test level at which performance reaches $P\%$ of target responses, as if the observer's responses were simply determined by whether $S_1 > S_2$ or $S_2 > S_1$. In SDT terms, performance on these two-alternative forced-choice (2AFC) tasks is modeled as shown in Fig. 1 but with a “symmetrical” criterion that implies $\delta_2 = \delta_1 = 0$ (Green & Swets, 1966, pp. 43–45; see Appendix A). Threshold is thus defined as the test level at which $\text{Prob}(S_t > S_s) = P/100$, where S_t and S_s are perceived magnitudes of test and standard, invariant with order of presentation. If this holds, a method that seeks the $P\%$ point on the psychometric function for detection will find it at $x = \mu^{-1}(\sqrt{2}\Phi^{-1}(P/100))$ whereas one that seeks the $P\%$ point on the psychometric function for discrimination will find it at $x = \mu^{-1}(\mu(x_s) + \sqrt{2}\Phi^{-1}(P/100))$. Test levels used in the illustration of Fig. 1 are those that result from these expressions with $P = 83.15$ in the detection task and $P = 79.38$ in the discrimination task, which are the percentage points tracked by the up–down method that will be used below. With order effects, Fig. 1 makes clear that the psychometric function for neither presentation order under the binary response format evaluates to $P/100$ at those test levels. Put another way, the 83.15% point on none of the two psychometric functions for detection (or the 79.83% point on none of the two psychometric functions for discrimination) is an estimate of threshold. It is also obvious that a psychometric function averaged over presentation orders will still not evaluate to $P/100$ at the test level satisfying the definition of threshold.

Simulation study

This study aims at determining how adaptive threshold estimates in detection and discrimination tasks vary with the magnitude of order effects caused by the width $\delta_2 - \delta_1$ and location $(\delta_1 + \delta_2)/2$ of the interval of perceptual uncertainty and by the response bias ξ . The reference for comparison is the true threshold in the ideal scenario assumed for the design of the adaptive method, which is captured by the indecision model when $\delta_2 = \delta_1 = 0$.

Design and conditions

We used the scenario in the illustration of Fig. 1, that is, a psychophysical function given by Eq. 1 with $\alpha = -1.05$ and $\beta = 0.08$ in the context of visual contrast perception, but this choice is inconsequential because the outcomes of dependable adaptive methods are hardly affected by the parameters of the true psychophysical function or, identically, by the slope and location of the psychometric function used to generate responses. For all replicates in all simulation conditions (see below), the sensory components of observed performance (i.e., the parameters of the psychophysical function and the variance of perceived magnitude) were fixed and so was the actual *sensory threshold* defined as the test level at which the percentage of *target judgments* (not to be confused with target responses) is P . Because estimates obtained in individual applications of an adaptive method are subject to sampling variability, distributions of estimates from 3000 replicates in each condition were used in the comparison of threshold estimates with the true threshold.

The simulation study comprised the factorial combination of four factors. The first factor was the type of task (detection or discrimination); the second factor was the center $c = (\delta_1 + \delta_2)/2$ of the interval of perceptual uncertainty, located from -1.6 to 1.6 in steps of 0.2 ; the third factor was the width $w = \delta_2 - \delta_1$ of the interval of perceptual uncertainty, ranging from 0 to 4 in steps of 0.4 ; and the fourth factor was the magnitude of response bias ξ , ranging from 0 to 1 in steps of 0.1 .

The adaptive method is the widely used 3–down/1–up staircase with steps up Δ^+ and steps down Δ^- in the optimal ratio (García-Pérez, 1998, 2011). Specifically, $\Delta^-/\Delta^+ = 0.7393$ makes the average-of-reversals estimator converge on the 83.15% point on the psychometric function for detection whereas $\Delta^-/\Delta^+ = 0.968$ makes it converge on the 79.38% point on the psychometric function for discrimination (see table I in García-Pérez, 2011). The size of the step down was $\Delta^- = 0.15$. The first trial administered an easy test level ($x = -0.5$ for detection and $x = -0.2$ for discrimination from $x_s = -0.7$), used the 1–down/1–up rule until the first reversal, switched to the 3–down/1–up rule for two more reversals, and continued unchanged for another 18 reversals. The stopping rule thus implied the usual completion of a fixed number of reversals and the threshold estimate was the average of these latter 18 reversals.

Data generation

For each replicate in each condition, a putative order of presentation of test and standard was first decided with equiprobability on each trial. Then, the response on that trial at the test level determined by the staircase rules was generated by the indecision model with the applicable values of δ_1 , δ_2 , and ξ . A judgment was first generated by a trinomial random process with the probabilities given by Eqs. 2a–2c or 2d–2f contingent on whether the test was nominally presented first or second, and note that this degenerates to a binomial process when $\delta_2 = \delta_1$. Note also that, for use in Eqs. 2, $\mu(x_s) = \mu(-0.7) = 5.074$ in the discrimination task whereas $\mu(x_s) = \mu(-\infty) = 0$ in the detection task (in an abuse of notation, given that the detection task implies a standard with 0% contrast whose logarithm goes to minus infinity). When the outcome of the trinomial was a U judgment, it was translated into an F or an S response with a Bernoulli process that produced F responses with probability ξ . Whether the result of authentic judgments or guesses, all F and S responses were expressed as correct (in detection tasks; ‘test higher’ in discrimination tasks) or incorrect (in detection tasks; ‘standard higher’ in discrimination tasks) as applicable to subsequently implement the up–down rule.

Results and discussion

Figure 2 presents detailed results under four illustrative conditions in the detection task. The leftmost panels show the decision space and the distribution of D when the test is presented first or second at level $x = -1.021$ so that, by Eq. 1, the mean of D is -1.358 when the test is first and 1.358 when it is second. The center panels show the staircase track for one of the replicates and the threshold estimate obtained on this occasion. The rightmost panel shows the distribution of threshold estimates across 3000 replicates. Figure 2a confirms the proper performance of this adaptive method in the absence of order effects and with an unbiased criterion (i.e., $\delta_2 = \delta_1 = \delta = 0$). In this scenario, threshold estimates reflect sensory processes and the dual-presentation task is criterion free: The observer’s response unmistakably indicates the stimulus for which perceived magnitude was larger (i.e., whether $S_1 > S_2$ or $S_2 > S_1$). Because the assumptions of the adaptive method hold in this scenario, threshold estimates are tightly distributed around the true threshold.

Figure 2b shows results when there is still no indecision but the criterion is displaced (i.e., $\delta_2 = \delta_1 = \delta = 1.6$). Order effects arise here from a decisional bias whereby observers respond S only upon relatively strong evidence that the magnitude of the second stimulus is larger. This scenario violates the assumption that the response depends only on whether or not $S_1 > S_2$. The sensory process is the same as in Fig. 2a and renders $\text{Prob}(S_t > S_s) = 0.8315$ at $x = -1.021$. Yet, with a displaced

criterion, the probability of a correct response when the test is second is 0.4320 whereas that when the test is first is 0.9818 instead. The 2AFC task is no longer criterion free in this scenario. The consequence on performance of the up–down method can be seen in the rightmost panel of Fig. 2b: The distribution of estimates shifts away from the true threshold. The average estimate is -0.9385 , where the psychometric functions for correct responses in test-first and test-second presentations respectively evaluate to 0.9964 and 0.6653 by Eqs. 3a and 3d. The probability of a correct response averaged across presentation orders is thus 0.8308 so that an adaptive method with randomization of presentation order seems to track its defining target (i.e., $P = 83.15$) on the average psychometric function. Yet, threshold is misestimated in that -1.021 and not -0.9385 is the stimulus level at which $\text{Prob}(S_t > S_s) = 0.8315$.

Figure 2c shows a distinctly different scenario that is nevertheless functionally equivalent to that in Fig. 2b. An interval of perceptual uncertainty now exists that is centered (i.e., $\delta_2 = -\delta_1 = 1.6$) but F responses are never given when uncertain (i.e., $\xi = 0$). Thus, U judgments are always expressed as S responses and, for all purposes, the observer operates as if $\delta_2 = \delta_1 = 1.6$. Figure 2d illustrates a slightly different scenario also involving $\delta_2 = -\delta_1 = 1.6$ but for an observer who guesses evenly when undecided (i.e., $\xi = 0.5$). It should be stressed that $\xi = 0.5$ removes order effects if and only if $\delta_2 = -\delta_1$ as assumed here. Then, psychometric functions do not differ across presentation orders but the presence of guesses makes them shallower than they would have been with $\delta_2 = \delta_1 = 0$ and, hence, the distribution of threshold estimates is also displaced away from the true threshold.

These results already reveal that threshold estimates obtained with an adaptive method for binary responses are meaningless without knowledge of the decisional and response processes by which the observer operated. Only under the conditions in Fig. 2a do adaptive threshold estimates carry their presumed meaning, namely, the stimulus level at which $\text{Prob}(S_t > S_s) = 0.8315$; under the conditions of Figs. 2b–2d (and all others discussed below), adaptive threshold estimates are contaminated by decisional and response processes whose influence cannot be discerned because adaptive data do not inform of them. In actual practice, then, using adaptive methods only provides estimates whose resemblance to the actual threshold that is sought remains unknown. Thus, differences in threshold estimates across conditions in an experimental study are only indicative of *some* effect of the corresponding manipulations, with no possibility of determining whether they reflect actual differences in sensitivity (the usual interpretation) or instead differences in decisional or response processes whose influence was supposedly removed by the 2AFC format.

Full results for detection and discrimination are shown in Figs 3a and b, respectively. Distributions of estimates are

Fig. 2 Sample simulation results for a detection task in scenarios (rows) with parameters indicated at the top left of the center panel. In each row, the left panel shows the decision space and the distribution of D when a test at $x = -1.021$ is presented in the first or the second interval. The center panel shows the staircase track for one of the replicates. Each symbol indicates the interval in which the test was presented and the (ternary) judgment made by the observer with the color codes used in Fig. 1. These judgments were translated into binary responses as indicated in the text. Note that black or grey symbols (U judgments) can only occur in parts c and d, and these resulted in guessed F or S responses that were correct or incorrect at random. Thin vertical lines demarcate the three

parts of each run, with threshold estimated from the 18 reversals in the final part. Small tick marks on the right vertical axis identify the levels at which each of these reversals occurred and the horizontal line across the panel indicates the average-of-reversals estimate. The right panel shows the distribution of estimates across 3000 replicates, with a vertical line at the actual location of the threshold defined as the stimulus level $x_T = -1.021$ at which the probability of a correct judgment is 0.8315. The sketch at the top of each panel summarizes the distribution by its mean (symbol) and an interval spanning one standard deviation on each side of the mean

sketched as they were at the top of each panel in the rightmost column of Fig. 2. Note that results in Fig. 2 are also reproduced in Fig. 3a. The scenario in Fig. 2a involves $(c, w) = (0, 0)$ and ξ is irrelevant here; in Fig. 2b, $(c, w) = (1.6, 0)$ and ξ is also irrelevant; in Fig. 2c, $(c, w, \xi) = (0, 3.2, 0)$; and in Fig. 2d, $(c, w, \xi) = (0, 3.2, 0.5)$. Adequate performance would show in Fig. 3 in the form of symbols sitting on the horizontal line in each panel, whose ordinate is the actual threshold by the definition that $\text{Prob}(S_1 > S_2) = P/100$. Far from it, large and systematic departures from this expectation can be noticed and the pattern of misestimation is analogous in detection and discrimination tasks except for a seemingly smaller magnitude in the latter case.

First note that the average-of-reversals estimator meets expectations when $\delta_2 = \delta_1 = 0$ so that $w = 0$ and $c = 0$ (bottom-most symbol in the string at the central abscissa in each panel),

a condition of absence of order effects due to lack of indecision by which response bias ξ also does not play any role. As w increases while $c = 0$ (symbols further up at the central abscissa), order effects caused by indecision and the ensuing response bias render a progressively larger misestimation of threshold. Generally, the misestimation is increasingly larger as w increases for intervals whose center is progressively more displaced (i.e., towards either extreme of the horizontal axis). There are nevertheless values of c at which increasing values of w render threshold estimates that move towards the true threshold rather than away from it (e.g., rightmost end of the panels for $\xi = 0$ and leftmost end of the panels for $\xi = 1$). This is easily understandable as a consequence of extreme bias coupled with an interval of perceptual uncertainty of appropriate width and location. Consider the condition marked with an arrow in the top panel of Fig. 3a, in which $(c, w, \xi) = (1.6,$

Fig. 3 Results for detection (a) and discrimination (b) tasks across simulation conditions. Symbols indicate the average threshold estimate across 3000 replicates under the applicable value of response bias ξ (indicated in each panel), location $(\delta_1 + \delta_2)/2$ of the center of the interval of perceptual uncertainty (horizontal axis), and width $\delta_2 - \delta_1$ of the interval (string of vertical symbols at each abscissa in each panel, unmarked). Vertical lines through each symbol sketch the standard deviation of the distribution; superposition hampers their individual

identification but they are all virtually of the same size. A dashed curve joins results obtained when the width of the interval of perceptual uncertainty is null (i.e., $\delta_2 - \delta_1 = 0$); results for increasingly larger widths (up to 4 in steps of 0.4) progressively move away from the dashed curve. The horizontal line in each panel indicates the true location of threshold ($x_T = -1.021$ in the detection task and $x_T = -0.607$ in the discrimination task). An arrow in the upper left panel marks a condition discussed in the text

3.2, 0) so that $\delta_1 = 0$ and $\delta_2 = 3.2$. As discussed in the analogous example of Fig. 2c, $\xi = 0$ means that the observer never responds F when undecided so that the upper boundary at $\delta_2 = 3.2$ is inoperative. Then, an S response is always given when $D > \delta_1 = 0$ whereas an F response is always given when $D < \delta_1 = 0$, functionally degenerating to a scenario in which $\delta_2 = \delta_1 = 0$. Chances that such a coalition of parameters will occur in practice are slim, besides the fact that one would never be capable of identifying whether this was the case.

Figure 4 plots results in a form that clarifies the effect of response bias. Results are shown for the discrimination task only because results for detection tasks are analogous. Obviously, there is no room for response bias in the absence of an interval of perceptual uncertainty (i.e., $\delta_2 = \delta_1 = \delta$; Fig. 4a) and threshold is only increasingly overestimated as δ moves away from 0. The influence of response bias is

increasingly larger as the width of the interval of perceptual uncertainty increases (Figs. 4b and 4c), although threshold can be estimated without bias for selected combinations of c , w , and ξ that are impossible to identify in empirical practice.

These systematic changes in threshold estimates are easily noticeable on average across multiple replicates of each condition (symbols in Figs. 3 and 4) but the variability of estimates across replicates (error bars) is relatively large and overlap across conditions is considerable. In consequence, the visible trends in Fig. 4 as response bias changes may not be noticeable in individual estimates empirically obtained through manipulation of response bias, and the order of individual estimates for different values of response bias may differ from what Fig. 4 shows the order to be on average. Also, although symbols in Fig. 4 indicate that adaptive estimates are on average higher than the true threshold (solid

Fig. 4 Simulation results for discrimination tasks plotted as a function of the location $(\delta_1 + \delta_2)/2$ of the interval of perceptual uncertainty (horizontal axis) for intervals of width 0 (a), 2 (b), or 4 (c). The vertical string of symbols at each abscissa pertain to values of response bias ξ , ranging from 0 to 1 in steps of 0.1. A dashed curve connects results for $\xi = 0$; results for increasingly larger values of ξ move progressively further away from the dashed curve. The solid horizontal line indicates

the true threshold. The vertical scale on the left (common to all panels) presents results in terms of threshold estimates, as in Fig. 3; the vertical scale on the right (also common to all panels) expresses results in terms of the percentage by which estimates exceed the true distance between the standard level ($x_s = -0.7$) and the true threshold ($x_T = -0.607$) and, thus, the percentage by which the difference limen (or discrimination threshold) is overestimated

horizontal line), error bars extending below this line indicate that estimates can occasionally be lower than the true threshold due to sampling error.

The variability of estimates across replicates decreases as the number of reversals increases (see, e.g., Alcalá-Quintana & García-Pérez, 2007; García-Pérez, 1998, 2000, 2001, 2011) and, hence, the overlap of distributions of estimates in Figs. 3 and 4 should accordingly shrink. We repeated the simulations with stopping rules that required 36 or 72 reversals, which produced increasingly narrower distributions without any meaningful change in their means (results not shown). Thus, investing more time to collect data from more reversals does not help to produce estimates that are unaffected by decisional and response processes.

Our simulation results demonstrate systematic increases in threshold estimates obtained with adaptive methods under violation of the assumption that performance is governed by the same psychometric function irrespective of the order of presentation of test and standard stimuli in each trial. Figures 3 and 4 show that the magnitude of the effect is determined by the differences in psychometric functions across presentation orders, but the chance of an after-the-fact correction of this influence does not exist because adaptive threshold estimation methods do not collect the information that would be needed for this purpose. The results presented thus far were obtained for a common true value of threshold across conditions, which suffices to demonstrate the influence of the contaminating factors. In empirical practice, true thresholds will vary across observers and one may wonder whether there is still any usable trace of variations in true threshold across such contaminated threshold estimates. Although this seems unlikely in the light of results in Figs. 3 and 4, we conducted an additional simulation for the discrimination task in Fig. 4 with 3000 replicates and random parameter values. Specifically, differences in true threshold across replicates were created by randomly drawing the value of parameter β in Eq. 1 from a

uniform distribution on $[0.045, 0.125]$ whereas differences in psychometric functions for each presentation order across replicates were created by drawing the center of the interval of perceptual uncertainty at random from a uniform distribution on $[-2, 2]$, the width of the interval from a uniform distribution on $[1, 5]$, and the response bias ξ from a uniform distribution on $[0.1, 0.9]$. For the rest, the simulation proceeded exactly as described earlier for discrimination tasks and we investigated the relationship between true and estimated thresholds. Not surprisingly, the concordance correlation coefficient (Lin, 1989) was a meager 0.19.

In sum, the average-of-reversals estimator converges away from the true threshold in a systematic manner that reflects the influence of the decisional and response processes responsible for order effects. These influences cannot be identified in empirical use of adaptive methods because data do not inform of the decisional and response processes by which observers operate. As a result, threshold estimates obtained under the influence of these extraneous factors only bear an insignificant correlation with true thresholds. In contrast, collecting ternary dual-presentation data and analyzing them under the indecision model allows separating the sensory, decisional, and response processes underlying observed data to obtain accurate estimates of the parameters describing these processes and, in turn, accurate estimates of thresholds (see García-Pérez & Alcalá-Quintana, 2017). The empirical validity of the simulation results presented in this section is investigated next.

Empirical study

This study investigates the extent to which the preceding simulation results hold in empirical practice. For this purpose, threshold estimates obtained with adaptive methods need to be compared with their “true” counterparts, although the latter are never available in an empirical study and must be

estimated with a dependable procedure. Thus, on the one hand, adaptive 3–down/1–up staircases were used to estimate observers' thresholds with manipulation of response bias in separate runs. On the other hand, psychometric functions for each presentation order were fitted to data collected with the ternary response format, which allows estimating all the parameters of the indecision model from which an estimate of the “true” threshold can be obtained. (For evidence on the accuracy of thresholds estimated in this way, see García-Pérez & Alcalá-Quintana, 2017). Data were collected across several sessions but trials for either purpose were randomly intermixed so that observers could not possibly tell which responses were used for which purpose. Without loss of generality, the task involved discrimination of the length of horizontal lines. The protocol of the study was approved by the institutional ethics committee.

Observers

Twenty experienced psychophysical observers with normal or corrected-to-normal vision participated in the study. All of them signed a consent form prior to their participation and, except for the authors (observers #1 and #2), all were naïve as to the goals of the study. Authors who also act as observers are commonplace in psychophysical research because awareness of the goals of the study is immaterial when observers cannot willingly influence the results. In any case, our observer-by-observer analyses allow the identification of any unusual performance eventually occurring for author observers.

Apparatus and stimuli

Stimuli were presented on a Philips Brilliance 4K UltraHD BDM4350UC/00 LCD display, which has an image area of 94×53 cm, a resolution of 3840×2160 pixels, and a frame rate of 60 Hz. At a viewing distance of 120 cm, the image area subtended 42.8×24.9 deg of visual angle (dva). Stimuli were black horizontal lines (luminance: 0.5 cd/m^2) on a gray background (luminance: 58 cd/m^2) that covered the entire image area. Line width was 8 pixels (0.2 cm; 0.1 dva). The length of the standard line was 800 pixels (19.6 cm; 9.3 dva) and the length of the test line varied across trials.

Standard and test lines were displayed simultaneously on each trial, one on the left side and the other on the right side, until the observer's response was entered. Lines were displayed such that the right end of the left line and the left end of the right line were separated horizontally by 100 pixels (2.4 cm; 1.2 dva); in addition, the left line was 25 pixels (0.6 cm; 0.3 dva) above the center of the image area and the right line was 25 pixels below it. This whole configuration was positioned such that the center of the horizontal gap between the lines was at the center of the image area, which did not provide any clue to the observers given the large distance

between the outer end of each line and the nearest vertical edge of the monitor. Which line (standard or test) was displayed on the left was randomized on each trial under the constraints described in the next section. Incidentally, given the simultaneous presentation at two positions, our former designation of responses as F (first) and S (second) are hereafter replaced with L (left) and R (right).

All experimental events were controlled by a computer running custom MATLAB scripts that called psychtoolbox-3 (<http://psychtoolbox.org>) functions.

Procedure

Testing took place in a dark room and observers kept a viewing distance of 120 cm without head restraint. Observers took a practice session of self-determined length until they gained familiarity with the task and the response interface. Data were then collected in three sessions separated by short breaks whose duration was at each observer's discretion. Each session presented trials from a single up–down staircase randomly interwoven with trials deployed for estimating psychometric functions. Trials of either type were indistinguishable by their design and the observer's task was to indicate which of the two lines (left or right) was subjectively longer or else that both were subjectively equal. This ternary response format is necessary for fitting psychometric functions under the indecision model and its apparent inadequacy for (binary) up–down staircases served two important goals. One was to make trials of either type indistinguishable by the response format; the other was to manipulate response bias. In fact, the observer's response bias would be unknown with a binary response format, as they would guess when undecided without leaving any trace that this occurred. Rather than asking observers to vary their bias in different runs while relying blindly on their compliance, asking them to report indecision instead allows a precise implementation of response bias with pre-selected strength. Specifically, U responses to trials from up–down runs were immediately transformed into L or R responses at random with the probability of L being $\xi = 0, 0.5$, or 1 in separate sessions. Then, for all practical purposes, up–down runs used a binary response format with response bias under control (for an analogous implementation of this strategy, see Alcalá-Quintana & García-Pérez, 2011). The order of these sessions (which differed only in the implemented value of ξ in up–down runs) was counterbalanced across observers in a Latin square. The order of administration of the three sequences in the Latin square was newly randomized for each consecutive set of three observers and only one of the sequences (0.5–1–0) in the last Latin square was omitted due to the size of the sample.

Adaptive staircases were set up as described above for the simulations. They started at an easily discriminable test length (935 pixels; compared to the standard length of 800 pixels),

used the simple 1–down/1–up rule until the first reversal, and then switched to the 3–down/1–up rule for another 20 reversals of which only the last 18 were averaged to estimate threshold. The sizes of the steps down and up were, respectively, $\Delta^- = 30$ pixels and $\Delta^+ = \Delta^-/0.968 = 31$ pixels. The location (left or right) of test and standard was randomly determined on each trial by an equiprobable Bernoulli process. Up–down staircases were identical in these respects in all sessions. The only difference across sessions was the probability ξ with which U responses were transformed into L or R responses. The response on each trial was immediately scored as ‘test longer’ or ‘standard longer’ for application of the placement rule for the next trial. The total number of trials to termination ranged from 56 to 111 with a median of 77.

Data for estimating psychometric functions were collected across the three sessions in which up–down staircases ran concurrently. The length of the test line on each trial was selected via an adaptive method of constant stimuli (AMOCs) that estimates monotonic and nonmonotonic psychometric functions (García-Pérez, 2014a; García-Pérez & Alcalá-Quintana, 2005). AMOCs uses staircases with steps up and down of the same size to ensure repeated presentation of relevant stimulus levels. Each session deployed 176 trials randomly interwoven from sixteen 11-trial staircases. Eight of them governed test presentations on the left and the other eight governed test presentations on the right, and the two sets were otherwise identical. In each set, the starting test length was 710 pixels for four of the staircases and 890 pixels for the other four, and all of them proceeded with step sizes up and down of 20 pixels. Test length decreased by one step after a ‘test longer’ response (L when the test was on the left or R when it was on the right), increased by one step after a ‘standard longer’ response (R when the test was on the left or L when it was on the right), and increased or decreased by two steps at random with equiprobability after a U response. To prevent excursions outside the relevant range of test lengths due to potential key-press errors, implementation of these rules was constrained by hard bounds placed at a minimum length of 710 pixels and a maximum length of 890 pixels. Test lengths that would fall outside this range defaulted to the corresponding bound without otherwise altering the general placement rule. Across all 528 trials, this eventuality never occurred for six observers, once for seven observers, 2–6 times for five observers, and 12 and 20 times respectively for the remaining two observers.

With the setup of up–down staircases, test and standard were rarely (if ever) identical in length. Also, the starting point and step sizes used in AMOCs ensured that standard and test never had the same length. To prevent observers from seeking refuge in the U response, the instructions emphasized that the two lines had different length and that observers should try hard to tell which of the two lines was longer although they could sometimes be unable to tell. Observers were asked to

switch gaze between the lines to make their judgment. They then used a numeric keypad to give L, R, or U responses by pressing the left-arrow, right-arrow, or up-arrow keys, respectively. The session was self-paced and the next trial did not start until a response had been entered. Stimuli were removed immediately afterwards and the image area displayed background luminance for 1000 ms before stimuli for the next trial were displayed. The total duration of a session ranged between 9 and 28 min with a median of 13 min.

Data pre-processing and analysis

Data analyses were conducted at the individual level. Adaptive up–down runs provided three separate estimates of the threshold x_T at which the percentage of ‘test longer’ responses is 79.38%, one for each of the three implemented magnitudes of response bias. Each estimate was obtained as the average of the last 18 reversals in the corresponding staircase.

Because data for estimating psychometric functions were collected across three separate sessions, nonparametric tests of equality of psychometric functions across sessions were first conducted with the generalized Berry–Mielke statistic (see García-Pérez & Núñez-Antón, 2018). Although there is no reason to think that performance would differ across sessions, this is a safety check before aggregating data. Psychometric functions from the indecision model with $\sigma = 1$ were subsequently fitted to each observer’s aggregated data (from $3 \times 176 = 528$ trials) by searching for maximum-likelihood estimates of model parameters with a FORTRAN version of the software in García-Pérez and Alcalá-Quintana (2017), which runs substantially faster for a more efficient exploration of the parameter space. The fitted model allowed for error/lapse parameters as described in García-Pérez and Alcalá-Quintana (2017), which capture the eventuality that some judgments are misreported. Error parameters capturing misreports of U judgments are relevant in this context because they inform of any residual guessing behavior despite the allowance for U responses. As it turned out, these parameter values were large for only three observers (see below) whereas error parameters reflecting the probabilities of misreporting L or R judgments were negligible for all observers except two (see below). We also fitted the model without error parameters and confirmed that the fit was meaningfully poorer for the observers who showed signs of residual guessing by the previous analysis. We are thus reporting estimates from the analysis with error parameters.

Besides error parameters, fitting the model only amounts to estimating parameters β , δ_1 , and δ_2 for each observer, because the psychophysical function does not differ for test and standard and also because parameter α does not play any role in discrimination tasks and can be set to any convenient value (we set $\alpha = 500$; for a discussion of how parameter α cancels

out in discrimination tasks, see figure 1 in García-Pérez & Alcalá-Quintana, 2017). Goodness of fit was assessed with the loglikelihood-ratio statistic G^2 . Estimated parameters of the psychophysical function were used to estimate the threshold x_T at which $\text{Prob}(S_t > S_s) = 0.7938$ via $x_T = \mu^{-1}(\mu(800) + \sqrt{2}\Phi^{-1}(0.7938))$, as described in Appendix A.

This pre-processing renders for each observer four separate estimates of the threshold x_T . Three of them come from up-down staircases with randomization of presentation order and under different magnitudes of response bias. The fourth comes from psychometric functions fitted under the indecision model, which is a theoretically appropriate estimate of the true threshold. The subsequent comparison of these four estimates aims at identifying traces of the patterns revealed in the simulation study. Specifically, when estimates are plotted in the form of Fig. 4 using the center and width of the interval of perceptual uncertainty estimated for each observer, their vertical arrangement should show traces of the corresponding pattern described by simulation results. Nevertheless, it should be kept in mind that simulation results used the actual threshold (as opposed to an estimate obtained by fitting the indecision model) and average threshold estimates across 3000 replicates in each condition (as opposed to that obtained in a single run for each of the selected values of ξ). Hence, the pattern of results of this empirical study cannot be as neat as that shown for simulation results in Fig. 4.

Results and discussion

Figure 5 shows staircase tracks and threshold estimates for two observers in each condition of response bias; plots for all observers are shown in Supplementary Figure S1. Estimated threshold is indicated by the ordinate of the horizontal line and its value is printed at the top-right corner of each panel. Note that ‘test longer’ responses (dark and pale blue symbols) prevail above the estimated threshold, ‘standard longer’ responses (dark and pale red symbols) prevail below

the standard length of 800 pixels, and U responses (black and gray symbols, which were transformed into L or R responses as applicable under each condition) prevail around the standard length. Note also that threshold estimates increase as ξ increases from 0 to 1 in the top row, whereas they instead decrease with increasing ξ in the bottom row. This ordered pattern did not occur equally cleanly for all observers (see Supplementary Figure S1), not only because estimates from a single run in each condition are affected by sampling error but also because the presence and form of these trends is linked to the width and location of each observer’s interval of perceptual uncertainty (see below).

Regarding AMOCS data, tests of equality of psychometric functions supported the aggregation of data across sessions to estimate psychometric functions. Specifically, these tests used the generalized Berry–Mielke statistic for three populations (three sessions), three response categories (L, U, and R), and ten strata (the number of test lengths at which data were collected) separately for each presentation order. Thus, a total of $2 \times 20 = 40$ size-0.05 tests were conducted and equality of psychometric functions across sessions was rejected in only seven of them (17.5%). For the record, equality was rejected for observers #1 and #5 under both orders of presentation, for observers #8 and #18 under test-right presentations, and for observer #19 under test-second presentations. Decomposition of the Berry–Mielke statistic revealed that rejection was due to discrepancies in only one of the ten strata. Thus, AMOCS data were aggregated across sessions for all observers, but note that excluding these five observers does not alter the conclusions of our observer-by-observer analyses.

Figure 6 shows data and fitted functions for each observer. In general, the path of the data is well captured by the fitted psychometric functions. Size-0.05 goodness-of-fit tests rejected the fit for only two observers (#10 and #20; grayed panels in Fig. 6), which is within the rejection rate of a true null, and note that fitted functions do justice to the data also in these cases. For observer #10, misfit appears to be due to a few U responses in test-left presentations at short test lengths where R

Fig. 5 Staircase tracks and threshold estimates (ordinate of the *horizontal line* in each panel) for two observers (*rows*) under each condition of response bias (*columns*). Graphical conventions as in analogous plots in Fig. 2

Fig. 6 Data and fitted psychometric functions under each presentation order (see the legend at the top) for each observer. The solid vertical line in each panel indicates the standard length; the dashed vertical line is the estimated threshold x_T defined as the test length at which $\text{Prob}(S_L > S_S) =$

0.7938, which is not linked to any percentage point on any psychometric function. Grayed panels indicate that the fit was rejected, but note that the fitted functions do justice to the data despite this eventuality

responses are expected (see the dark red and black data points on the far left of the panel for this observer); for observer #20, misfit appears caused by a few misreports at large test lengths. A formal inquiry via residual analysis (see García-Pérez, Núñez-Antón, & Alcalá-Quintana, 2015) confirmed the validity of this surmise. Similar suspect misreports were given by observer #19 without causing misfit. Error parameters in the model flagged misreports of L or R judgments only for observers #10 and #19; misreport of U judgments (residual guessing) is discussed next.

All observers give U responses when test length is around the standard length (solid vertical line in each panel), although some give them frequently and over a broad range (e.g., #1 and #3) whereas others do it less frequently and over a narrower range (e.g., #2, #4, and #5). The latter three observers are those for whom the analysis flagged residual guessing, consistent with manifest characteristics of their data. Note that, for these observers, the ordinate of the estimated psychometric functions for U responses (black and gray curves) hardly ever exceeds 0.3 at test lengths closest to the standard (790 and 810 pixels) so that these observers seem prone to give L or R instead of U responses in these cases. These observers must have behaved identically on concurrent up-down runs and, then, they tampered with our manipulation of response bias because they implemented their own response bias. Yet, the overall response bias in up-down runs

for them has only shrunk linearly to the range from somewhere above zero to somewhere below one.

Regarding order effects, it is clearly apparent that many observers' data are somewhere from meaningfully to substantially displaced laterally across presentation orders (e.g., #6–#15 and #18), indicating intervals of perceptual uncertainty that are not centered. In contrast, other observers' data are virtually superimposed (e.g., #1, #4, #16, #17, and #20), indicative of centered intervals. Note also that, relative to data for test-left presentations (dark colors), those for test-right presentations (pale colors) are displaced rightwards in some cases (e.g., #6, #8, #9, #10, #12, #14, and #18) and leftwards in others (e.g., #7, #11, and #15), reflecting intervals of perceptual uncertainty whose center is accordingly displaced to the left or to the right of $D = 0$. Table 1 lists estimated model parameters, goodness-of-fit statistics, and p values across observers, along with the thresholds estimated from these psychometric functions and those obtained in each of the up-down runs.

Figure 7 plots threshold estimates against the center (top panel) and width (bottom panel) of each observer's estimated interval of perceptual uncertainty. In agreement with simulation results, estimates from up-down runs (circles) are virtually always higher than that obtained from psychometric functions (short horizontal dash) and the magnitude of overestimation varies across observers. Adaptive estimates for different values of ξ , are not always ordered as average results in

Table 1. Model parameters (β , δ_1 , and δ_2), fit measures (G^2 statistic with 25 degrees of freedom and p value), and threshold estimates for each observer. Threshold estimated from psychometric functions is denoted $x_T^{(P)}$ whereas those estimated in adaptive up–down runs are denoted

$x_T^{(\xi)}$, for $\xi \in \{0, 0.5, 1\}$. Error parameters are excluded from this list. Entries in bold type identify observers for whom the G^2 statistic was significant at the 5% level.

Obs.	β	δ_1	δ_2	G^2	p	$x_T^{(P)}$	$x_T^{(0)}$	$x_T^{(0.5)}$	$x_T^{(1)}$
1	11.13	−1.925	1.505	17.54	0.861	812.90	815.39	818.72	822.22
2	14.29	−0.842	0.336	18.77	0.808	816.56	815.72	819.94	818.22
3	15.94	−2.659	2.210	28.60	0.281	818.48	832.00	852.00	841.87
4	8.36	−2.193	2.361	12.60	0.981	809.69	823.78	818.22	820.78
5	15.92	−1.178	1.211	25.34	0.444	818.45	831.89	821.39	831.00
6	18.74	−1.807	2.682	33.33	0.123	821.72	830.33	847.67	870.44
7	24.58	−1.685	−0.216	28.34	0.293	828.49	872.61	849.22	820.78
8	15.46	−1.328	2.928	26.32	0.391	817.92	816.83	828.67	874.78
9	14.98	−0.289	3.742	25.85	0.415	817.36	828.44	843.44	868.22
10	22.08	−1.641	2.809	44.97	0.008	825.60	848.44	862.00	868.89
11	21.59	−2.916	0.038	23.29	0.561	825.03	878.94	843.83	835.33
12	35.47	−0.609	1.467	33.42	0.121	841.12	830.89	849.83	873.44
13	13.96	−0.423	1.292	10.16	0.996	816.18	814.06	817.56	844.56
14	16.61	−2.179	3.039	12.81	0.979	819.26	829.33	865.72	850.78
15	15.86	−5.024	2.210	20.73	0.707	818.38	874.78	875.67	849.44
16	19.54	−1.628	1.746	32.92	0.133	822.65	844.17	840.78	847.78
17	13.78	−2.013	2.073	22.46	0.609	815.97	817.67	841.44	828.17
18	15.74	−1.133	2.024	24.21	0.508	818.25	828.33	843.78	839.56
19	17.16	−1.723	1.363	36.09	0.070	819.89	843.44	830.94	871.11
20	20.59	−1.963	2.100	37.70	0.050	823.87	860.00	834.33	837.67

Fig. 4 indicated (i.e., $x_T^{(0)} > x_T^{(0.5)} > x_T^{(1)}$ when $c < 0$ and $x_T^{(0)} < x_T^{(0.5)} < x_T^{(1)}$ when $c > 0$), but this is understandable from the overlap of the distributions in Fig. 4 and the use of a single estimate per condition in this study. Yet, violations of the expected order are never seen when $|c| > 0.4$ because in these cases the distributions of estimates are sufficiently separated for $\xi \in \{0, 0.5, 1\}$ when $w \geq 2$ (see Fig. 4), as is the case for almost all observers (see the abscissae of data points in the bottom panel of Fig. 7). On the other hand, disorder of adaptive estimates is more frequent when $|c| < 0.4$ because in these cases the distributions of estimates across values of ξ are heavily superimposed (see Fig. 4).

It should also be noted that the estimated width of the interval of perceptual uncertainty exceeds 4 units for half of these observers (see the abscissae of data points in the bottom panel of Fig. 7). Simulation results were presented in Fig. 4 for widths up to only 4 units (see the rightmost panel in Fig. 4) and, then, the figures of percentage excess by which threshold is overestimated in empirical practice can be larger than what Fig. 4 indicated, namely, overestimates of the difference limen that are up to 250% larger than its true value. This is sufficiently well appreciated for the observer (#15) whose data are plotted at the rightmost abscissa in the bottom panel of Fig. 7. For this observer, $x_T^{(P)} = 818.38$ (see Table 1) whereas adaptive

threshold estimates vary with response bias between 849.44 and 875.67 (see Table 1). The difference limen is, thus, 18.38 pixels but adaptive threshold estimates place it somewhere between 49.44 and 75.67 or, in percentage terms, somewhere between 169% and 312% above its true value. Repeating these computations for the remaining observers reveals that order effects and response bias caused by the binary format substantially distort threshold estimates obtained with adaptive methods that ignore this reality.

We conducted a more stringent test of the validity of the indecision model as an integral account of psychophysical performance. Because AMOCS and adaptive up–down trials were intermixed and indistinguishable, the observers' approach must have been identical in both cases. Then, psychometric functions estimated from AMOCS data (with separate treatment of undecided responses) and adaptive threshold estimates (with manipulation of response bias when undecided) must reflect a common underlying reality. Specifically, if the psychometric functions estimated from AMOCS data describe the observers' sensory, decisional, and response processes, this description should predict within sampling error the thresholds estimated in up–down runs in which the same sensory and decisional processes participated but the response process was manipulated. The prediction is obtained in three steps. Firstly, the probability of U responses must be diverted

Fig. 7 Threshold estimates as a function of the center (*top panel*) and width (*bottom panel*) of the interval of perceptual uncertainty estimated from AMOCS data. The four threshold estimates for each observer are joined by a vertical dashed line for clarity. As indicated at the left in the bottom panel, the short horizontal dash is the estimate $x_T^{(P)}$ obtained from

toward L or R responses according to the manipulated value of ξ in each run, which calls for Eq. 3 above. Secondly, adaptive runs with randomization of presentation order track their target point on the average psychometric function Ψ_{TL} for ‘test longer’ responses across presentation orders, which is given by

$$\Psi_{TL}(x) = \frac{\Psi_{F,1}^*(x) + \Psi_{S,2}^*(x)}{2} = \frac{p_{F,1}(x) + \xi p_{U,1}(x) + p_{S,2}(x) + (1-\xi) p_{U,2}(x)}{2}, \quad (4)$$

where the functions on the right-hand sides are defined in Eqs. 2 and 3 and evaluated with the parameter values estimated for each observer from AMOCS data, and with ξ valued as it applies to the up–down run for which the prediction is made. Finally, the predicted threshold in the corresponding adaptive run is the 79.38% point on Ψ_{TL} . Note that predictions vary for each observer across values of ξ , despite the fact that all of them originate from the same psychometric functions in Eqs. 2 with the single set of parameter values estimated from AMOCS data.

Figure 8 shows scatter plots of these predicted thresholds against the actual estimates obtained in each run. Data points gather tightly around the diagonal identity line in all panels

the fitted psychometric functions whereas circles represent estimates $x_T^{(\xi)}$ from adaptive up–down runs in which response bias ξ was 0 (solid circles), 0.5 (gray circles), or 1 (open circles). Data from the two authors (observers #1 and #2) are identified with red symbols

and the agreement is indicated by the value of the concordance correlation coefficient ρ_c (Lin, 1989; see also Lin, Hedayat, Sinha, & Yang, 2002). The agreement is remarkable despite sampling error and on consideration that the simple average in Eq. 4 implies identical numbers of trials with test and standard in each of the two orders of presentation at each test level. Randomization of presentation order in actual adaptive runs does not ensure these conditions and this adds error to predictions under extreme bias (i.e., when $\xi = 0$ or $\xi = 1$). Yet, threshold estimates obtained in separate adaptive runs with manipulation of response bias are predicted reasonably accurately from each observer’s sensory, decisional, and response processes as these manifest in psychometric functions separately estimated to capture order effects.

In sum, these results corroborate the empirical validity of our simulation results. In practical terms, these results confirm that adaptive threshold estimates do not portray isolated sensory processes, as their value is heavily influenced by decisional and response processes whose operational characteristics are not identifiable in data collected with adaptive methods that only track percentage points on a putative psychometric function for binary responses with randomization of presentation order. In the presence of order effects, indecision, and response bias, adaptive methods track their target percentage point on the average of the psychometric functions

Fig. 8 Scatter plots of thresholds predicted from psychometric functions via Eq. 4 against those obtained in each of the adaptive runs with manipulation of response bias (*panels*). Data from the two authors

(observers #1 and #2) are identified with *red symbols*. The *dashed diagonal* in each panel is the identity line. The upper-left corner in each panel shows the value of the concordance correlation coefficient ρ_c

that hold for each presentation order, which does not offer a faithful portrait of the isolated sensory process of interest.

Discussion

The psychometric function expresses how the probability of a target response varies with stimulus level. Some psychophysical studies measure it in full but many others only estimate percentage points on it that are meaningful because they reflect landmarks such as detection or discrimination thresholds, or because they are useful for estimating its slope indirectly. Many adaptive methods have been proposed to achieve these goals efficiently and effortlessly so that quick estimates can be obtained for comparisons across experimental conditions in within- or between-subjects studies. Yet, the comparison is warranted as long as the stimulus level at which performance reaches some percentage of target responses relates to a common indicator of sensory processing. In turn, this requires assurance that the stimulus level at which the percentage of target responses is P can be unmistakably interpreted as that at which $\text{Prob}(S_t > S_s) = P/100$ despite the fact that the perceived magnitudes S_t and S_s are not observable and only the outcomes of a response process that uses them can be recorded. When responses depend exclusively on whether or not $S_t > S_s$, a link can be established between $\text{Prob}(S_t > S_s) = P/100$ and the observed percentage of target responses (see Appendix A). This link justifies the interpretation of percentage points as relating to a common indicator of sensory processing which is, in SDT terms, a fixed value of d' .

The assumptions required to establish such a link imply that the probability of the target response is a sole function of stimulus level. Yet, order effects ubiquitously found in 2AFC data reveal this assumption untenable: The psychometric function differs with the order of presentation of test and standard stimuli, thus shattering the framework under which observed responses on a psychophysical task can be linked to

unobservable perceived magnitudes. These order effects are a consequence of decisional and response processes conventionally ignored. Then, empirical use of adaptive methods for binary responses with randomization of presentation order across trials deprives the resultant estimates of the foundations that allow their interpretation as uncontaminated estimates of the operation of sensory processes.

These points were corroborated in our simulation and empirical studies. Simulation results indicated that extraneous ‘correct’ responses (in detection tasks) or ‘test longer’ responses (in discrimination tasks) given under a binary response format that forces observers to guess when undecided alter threshold estimates, which thus relate strongly to decisional and response processes and not just to sensory processes. The resultant estimate only indicates the stimulus level at which the percentage of target responses (with its multiple determinants) reaches some value, not the stimulus level at which the perceived magnitude of the test exceeds that of the standard with the stated probability. Our empirical study confirmed that adaptive threshold estimates vary with order effects and the magnitude of response bias as expected from simulation results, given the decisional processes separately estimated for each observer.

These results prove that threshold estimates obtained with adaptive methods under the binary response format are not valid indicators of sensory processing. This is true when these methods run with randomization of presentation order across trials, but also if presentation order is fixed. As Fig. 6 reveals, indecision is equally prevalent under both presentation orders (gray and black data points and curves) and, then, response bias upon guessing under binary responding still contributes to detaching observed responses from their sensory determinants. It should be noted that Kaernbach (2001) devised adaptive up–down methods for use with the ternary response format in which observers are allowed to express indecision. Yet, threshold continues to be estimated as a percentage point on a putative psychometric function for binary responses and the

width of the interval of perceptual uncertainty still affects these estimates, making them uninterpretable (see García-Pérez, 2002).

We also conducted a stringent test of the empirical validity of the indecision model used in our simulations, which consists of checking out that threshold estimates obtained in adaptive up–down runs with manipulation of response bias when undecided can be predicted from psychometric functions estimated to capture the natural operation of the observers' sensory, decisional, and response processes. The test supported the notion that adaptive threshold estimates do not reflect the isolated operation of sensory processes. Thus, adaptive threshold estimates obtained in empirical practice are meaningfully influenced by decisional and response processes that are supposedly eliminated by the 2AFC format. Unfortunately, disentangling these influences is impossible because data collected with adaptive threshold estimation methods do not inform of them. Estimating thresholds separately for each order of presentation via dual adaptive runs is also insufficient for this purpose because indecision and response bias still play a crucial role in the observed outcomes for each presentation order. In addition, obtaining a different threshold for each presentation order defies the notion that sensory processes act upon the displayed stimulus regardless of how many other stimuli may have ever been displayed or at what point in time.

Our empirical study used a single sensory modality (vision), type of stimulus (lines), physical dimension (length), task (discrimination), and adaptive method (3–down/1–up staircases), but there are reasons to think that our results will also hold for other choices in these respects. As regards other sensory modalities, types of stimuli, tasks, and physical dimensions, order effects have overwhelmingly been found across the board, implying that also in those cases the probability of the target response varies across presentation orders. Analogously, these results surely generalize to *m*AFC tasks (with $m > 2$) in which order effects have also been documented (e.g., Elfenbein, Small, & Davis, 1993; Johnson, Watson, & Kelly, 1984; Kim, Lee, & Lee, 2010). As regards other adaptive methods designed to track percentage points on an underlying psychometric function, the fact that the psychometric function differs across presentation orders also violates their founding assumption that stimulus level is the sole determinant of the probability of the target response.

Given that order effects in dual-presentation tasks are universal, the only sensible action in empirical practice seems to be to steer away from adaptive methods that cannot cope with them. This implies letting go of the illusory notion that thresholds can be quickly estimated in only a handful of trials, switching to methods that can effectively separate sensory, decisional, and response processes for uncontaminated estimation of the stimulus level at which $\text{Prob}(S_t > S_s) = P/100$. Fitting psychometric functions under the indecision model is a suitable approach to accomplishing the separation and user-

friendly software to accomplish this goal is freely available (García-Pérez & Alcalá-Quintana, 2017). This switch certainly requires a little extra work and burden on the observers but, as Brownlee (1957) replied to a complaint of this type, “that a valid experiment takes more work than an invalid experiment is irrelevant to a man who is wanting to make valid inferences” (quoted by Versace, 1967).

Open practices statement The data and materials for the empirical study are available from the corresponding author upon request. The study was not preregistered.

Acknowledgements This work was supported by grant PSI2015-67162-P from Ministerio de Economía y Competitividad.

Appendix

This appendix describes how percentage points on a psychometric function relate to conventional measures such as d' in SDT, the assumptions underlying this relation, how violation of these assumptions render uninterpretable percentage points, and how the ensuing difficulties can be circumvented.

Appendix Figure 9a sketches some of the SDT assumptions for a discrimination task. The perceived magnitude S_s of a standard stimulus with physical magnitude x_s is a normal random variable with mean $\mu(x_s)$ and variance σ^2 whereas the perceived magnitude S_t of a test stimulus with physical magnitude x_t is also a normal random variable with variance σ^2 but with mean $\mu(x_t)$ instead. In this illustration, x_s and x_t are such that $d' = \frac{\mu(x_t) - \mu(x_s)}{\sigma} = 1$ so that x_t is the threshold defined as the test level for which $d' = 1$. Naturally, threshold can be defined relative to a different d' (see, e.g., Klein, 2001, p. 1430), but this does not change anything meaningfully. A second assumption in SDT is that an observer's response in a 2AFC trial with instructions to report the stimulus with a higher (or lower) magnitude is determined by whether $S_t > S_s$ or $S_t < S_s$, which implies the decision variable $D = S_t - S_s$ and a decision boundary at $D = \delta$, with $\delta = 0$. In these conditions, D is a normally distributed random variable with mean $\mu(x_t) - \mu(x_s)$ and variance $2\sigma^2$ so that $\text{Prob}(S_t > S_s) = \text{Prob}(D > 0) = 0.7602$ when $d' = 1$ (Appendix Fig. 9b). Then, an adaptive method that seeks the 76.02% point on the psychometric function for “test higher” responses provides an estimate of the stimulus level for which $d' = 1$. Adaptive methods do not exist that estimate any desired percentage point accurately and efficiently (see García-Pérez, 1998, 2011; García-Pérez & Alcalá-Quintana, 2007). Then, the definition of threshold usually goes in reverse as the stimulus level whose d' is associated with the $P\%$ point that a given adaptive method can estimate. Note that $\text{Prob}(D > 0)$ implies $\text{Prob}\left(Z > \frac{0 - (\mu(x_t) - \mu(x_s))}{\sqrt{2}\sigma}\right) = \text{Prob}\left(Z > \frac{-d'}{\sqrt{2}}\right)$, where Z is the standard normal random variable. Hence,

Fig. 9 **a** Distributions of perceived magnitude S for a standard stimulus (gray curve) and a test stimulus (black curve) whose physical magnitudes are such that $d' = 1$. **b** Decision space assumed in conventional SDT, which is also that assumed for the design of adaptive methods that seek to estimate the stimulus level at which $\text{Prob}(S_t > S_s) = P/100$. The curve shows the distribution of the difference between the perceived

magnitudes of test and standard stimuli in **a** and, thus, the grayed area is the value of P associated with a test stimulus for which $d' = 1$. **c** Realistic decision space identified in empirical studies and captured by the indecision model, where the operating boundaries δ_1 and δ_2 are typically displaced away from 0 to different extents in each direction

$\text{Prob}(D > 0) = P/100$ implies $\Phi\left(\frac{d'}{\sqrt{2}}\right) = P/100$, from which $d' = \sqrt{2}\Phi^{-1}(P/100)$. Order of presentation of standard and test is immaterial when $\delta = 0$ and, then, any percentage point that an adaptive method can track is associated with a fixed and known value of d' in these conditions.

If $\delta \neq 0$ under otherwise identical conditions, the one-to-one relation between P and d' breaks down. In such a case, $\text{Prob}(D > \delta) \neq \text{Prob}(S_t > S_s)$ and $\text{Prob}(D > \delta) = P/100$ implies $\Phi\left(\frac{d' - \delta}{\sqrt{2}\sigma}\right) = P/100$, from which d' cannot be extracted without knowledge of the value of δ . Although this knowledge (which the adaptive method does not provide) might allow determining the d' associated with the estimate, variations in the value of δ across observers or experimental conditions would make threshold estimates incomparable because they would relate to different values of d' in each case. The only way around this difficulty is via an alternative procedure that recovers the test level at which $\text{Prob}(S_t > S_s) = P/100$ irrespective of the value of δ . This is impossible when data are collected with an adaptive method that estimates the $P\%$ point, for reasons that go beyond the fact that these methods cannot estimate δ and that will become apparent in the following discussion of an alternative approach.

Empirical evidence indicates that the decision space is actually partitioned as in the illustration of Appendix Fig. 9c, in which the interval $[\delta_1, \delta_2]$ is often not centered on 0 and where values of D lead to undecided judgments. Adaptive methods that provide estimates of some percentage point on the psychometric function also cannot identify this scenario. Yet, fitting the indecision model allows detaching the stimulus level at which $\text{Prob}(S_t > S_s) = P/100$ from percentage points on a psychometric function whose form is affected by decisional and response processes that taint the isolated portrait of sensory processes only observable if $\delta_2 = \delta_1 = 0$. Specifically, $\text{Prob}(S_t > S_s) = P/100$ implies $\text{Prob}\left(Z > -\frac{\mu(x_t) - \mu(x_s)}{\sqrt{2}\sigma}\right) = P/100$ so that $\frac{\mu(x_t) - \mu(x_s)}{\sqrt{2}\sigma} = \Phi^{-1}(P/100)$ and, then, $x_t = \mu^{-1}(\mu(x_s) + \sqrt{2}\sigma\Phi^{-1}(P/100))$. Thus, irrespective of the values of δ_1 and δ_2 (which are nevertheless

estimated along the way) and also irrespective of differences in psychometric functions across presentation orders, the test level x_T at which $\text{Prob}(S_t > S_s) = P/100$ (i.e., the threshold defined relative to $d' = \sqrt{2}\Phi^{-1}(P/100)$) is obtained from the parameters of the psychophysical function μ estimated by fitting the model. In the expression that estimates x_T from the psychophysical function μ , σ acts as an inconsequential scale factor that only needs to be set to the value used upon fitting the model, and $\sigma = 1$ is the most convenient choice for this purpose.

References

- Alcalá-Quintana, R., & García-Pérez, M. A. (2007). A comparison of fixed-step-size and Bayesian staircases for sensory threshold estimation. *Spatial Vision*, 20, 197–218. <https://doi.org/10.1163/156856807780421174>
- 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. <https://doi.org/10.1080/17470218.2010.540018>
- Bausenhardt, K. M., Dyjas, O., & Ulrich, R. (2015). Effects of stimulus order on discrimination sensitivity for short and long durations. *Attention, Perception, & Psychophysics*, 77, 1033–1043. <https://doi.org/10.3758/s13414-015-0875-8>
- Brownlee, K. A. (1957). The principles of experimental design. *Industrial Quality Control*, 13, 12–20.
- Cai, M. B., & Eagleman, D. M. (2015). Duration estimates within a modality are integrated sub-optimally. *Frontiers in Psychology*, 6, 1041. <https://doi.org/10.3389/fpsyg.2015.01041>
- Dyjas, O., Bausenhardt, K. M., & Ulrich, R. (2012). Trial-by-trial updating of an internal reference in discrimination tasks: Evidence from effects of stimulus order and trial sequence. *Attention, Perception, & Psychophysics*, 74, 1819–1841. <https://doi.org/10.3758/s13414-012-0362-4>
- Dyjas, O., Bausenhardt, K. M., & Ulrich, R. (2014). Effects of stimulus order on duration discrimination sensitivity are under attentional control. *Journal of Experimental Psychology: Human Perception and Performance*, 40, 292–307. <https://doi.org/10.1037/a0033611>
- Dyjas, O., & Ulrich, R. (2014). Effects of stimulus order on discrimination processes in comparative and equality judgements: Data and models. *Quarterly Journal of Experimental Psychology*, 67, 1121–1150. <https://doi.org/10.1080/17470218.2013.847968>

- Ellenbein, J. L., Small, A. M., & Davis, J. M. (1993). Developmental patterns of duration discrimination. *Journal of Speech, Language, and Hearing Research*, 36, 842–849. <https://doi.org/10.1044/jshr.3604.842>
- Ellinghaus, R., Gick, M., Ulrich, R., & Bausenhardt, K. M. (2019). Decay of internal reference information in duration discrimination: Intertrial interval modulates the Type B effect. *Quarterly Journal of Experimental Psychology*, 72, 1578–1586. <https://doi.org/10.1177/1747021818808187>
- Ellinghaus, R., Ulrich, R., & Bausenhardt, K. M. (2018). Effects of stimulus order on comparative judgments across stimulus attributes and sensory modalities. *Journal of Experimental Psychology: Human Perception and Performance*, 44, 7–12. <https://doi.org/10.1037/xhp0000495>
- Fechner, G. T. (1966). *Elements of Psychophysics, Vol. I*. New York, NY: Holt. (Original work published 1860).
- Fernberger, S. W. (1913). On the relation of the methods of just perceptible differences and constant stimuli. *Psychological Monographs*, 14, 1–81. <https://doi.org/10.1037/h0093068>
- Fernberger, S. W. (1914a). A simplification of the practice of the method of constant stimuli. *American Journal of Psychology*, 25, 121–130. <https://doi.org/10.2307/1413025>
- Fernberger, S. W. (1914b). On the elimination of the two extreme intensities of the comparison stimuli in the method of constant stimuli. *Psychological Review*, 21, 335–355. <https://doi.org/10.1037/h0072494>
- Fernberger, S. W. (1916). The influence of mental and physical work on the formation of judgments in lifted weight experiments. *Journal of Experimental Psychology*, 1, 508–532. <https://doi.org/10.1037/h0075134>
- Fernberger, S. W. (1920). Interdependence of judgments within the series for the method of constant stimuli. *Journal of Experimental Psychology*, 3, 126–150. <https://doi.org/10.1037/h0065212>
- Fernberger, S. W. (1921). An experimental study of the “stimulus error”. *Journal of Experimental Psychology*, 4, 63–76. <https://doi.org/10.1037/h0071167>
- García-Pérez, M. A. (1998). Forced-choice staircases with fixed step sizes: Asymptotic and small-sample properties. *Vision Research*, 38, 1861–1881. [https://doi.org/10.1016/S0042-6989\(97\)00340-4](https://doi.org/10.1016/S0042-6989(97)00340-4)
- García-Pérez, M. A. (2000). Optimal setups for forced-choice staircases with fixed step sizes. *Spatial Vision*, 13, 431–448. <https://doi.org/10.1163/156856800741306>
- García-Pérez, M. A. (2001). Yes–no staircases with fixed step sizes: Psychometric properties and optimal setup. *Optometry and Vision Science*, 78, 56–64. <https://doi.org/10.1097/00006324-200101010-00015>
- García-Pérez, M. A. (2002). Properties of some variants of adaptive staircases with fixed step sizes. *Spatial Vision*, 15, 303–321. <https://doi.org/10.1163/15685680260174056>
- García-Pérez, M. A. (2011). A cautionary note on the use of the adaptive up–down method. *Journal of the Acoustical Society of America*, 130, 2098–2107. <https://doi.org/10.1121/1.3628334>
- García-Pérez, M. A. (2014a). Adaptive psychophysical methods for nonmonotonic psychometric functions. *Attention, Perception, & Psychophysics*, 76, 621–641. <https://doi.org/10.3758/s13414-013-0574-2>
- García-Pérez, M. A. (2014b). Does time ever fly or slow down? The difficult interpretation of psychophysical data on time perception. *Frontiers in Human Neuroscience*, 8, 415. <https://doi.org/10.3389/fnhum.2014.00415>
- García-Pérez, M. A., & Alcalá-Quintana, R. (2005). Sampling plans for fitting the psychometric function. *Spanish Journal of Psychology*, 8, 256–289. <https://doi.org/10.1017/S113874160000514X>
- García-Pérez, M. A., & Alcalá-Quintana, R. (2007). Bayesian adaptive estimation of arbitrary points on a psychometric function. *British Journal of Mathematical and Statistical Psychology*, 60, 147–174. <https://doi.org/10.1348/000711006X104596>
- 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. <https://doi.org/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. <https://doi.org/10.3758/APP.72.4.1155>
- García-Pérez, M. A., & Alcalá-Quintana, R. (2011a). Interval bias in 2AFC detection tasks: Sorting out the artifacts. *Attention, Perception, & Psychophysics*, 73, 2332–2352. <https://doi.org/10.3758/s13414-011-0167-x>
- García-Pérez, M. A., & Alcalá-Quintana, R. (2011b). Improving the estimation of psychometric functions in 2AFC discrimination tasks. *Frontiers in Psychology*, 2, 96. <https://doi.org/10.3389/fpsyg.2011.00096>
- García-Pérez, M. A., & Alcalá-Quintana, R. (2012a). On the discrepant results in synchrony judgment and temporal-order judgment tasks: A quantitative model. *Psychonomic Bulletin & Review*, 19, 820–846. <https://doi.org/10.3758/s13423-012-0278-y>
- García-Pérez, M. A., & Alcalá-Quintana, R. (2012b). Response errors explain the failure of independent-channels models of perception of temporal order. *Frontiers in Psychology*, 3, 94. <https://doi.org/10.3389/fpsyg.2012.00094>
- García-Pérez, M. A., & Alcalá-Quintana, R. (2013). Shifts of the psychometric function: Distinguishing bias from perceptual effects. *Quarterly Journal of Experimental Psychology*, 66, 319–337. <https://doi.org/10.1080/17470218.2012.708761>
- García-Pérez, M. A., & Alcalá-Quintana, R. (2017). The indecision model of psychophysical performance in dual-presentation tasks: Parameter estimation and comparative analysis of response formats. *Frontiers in Psychology*, 8, 1142. <https://doi.org/10.3389/fpsyg.2017.01142>
- García-Pérez, M. A., & Alcalá-Quintana, R. (2018). Perceived temporal order and simultaneity: Beyond psychometric functions. In A. Vatakis, F. Balci, M. Di Luca, & Á. Correa (Eds.), *Timing and Time Perception: Procedures, Measures, and Applications* (pp. 263–294). Leiden, Netherlands: Brill. https://doi.org/10.1163/9789004280205_013
- García-Pérez, M. A., & Alcalá-Quintana, R. (2019). The do’s and don’ts of psychophysical methods for interpretability of psychometric functions and their descriptors. *Spanish Journal of Psychology*, 22, e56. <https://doi.org/10.1017/sjp.2019.49>
- García-Pérez, M. A., Giorgi, R. G., Woods, R. L., & Peli, E. (2005). Thresholds vary between spatial and temporal forced-choice paradigms: The case of lateral interactions in peripheral vision. *Spatial Vision*, 18, 99–127. <https://doi.org/10.1163/1568568052801591>
- García-Pérez, M. A., & Núñez-Antón, V. (2018). Nonparametric tests for equality of psychometric functions. *Behavior Research Methods*, 50, 2226–2255. <https://doi.org/10.3758/s13428-017-0989-0>
- García-Pérez, M. A., Núñez-Antón, V., & Alcalá-Quintana, R. (2015). Analysis of residuals in contingency tables: Another nail in the coffin of conditional approaches to significance testing. *Behavior Research Methods*, 47, 147–161. <https://doi.org/10.3758/s13428-014-0472-0>
- García-Pérez, M. A., & Peli, E. (2014). The bisection point across variants of the task. *Attention, Perception, & Psychophysics*, 76, 1671–1697. <https://doi.org/10.3758/s13414-014-0672-9>
- García-Pérez, M. A., & Peli, E. (2015). Aniseikonia tests: The role of viewing mode, response bias, and size–color illusions. *Translational Vision Science & Technology*, 4(3)9. <https://doi.org/10.1167/tvst.4.3.9>

- García-Pérez, M. A., & Peli, E. (2019). Psychophysical tests do not identify ocular dominance consistently. *i-Perception*, 10(2). <https://doi.org/10.1177/2041669519841397>
- Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. New York, NY: Wiley.
- Greenberg, M. G. (1965). A modification of Thurstone's law of comparative judgment to accommodate a judgment category of "equal" or "no difference". *Psychological Bulletin*, 64, 108–112. <https://doi.org/10.1037/h0022283>
- Hegelmaier, F. (1852). Ueber das Gedächtniss für Linear-anschauungen. *Archiv für physiologische Heilkunde*, 11, 844–853. <https://hdl.handle.net/2027/mdp.39015013754380>
- Hellström, Å. (1985). The time-order error and its relatives: Mirrors of cognitive processes in comparing. *Psychological Bulletin*, 97, 35–61. <https://doi.org/10.1037/0033-2909.97.1.35>
- Johnson, D. M., Watson, C. S., & Kelly, W. J. (1984). Performance differences among the intervals in forced-choice tasks. *Perception & Psychophysics*, 35, 553–557. <https://doi.org/10.3758/BF03205952>
- Kaernbach, C. (2001). Adaptive threshold estimation with unforced-choice tasks. *Perception & Psychophysics*, 63, 1377–1388. <https://doi.org/10.3758/BF03194549>
- Kim, M.-A., Lee, Y.-M., & Lee, H.-S. (2010). Comparison of d' estimates produced by three versions of a duo-trio test for discriminating tomato juices with varying salt concentrations: The effects of the number and position of the reference stimulus. *Food Quality and Preference*, 21, 504–511. <https://doi.org/10.1016/j.foodqual.2010.01.005>
- Klein, S. A. (2001). Measuring, estimating, and understanding the psychometric function: A commentary. *Perception & Psychophysics*, 63, 1421–1455. <https://doi.org/10.3758/BF03194552>
- Laming, D., & Laming, J. (1992). F. Hegelmaier: On memory for the length of a line. *Psychological Research*, 54, 233–239. <https://doi.org/10.1007/BF01358261>
- Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Perception & Psychophysics*, 63, 1279–1292. <https://doi.org/10.3758/BF03194543>
- Lin, L., Hedayat, A. S., Sinha, B., & Yang, M. (2002). Statistical methods for assessing agreement: Models, issues, and tools. *Journal of the American Statistical Association*, 97, 257–270. <https://doi.org/10.1198/016214502753479392>
- Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45, 255–268. <https://doi.org/10.2307/2532051>
- Mayer, K. M., Di Luca, M., & Ernst, M. O. (2014). Duration perception in crossmodally-defined intervals. *Acta Psychologica*, 147, 2–9. <https://doi.org/10.1016/j.actpsy.2013.07.009>
- Nachmias, J. (2006). The role of virtual standards in visual discrimination. *Vision Research*, 46, 2456–2464. <https://doi.org/10.1016/j.visres.2006.01.029>
- Nesti, A., Barnett-Cowan, M., MacNeilage, P. R., & Bühlhoff, H. H. (2014). Human sensitivity to vertical self-motion. *Experimental Brain Research*, 232, 303–314. <https://doi.org/10.1007/s00221-013-3741-8>
- Olson, C. L., & Ogilvie, J. C. (1972). The method of constant stimuli with two or more categories of response. *Journal of Mathematical Psychology*, 9, 320–338. [https://doi.org/10.1016/0022-2496\(72\)90022-3](https://doi.org/10.1016/0022-2496(72)90022-3)
- Saito, T. (1994). Psychological scaling of the asymmetry observed in comparative judgement. *British Journal of Mathematical and Statistical Psychology*, 47, 41–62. <https://doi.org/10.1111/j.2044-8317.1994.tb01024.x>
- Self, M. W., Mookhoek, A., Tjalmma, N., & Roelfsema, P. R. (2015). Contextual effects on perceived contrast: Figure-ground assignment and orientation contrast. *Journal of Vision*, 15(2). <https://doi.org/10.1167/15.2.2>
- Stewart, R. M. (1914). The effect of intensity and order on the apparent pitch of tones in the middle range. *Psychological Monographs*, 16, 157–160. <https://doi.org/10.1037/h0093076>
- Stott, L. H. (1935). Time-order errors in the discrimination of short tonal durations. *Journal of Experimental Psychology*, 18, 741–766. <https://doi.org/10.1037/h0057714>
- Treisman, M., & Watts, T. R. (1966). Relation between signal detectability theory and the traditional procedures for measuring sensory thresholds: Estimating d' from results given by the method of constant stimuli. *Psychological Bulletin*, 66, 438–454. <https://doi.org/10.1037/h0020413>
- Treutwein, B. (1995). Adaptive psychophysical procedures. *Vision Research*, 35, 2503–2522. [https://doi.org/10.1016/0042-6989\(95\)00016-X](https://doi.org/10.1016/0042-6989(95)00016-X)
- Ulrich, R., & Vorberg, D. (2009). Estimating the difference limen in 2AFC tasks: Pitfalls and improved estimators. *Attention, Perception, & Psychophysics*, 71, 1219–1227. <https://doi.org/10.3758/APP.71.6.1219>
- Urban, F. M. (1908). *The Application of Statistical Methods to the Problems of Psychophysics*. Philadelphia, PA: The Psychological Clinic Press.
- Urban, F. M. (1910). The method of constant stimuli and its generalizations. *Psychological Review*, 17, 229–259. <https://doi.org/10.1037/h0074515>
- Versace, J. (1967). Experimental validation of dynamic models. *SAE Technical Paper 670896*, 41–44. <https://doi.org/10.4271/670896>
- Vickers, D. (1975). Where Angell feared to tread: Response time and frequency in three-category discrimination. In P. M. A. Rabbitt, & S. Dornic (Eds.), *Attention and Performance V* (pp. 455–469). New York, NY: Academic Press.
- von Castell, C., Hecht, H., & Oberfeld, D. (2017). Measuring perceived ceiling height in a visual comparison task. *Quarterly Journal of Experimental Psychology*, 70, 516–532. <https://doi.org/10.1080/17470218.2015.1136658>
- 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. <https://doi.org/10.1037/h0034736>
- Woodrow, H. (1935). The effect of practice upon time-order errors in the comparison of temporal intervals. *Psychological Review*, 42, 127–152. <https://doi.org/10.1037/h0063696>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.