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)

MIGUEL A. GARCÍA-PÉREZ AND ROCÍO ALCALÁ-QUINTANA
Universidad Complutense, Madrid, Spain

Lapid, Ulrich, and Rammsayer (2008) reported that estimates of the difference limen (DL) from a two-alternative forced choice (2AFC) task are higher than those obtained from a reminder task. This article reanalyzes their data in order to correct an error in their estimates of the DL from 2AFC data. We also extend the psychometric functions fitted to data from both tasks to incorporate an extra parameter that has been shown to allow obtaining accurate estimates of the DL that are unaffected by lapses. Contrary to Lapid et al.'s conclusion, our reanalysis shows that DLs estimated with the 2AFC task are only minimally (and not always significantly) larger than those estimated with the reminder task. We also show that their data are contaminated by response bias, and that the small remaining difference between DLs estimated with 2AFC and reminder tasks can be reasonably attributed to the differential effects that response bias has in either task as they were defined in Lapid et al.'s experiments. Finally, we discuss a novel approach presented by Ulrich and Vorberg (2009) for fitting psychometric functions to 2AFC discrimination data.

In the typical discrimination experiment for estimation of the difference limen (DL) or the point of subjective equality (PSE), observers are confronted with a series of trials, each of which consists of two presentations. One of the presentations displays a standard stimulus whose magnitude s is fixed throughout the experiment, whereas the other displays a comparison stimulus whose magnitude x varies across trials and that, on any given trial, can be anywhere from well below to well above the magnitude of the standard. On each trial, the observer is asked to report the presentation in which the stimulus had a higher magnitude. The series of trials can be designed and arranged according to diverse criteria, yielding psychophysical methods that are given different names.

Lapid, Ulrich, and Rammsayer (2008) have addressed experimentally the question of whether two of these methods—namely, the reminder task and the two-alternative forced choice (2AFC) task (thoroughly described in the next section)—yield comparable DL estimates. Conventional wisdom has it that both tasks should yield the same outcomes within sampling error, but in a series of experiments, Lapid et al. gathered evidence indicating that the 2AFC task produces consistently larger DLs than the reminder task. In other words, the discrimination performance of observers seemed to be significantly worse in the 2AFC than in the reminder task. Lapid et al. also reported

evidence to the effect that discrimination performance in the reminder task is significantly better when the standard stimulus is presented in the first interval than when it is presented in the second, and proposed a moving-average model to explain some of their results. In particular, the model explains the time-order effect (i.e., differences in performance associated with presentation order) through a sensation-weighting rule analogous to that proposed by Hellström (1979, 1985, 2003) but in which the observer is further assumed to use an implicit standard that is continuously updated through sensory experience gathered along a subset of the immediately preceding trials.

This article reanalyzes the data from Lapid et al.'s (2008) experiments, with two goals. Our first goal is to correct an error in their estimates of the DL from the 2AFC task. This error arises from their inappropriate choice of the psychometric function to fit to 2AFC data, with the consequence that DL estimates from the 2AFC task were spuriously higher than they should be. Our reanalysis shows that DLs estimated with the 2AFC task are only minimally (and not always significantly) larger than those estimated with the reminder task. Our second goal is to show that response bias (defined as an observer's preference toward one of the two intervals in a trial) contaminates DL estimates differently in the reminder and in the 2AFC task, with the consequence that DL estimates from the 2AFC task again

M. A. García-Pérez, miguel@psi.ucm.es

become higher than they should be. We discuss a model that explains interval bias (defined as differences in performance when the comparison stimulus is presented in the first or second intervals) by incorporating response bias when observers guess at random on trials in which standard and comparison stimuli appear identical to the observer (Alcalá-Quintana & García-Pérez, 2009; García-Pérez & Alcalá-Quintana, 2009a). This model is shown to account for the pattern of results across Lapid et al.'s experiments, and it also prescribes how to conduct discrimination experiments so as to eliminate the contaminating effects of response bias. We finally discuss the utility of a novel approach described by Ulrich and Vorberg (2009) for fitting psychometric functions to 2AFC discrimination data that are potentially affected by response bias.

THE TWO TASKS

The reminder task used by Lapid et al. (2008) consisted of two intervals designed in such a way that the standard stimulus was always presented in the first interval, with the comparison stimulus always presented in the second. Nonetheless, their Experiments 5 and 6 used what they referred to as an "irregular condition," in which the standard was always presented in the second interval. The observer had to report whether the stimulus in the second interval had a higher (or a lower) magnitude than that in the first.¹ The resultant data, in the form of proportion of times that the comparison stimulus was reported to have higher magnitude as a function of its actual magnitude, were used to fit a psychometric function that ranges from 0 to 1. Finally, half the difference between the 75% and the 25% points (i.e., the magnitude of the comparison stimulus at which the probability is, respectively, .75 and .25 on the fitted psychometric function) was taken to be the estimate of the DL from the reminder task.

The 2AFC task used by Lapid et al. (2008) differed from the reminder task in two respects. One was that the order of presentation of standard and comparison stimuli was randomized, so that on any given trial either stimulus could be presented in the first interval with the same probability. The second difference was that the comparison stimulus never had magnitudes on both sides of the standard; in most of their experiments, the comparison always had higher magnitude than the standard, and only in their Experiment 3 did it always have lower magnitude.² Observers had to report which of the two intervals presented a stimulus with a larger magnitude. The data were also expressed in the form of proportion correct, which is equivalent to the proportion of times that the comparison stimulus was reported to be larger, because it was actually larger (except in their Experiment 3, where the definition of a correct response was modified accordingly). These proportions correct, as a function of the difference between the magnitude of the comparison and that of the standard, were used to fit a psychometric function that ranged from .5 to 1, and which differed from that fitted to data from the reminder task by mere vertical compression and displacement so as to accommodate this restricted range. Finally, the DL was estimated as the 75% point on the fitted function. As

explained in the next section, this definition of the DL and the scaled form of the psychometric function that was fitted to 2AFC data are both inadequate.

FITTING A PSYCHOMETRIC FUNCTION TO 2AFC DISCRIMINATION DATA

The inappropriateness of the psychometric function fitted by Lapid et al. (2008) to 2AFC data is best described via comparison with the function that they fitted to data from the reminder task. This latter psychometric function was

$$\Psi(x; a, b) = \frac{1}{1 + \exp[-(x - a) / b]}, \quad (1)$$

which has a lower asymptote at 0 and an upper asymptote at 1. This function gives the probability that an observer would report the comparison stimulus to have a larger magnitude than the standard (whose magnitude is s) as a function of the magnitude x of the comparison. Thus, when $x \ll s$, the probability that the observer responds "larger" is virtually zero, and when $x \gg s$, the probability that the observer responds "larger" is virtually unity. Also, the case $x = s$ defines the point of objective equality (POE) because then comparison and standard stimuli have the same physical magnitude; if they are also perceived to be equal by the observer, Equation 1 has its 50% point precisely at $x = a = s$ —that is, at the level of the standard stimulus. In any case, parameter a defines the PSE, whether or not its value matches the magnitude s of the standard stimulus and, then, the POE.

In a 2AFC task the situation is actually identical, despite the fact that in Lapid et al.'s (2008) experiments all the magnitudes of the comparison stimulus satisfied either $x > s$ or $x < s$. We will refer to 2AFC data gathered in these conditions as "one-sided data," for obvious reasons. But as far as the sensory system is concerned, the reminder task and the 2AFC task are identical, and only differ as to the particular set of comparison stimuli presented to the observer and also as to the phrasing of the question that they have to respond to on every trial.³ It should also be realized that a psychometric function ranging from 0 to 1 such as that in Equation 1 is a description of the observable outcomes of the sensory process and, thus, it should identically apply in the reminder and the 2AFC tasks. These functions may for various reasons have different parameters in reminder and 2AFC tasks, but in both tasks they must range from 0 to 1. Indeed, the 2AFC discrimination task has a peculiarity relevant only to the experimenter, who must be aware that "probability correct" as a label for the ordinate of the psychometric function of Equation 1 indicates "probability that the comparison is reported to have a larger magnitude" and not "probability that the observer reported the stimulus whose magnitude was actually larger." The reason is that a psychometric function that abides by the latter definition is not monotonic, as are those derived by Ulrich and Vorberg (2009). On this caveat, all theoretical analyses (e.g., Foley & Schwarz, 1998; García-Pérez & Alcalá-Quintana, 2007; Werkhoven & Snippe, 1996) concur in showing that the psychometric function that holds in 2AFC discrimination tasks with

above-threshold standards has the full range of Equation 1 when lapses (whether the result of actual lapses of attention or mere finger errors) are not considered. That is, also in the 2AFC task, when $x \ll s$ the probability that the observer reports the comparison stimulus to have a larger magnitude than the standard is virtually zero and, conversely, when $x \gg s$ the probability that the observer reports the comparison stimulus to have a larger magnitude than the standard is virtually unity. And these characteristics should not vary when the experimenter chooses to present comparison stimuli satisfying $x > s$ only, or $x < s$ only. The POE and the PSE are also defined here as described in the preceding paragraph for the reminder task.

Yet, Lapid et al. (2008) chose to fit their one-sided 2AFC data to the scaled function

$$\Psi(x; a, b) = 0.5 + \frac{0.5}{1 + \exp[-(x - a) / b]}, \quad (2)$$

whose inadequacy should be clear now but is further discussed next.⁴ First, Equation 2 has a lower asymptote at .5 and will never reach the low performance levels that left-side data would have shown, had they been collected. Indeed, this lower asymptote at .5 implies that chances are 50% that the observer reports the comparison to have a larger magnitude than the standard, even when $x \ll s$. Second, the fact that the lower asymptote of Equation 2 is at .5 further implies that the theoretical PSE lies at a comparison magnitude of $-\infty$ (in an obvious abuse of mathematical notation on our part) and that the theoretical performance level at the POE (when $x = s$) must be above 50%.

The reasons that may have led Lapid et al. (2008) to fit 2AFC discrimination data to Equation 2 are unclear. In an earlier article, Miller and Ulrich (2001) correctly stated that the psychometric function in a discrimination task “relates the probability of a certain response, r , to the value of a stimulus, s , along a certain physical dimen-

sion” (p. 1399). They noted that the response r is typically “comparison longer,” and they also correctly acknowledged that a psychometric function “must satisfy three conditions: (1) $F(s) \rightarrow 0$ as $s \rightarrow -\infty$, (2) $F(s) \rightarrow 1$ as $s \rightarrow \infty$, and (3) $F(s)$ is monotonically increasing with s .” However, a subsequent article by Ulrich and Miller (2004) was entirely based on the mistake of considering that the psychometric function for discrimination must have a lower asymptote at .5, a mistake perhaps arising from the true fact that the chance level in a 2AFC task is .5. In reality, the latter fact implies only that the psychometric function for discrimination is nonmonotonic and has a discontinuity at the PSE when regarded as expressing the probability of a correct response in the strict physical sense (i.e., the probability of reporting the stimulus whose physical magnitude is actually larger; see Ulrich & Vorberg, 2009). The mathematical inconvenience that nonmonotonicity brings around is generally avoided by considering the psychometric function to express instead the probability that the comparison stimulus is reported to have a larger magnitude, thus satisfying the three assumptions listed by Miller and Ulrich.

To illustrate the differences between Equations 1 and 2, consider the data in Figure 5 of Lapid et al. (2008), which we present here in Figure 1A. Our maximum-likelihood estimation procedure (see the Appendix) yields $\hat{a} = 534.48$ and $\hat{b} = 18.99$ in the fit of Equation 2 to these data. Fitting instead the proper psychometric function in Equation 1 yields $\hat{a} = 502.28$ and $\hat{b} = 26.73$. The consequence of these differences is that the DL estimated from the wrong psychometric function of Equation 2 is larger than that estimated from the proper psychometric function of Equation 1. Specifically, and for the data in Figure 1A, Equation 2 yields $DL = 534.48 - 500 = 34.48$ msec, whereas Equation 1 yields $DL = 26.73 \log(.75/.25) = 29.37$ msec. An error of 5.11 msec for this particular observer may seem negligible, considering that Lapid et al. reported an average difference

Figure 1. Data from the 2AFC task (symbols) and fitted psychometric functions (curves) for 2 of the participants in Experiment 2 of Lapid, Ulrich, and Rammsayer (2008). The dashed curve is the fit of Equation 2; the solid curve is the fit of Equation 1. Parameter estimates and derived difference limen (DL) are given next to each function. (A) Data for the participant whose results were plotted in Lapid et al.’s Figure 5. (B) Data for a different participant for whom DLs estimated from Equations 1 and 2 differ substantially.

of 27.1 msec between DLs estimated with the 2AFC task through Equation 2 (which averaged 59.3 msec in Lapid et al.'s [2008] Experiment 2) and DLs estimated with the reminder task through Equation 1 (which averaged 32.2 msec in their Experiment 1). In other words, a difference of only 5 msec between correctly and incorrectly estimated DLs from 2AFC data will not make up for the substantially lower DL estimates obtained with the reminder task. The error may nevertheless be larger for other observers, as illustrated in Figure 1B, where DLs estimated by fitting Equations 1 and 2 are, respectively, 17.62 msec and 47.43 msec, resulting in a sizeable difference of 29.81 msec.

Figure 1 helps clarify the impact that the inappropriate choice of psychometric function had on Lapid et al.'s (2008) estimates of the DL from 2AFC data through Equation 2. With Equation 1, the DL is defined as the distance between the 75% point and the PSE (i.e., the 50% point on the fitted psychometric function),⁵ and these two points move rigidly as the fitted psychometric function shifts laterally when the PSE does not match the POE (as in Figure 1B). In other words, from Equation 1, the DL is actually a pure measure of the width of the fitted psychometric function. With Equation 2, in contrast, the DL is arbitrarily defined as the distance between the 75% point and the POE (i.e., the actual magnitude of the standard stimulus), where the latter is a fixed reference point independent of lateral displacements of the fitted psychometric function. In other words, with Equation 2, the DL is determined only by the location of the fitted psychometric function, not by its width. This peculiar definition of the DL is needed, because Equation 2 actually sets the PSE at $-\infty$, as discussed above; then, only when the PSE under Equation 1 roughly equals the POE (as in Figure 1A, thus equating also the arbitrary reference point for Equation 2)

can the two functions provide similar estimates of the DL. It should finally be stressed that psychometric functions such as the one in Equation 1 are routinely fitted to one-sided 2AFC discrimination data (see, e.g., Figure 3 in Kontsevich, Chen, & Tyler, 2002; see also Figure 1 in Georgeson & Meese, 2006, and Figure 1 in Klein, 2006),⁶ although the function is often amended to incorporate an additional lapsing-rate parameter λ , as described in the Appendix.

Given that the DL from the 2AFC task may have been overestimated for each observer by Lapid et al. (2008), the question is whether the average DL, when properly estimated from the 2AFC task, will still be significantly larger than that estimated from the reminder task. Einat Lapid kindly provided us with raw data from their study, and we have carried out the analysis that answers this question, with the results reported next. We took the entire 2AFC data set (four separate blocks of 100 trials for each of 34 observers) from their Experiment 2 and fitted the psychometric functions in Equations 1 and 2 in each of the 136 cases.⁷ The DL arising from each function was then computed as described for the 2 sample cases shown in Figure 1. For each of the 136 cases, Figure 2A shows DLs estimated by fitting Equation 2 against DLs estimated from the same data by fitting Equation 1 instead. Data points lie around the diagonal identity line, although Equation 1 seems to yield DL estimates well above 100 msec in cases in which Equation 2 generally yields DL estimates below 100 msec (open symbols in Figure 2A), which is against all expectations created by the illustration in Figure 1. The reason for this characteristic becomes evident in Figure 3, which shows data and fitted psychometric functions for 8 of the cases just referred to. Clearly, the data are too noisy to indicate any psychomet-

Figure 2. Relation between 2AFC estimates of the difference limen (DL) obtained by fitting Equations 1 and 2 (A) or Equations A1 and 2 (B) to the same data. Each symbol comes from one of the 136 data sets (four conditions for each of 34 observers) in Experiment 2 of Lapid, Ulrich, and Rammsayer (2008). Only symbols within the inset square (representing estimated DLs valued at or below 100 msec with both equations) were considered valid for statistical analyses. Open symbols denote cases for which raw data and fitted psychometric functions are plotted in Figure 3.

Figure 3. Data from the 2AFC task (symbols) and fitted psychometric functions (curves) for 8 of the participants in Experiment 2 of Lapid, Ulrich, and Rammsayer (2008). The dashed curve is the fit of Equation 2; the solid curve is the fit of Equation 1. Parameter estimates and derived difference limens (DLs) from each equation are given in each panel. The lack of a clean increasing pattern in the data yields meaningless parameter estimates. Estimated DL from Equation 1 is above 140 msec in all cases, and much lower DLs are obtained in some cases by fitting Equation 2 instead, regardless of the shallow fitted psychometric functions and the lack of a clean increasing pattern in the data.

ric function at all. The lack of a clean increasing pattern in the data makes the fitted Equations 1 and 2 immensely shallow, a problem that affects estimates of the DL from Equation 1 (which are actual measures of the width of the function) much more than it affects estimates of the DL from Equation 2 (whose values are actually unaffected

by the width of the function). This is the reason that DL estimates from Equation 1 are comparatively larger than those from Equation 2, although both of them are equally meaningless because they come from extremely noisy data. There is actually no point in fitting a psychometric function to such unpatterned data. Inspection of data

and fitted psychometric functions for the entire set of 136 cases involved in Lapid et al.'s Experiment 2 revealed that unpatterned trends similar to those in the panels of Figure 3 render shallow psychometric functions for which the estimated b in Equation 1 always lies at or almost at the upper bound of the parameter space ($\hat{b} \approx 150$; see the Appendix), rendering spurious DLs larger than 100 msec. It should be noted that the DL for stimulus duration has been documented to be in the 5%–10% range for standard durations in the 150–1,500 msec range, whether in the visual, auditory, or tactile sensory modalities, and whether observers are trained or untrained (Westheimer, 1999; see also Buonomano, Bramen, & Khodadadifar, 2009). Thus, for the 500-msec standard duration in the experiments of Lapid et al., DLs would be expected to be in the 25- to 50-msec range. Therefore, because DLs of 100 msec or higher appear unrealistic in the light of previous research, we discarded all cases yielding this outcome in the fit of either equation, which left the 110 cases within the small inset square next to the origin in Figure 2A. It should be noted incidentally that Lapid et al. also discarded data from some of their experiments (specifically, 3 observers from their Experiment 1, 4 from their Experiment 2, and 1 from their Experiment 5; yet, Dr. Lapid could not tell us which particular observers those were so that we could also remove them for our analyses), although they only indicated that this was done because these observers' estimated DLs "were 2 SD s above the mean of the entire sample" (Lapid et al., 2008, p. 296).

With these data, we can come back to our initial question of whether DL estimates obtained by fitting the incorrect Equation 2 to 2AFC data are or are not spuriously inflated in comparison with those estimated by fitting Equation 1. For the 110 cases within the inset square of Figure 2A, the average DL ($\pm SE$) from Equation 2 is 52.69 ± 2.01 msec, whereas the average DL from Equation 1 is 49.37 ± 1.99 msec, for an average difference of only $52.69 - 49.37 = 3.32$ msec between fitting the incorrect Equation 2 and fitting the proper Equation 1. Thus, using Equation 1 to estimate the DL from 2AFC data does not seem to change the result much, but so far we have considered only the first part of our argument.

As discussed earlier, a DL estimate from Equation 1 is actually a measure of the psychometric function's width, which must then be estimated as precisely as possible. It has been shown that lapses spuriously produce broad psychometric functions when their mathematical form does not incorporate a free lapsing-rate parameter (Wichmann & Hill, 2001), as is the case with Equation 1. To circumvent this problem, we fitted instead Equation A1 in the Appendix, and Figure 2B plots DL estimates from the previous fit of Equation 2 against DL estimates obtained from Equation A1, which is an extension of Equation 1 that merely incorporates a free lapsing-rate parameter λ . Removal of unpatterned data sets (with the same criterion described above) rendered the 115 data points in the inset square, for which the average DL from Equation 2 was 53.65 ± 2.01 msec, whereas the average DL from Equation A1 was 39.55 ± 2.05 msec. This renders a final average difference of 14.10 msec between fitting the incorrect

Equation 2 and fitting the proper and complete Equation A1 to 2AFC data.

In sum, the wrong choice of a psychometric function to fit to 2AFC data, as well as the lack of a free lapsing-rate parameter, spuriously inflated DLs estimated by Lapid et al. (2008). The preceding reanalysis has shown that proper DL estimates from the 2AFC task (averaging 39.55 msec when noisy data were discarded) are meaningfully lower than those reported by Lapid et al. (which averaged 59.3 msec), and they are also lower than those obtained from our own fit of Equation 2 (which averaged 53.65 msec when unpatterned data were discarded also). How does this figure finally compare with DLs from the reminder task? Lapid et al. reported an average DL of 32.2 msec obtained by fitting Equation 1 to data from the reminder task in their Experiment 1, but the width of the fitted psychometric functions (and, thus, the estimated DL) was also potentially broadened by the lack of a lapsing-rate parameter. For this reason, we have carried out our own reanalysis (removing also unpatterned data and fitting Equation A1 instead), which rendered an average DL of 30.99 ± 1.17 msec. Thus, a difference of 8.56 ± 2.37 msec still remains between average DLs from 2AFC and reminder tasks when DLs are in both cases properly estimated through Equation A1. This difference is significant by an unconditional separate-variance two-sample t test [Hayes & Cai, 2007; Zimmerman, 2004; $t(182.81) = 3.61, p < .0005$, two-tailed]. In any case, DLs from the 2AFC task appear to be about 28% larger than those obtained with the reminder task, when Lapid et al. reported them to be about 84% larger. The next section will look into the reason for this residual difference, but we will close this section with a further discussion of the validity of DL estimates obtained by Lapid et al.

Lapid et al. (2008) reported in their Appendix A the results of simulations whose goal was "to rule out the possibility that the discrepancy of the DL estimates between the 2AFC and the reminder tasks is due to the estimation procedure applied" (p. 304). That possibility seemed to have actually been ruled out because in those simulations the average estimated DLs ended up being virtually identical in both tasks. The question then arises as to where these simulations stand with respect to our main topic in this section. What Lapid et al. did in their simulations was to generate artificial data across experiments that were identical to their empirical experiments, and then to subject the resultant data to the same analyses actually performed on their empirical data. In these simulations, then, data for the reminder task were generated using the psychometric function in Equation 1, whereas data for the 2AFC task were generated using the empirically inadequate psychometric function in Equation 2. The true DL was identical in both cases, and, also in both cases, it was recovered in an unbiased fashion by their estimation procedure. This outcome is a natural consequence of the well-known fact that, when data are not unreasonably scarce, the true parameters can always be accurately recovered when the generating and fitted models match—that is, when the target function whose parameters are estimated coincides with the function that actually generated the data. The reason that

Figure 4. Data from the reminder task (symbols) and fitted psychometric functions (curves) for 2 of the participants in Experiment 1 of Lapid, Ulrich, and Rammsayer (2008). The fitted curves are given by Equation 1. Parameter estimates and derived difference limens (DLs) are given in each panel. A dashed vertical line in each panel indicates the point of objective equality (POE) (at 500 msec). (A) Data for the participant whose results were plotted in Lapid et al.'s Figure 2. (Copyright 2008 by the Psychonomic Society, Inc.) (B) Data for a different participant for whom the point of subjective equality (PSE) is also substantially away from the POE but in the opposite direction.

Lapid et al.'s simulations are not adequate to address their declared purposes is that in their experiments with human observers 2AFC data were produced instead by a psychometric function that (like that in Equation A1) ranges from near 0 to near 1, whereas parameters were estimated for the psychometric function in Equation 2, which ranges from .5 to 1. Simulations carried out in these alternative conditions would have yielded results mimicking the empirical finding that DLs estimated from the 2AFC task are larger than those estimated from the reminder task. However, as was evident from our preceding analysis in this section, this sole factor cannot account for the full difference between DL estimates from reminder and 2AFC tasks that was reported by Lapid et al.; and the reason lies in an additional factor discussed in the next section.

EFFECTS OF RESPONSE BIAS IN REMINDER AND 2AFC TASKS

Experiment 1 of Lapid et al. (2008) revealed a characteristic that they seemed to bypass. Their Figure 2 (which we reproduce here as Figure 4A) shows the psychometric function of Equation 1 fitted to one of their observers' data. The standard stimulus had a duration of 500 msec and, strikingly, the PSE for this observer turned out to be 536 msec, whereas the DL was 19 msec.⁸ It seems remarkable that the PSE is above the POE by an amount almost twice the DL. The consequence, which is clearly apparent in Figure 4A, is that the psychometric function is substantially displaced to the right of the POE instead of being centered on it, as one might perhaps expect. Also remarkable is the fact that the average PSE reported by Lapid et al. for the entire group of observers was 502.2 msec, virtually right on the POE. Given that the observer whose data are reported in Figure 4A had the PSE substantially above the POE, the only way in which the average PSE can be virtually at the POE is if the psychometric function of

other observers is displaced also substantially but in the opposite direction (see Figure 4B for an actual case). Also, the standard error of the mean PSE was 1.7 msec, which is almost 50% larger than the standard error of the mean DL (1.2 msec). In other words, variations in PSE across observers overwhelm variations in DL; this is striking, because one might expect the PSE to lie always very close to the POE (and not to vary wildly above or below it across observers), whereas the DL might reasonably vary much more across observers. In the remainder of this section, we argue that the displaced psychometric functions and the large variability of PSE estimates (compared with that of DL estimates) in the reminder task reflect an identifiable form of response bias that can also be shown to differentially affect reminder and 2AFC tasks. To facilitate this demonstration, consider first some sample data from a different study in which the role of response bias in discrimination tasks was explicitly explored.

Figure 5 shows psychometric functions for 2 representative observers (rows) in each of three conditions (columns) in a study of visual contrast discrimination (Alcalá-Quintana & García-Pérez, 2009). Data were gathered with an adaptive method of constant stimuli very similar to that used by Lapid et al. (2008). Responses were collected for an overall number of 480 temporal 2AFC trials, in which the contrast of the comparison stimulus could be lower, higher, or equal to that of the standard stimulus whose log contrast was -1 . The order of presentation of standard and comparison stimuli across trials was randomized with the constraint that exactly half of the trials present the standard first and the other half present it second. Psychometric functions were then independently fitted to data from the subset of 240 trials in which the standard was presented first (with the results shown in the left column of Figure 5), to the subset of 240 trials in which the standard was presented second (center column of Figure 5), and to the entire set

Figure 5. Data (symbols) and fitted psychometric functions (curves) for 2 observers in Experiment 1 of Alcalá-Quintana and García-Pérez (2009). The left column includes the 240 trials in which the standard stimulus was presented in the first interval, the center column includes the 240 trials in which the standard stimulus was presented in the second interval, and the right column includes all 480 trials. A dashed vertical line in each panel indicates the point of objective equality (POE); a solid vertical line indicates the point of subjective equality (PSE). Estimated values for the difference limen (DL) and the PSE are given in each panel.

of 480 trials regardless of order of presentation (right column of Figure 5).

Two characteristics of these psychometric functions are relevant to the present discussion. First, when the standard was presented first (left column of Figure 5), as it was under the reminder task in Experiment 1 in Lapid et al. (2008), the psychometric function turned up systematically displaced, so that the PSE (vertical solid line in each panel) differed from the POE (vertical dashed line); in contrast, when the standard was presented second (center column of Figure 5), the psychometric function was equally displaced but in the opposite direction. Second, and as a result of the above, when trials of either type were aggregated (right column of Figure 5) the resultant psychometric function had the PSE at the precise location of the POE and, at the same time, it was broader than those obtained when presentation order was fixed for all trials. It should be stressed that all the data in Figure 5 were gathered in a single experimental session in which observers were not aware of which interval in each trial presented the standard and which one presented the comparison, so that there must have been a single psychometric function (which carries a unique PSE and a unique DL) governing the performance of each observer.

The implications of these results for the study of Lapid et al. (2008) are straightforward. The reminder task in their Experiment 1 presented the standard stimulus always in the first interval (thus, similar to the condition represented in the left column of Figure 5). In contrast,

the 2AFC task in their Experiment 2 presented instead the standard stimulus in a random interval across trials (thus similar to the condition represented in the right column of Figure 5 but involving only one-sided data), which yields relatively broader psychometric functions. And broader psychometric functions necessarily carry larger DLs. In consequence, DLs estimated from sessions in which the order of presentation of standard and comparison stimuli is randomized across trials (the 2AFC task of Lapid et al., 2008) must be larger than those estimated from sessions in which the order of presentation is fixed across trials (the reminder task of Lapid et al., 2008). Alcalá-Quintana and García-Pérez (2009) showed that this pattern of psychometric functions reveals contamination by response bias and that elimination of response bias renders psychometric functions that are identical (within sampling error), whether the standard stimulus is always presented first or second, or in an interval newly decided at random with equiprobability on each trial.

Whether this response bias has also contaminated the experiments of Lapid et al. (2008) is easy to check out, although not all of their experiments lend themselves to the necessary analysis. Splitting 2AFC data into two subsets, as in the example of Figure 5, is only feasible when a large number of trials has been administered. In the experiments of Lapid et al., the overall number of 2AFC trials was only 100 (too few),⁹ so that splitting them into two subsets would render unusable data. A sensible alternative to check for response bias would be to combine trials from a condition in-

volving the reminder task with the standard presented first with trials from a condition involving the same task but with the standard presented second. This strategy would further allow for a side-by-side comparison of the three resultant psychometric functions, as illustrated in Figure 5. Lapid et al.'s Experiments 5 and 6 included these two conditions for the reminder task, and they also included a separate 2AFC task that allows extending the side-by-side picture of performance with an independent psychometric function fitted to the additional (one-sided) 2AFC data.

We thus took the data from Experiments 5 and 6 of Lapid et al. (2008) and fitted Equation A1 in the four conditions just described. We fitted Equation A1 directly because we do not aim to reproduce the results of Lapid et al., but rather to obtain adequate and comparable PSE and DL estimates across conditions. Sample results of this fit are shown in Figure 6 for 4 of the participants (rows) in Experiment 5. Observers whose data are reported in the two top rows show evidence of response bias in opposite directions, such that in the standard-first condition (left-

Figure 6. Data (symbols) and fitted psychometric functions in Equation A1 (curves) for 4 observers in Experiment 5 of Lapid, Ulrich, and Rammsayer (2008). Left to right, the first, second, and fourth columns represent the three separate conditions for which data were independently gathered in this experiment; the third column merely includes aggregated data from the conditions in the first and second columns. Crossing dashed lines indicate the point of objective equality (POE) in each panel. Estimated values for the difference limen (DL) and the point of subjective equality (PSE) are also given in each panel. The two top rows depict data that are relatively clean and support the estimates; the two bottom rows depict data that are much noisier and that cast doubt on the validity of parameter estimates.

Figure 7. Relationship between estimates of the point of subjective equality (PSE) (A) and the difference limen (DL) (B) from the reminder task when the standard is presented first (abscissa) or second (ordinate). Data come from Lapid, Ulrich, and Rammsayer's (2008) Experiments 5 (solid symbols) and 6 (open symbols). The diagonal line in each panel indicates the expected location of the data in the presence of response bias.

most column) their PSEs are displaced in opposite directions from the POE, a displacement that reverses for each observer in the standard-second condition (second column from the left) and disappears for aggregated data (third column from the left). Observers whose data are reported in the two bottom rows illustrate that the data were not always as clean as for the observers in the two top rows, and that parameter estimates (and, thus, PSE and DL estimates) are suspect when noisy data indicate that the observer's performance is inconsistent. The rightmost column in Figure 6 shows results for the one-sided 2AFC task also performed by each of these observers, which usually appear noisier.

Figure 7 shows the complete picture of response bias, as indicated by the relation between PSEs and DLs estimated from the two versions of the reminder task. It is useful to recall that response bias manifests in PSE estimates on opposite sides of the POE when the standard is presented first or second, and in similar estimates of the DL in both cases. The solid lines in Figures 7A and 7B show these expectations, and PSE and DL estimates from Equation A1 for Experiments 5 (solid symbols) and 6 (open symbols) seem to support them. The product-moment correlation between PSE estimates in Figure 7A was $-.491$ for the 22 data points from Experiment 5 and $-.687$ for the 17 data points from Experiment 6,¹⁰ both of which are significantly lower than zero ($p < .05$ in both cases; left-tailed test). On the other hand, the correlation between DL estimates in Figure 7B was $.517$ for the data from Experiment 5 and $.387$ for the data from Experiment 6, where only the former is significantly greater than zero ($p < .01$; right-tailed test).

To complete this picture, Figure 8 shows the relation of DL estimated from the 2AFC task with DL estimated from the "regular" reminder task (Figure 8A), from the "irregular" reminder task (Figure 8B), and from aggregated data from both versions of the reminder task (Fig-

ure 8C). For Experiments 5 and 6, product-moment correlations were $.431$ and $.556$ in Figure 8A (both of which are significantly greater than zero, $p < .05$), $-.074$ and $.027$ in Figure 8B (nonsignificant), and $.289$ and $.294$ in Figure 8C (nonsignificant).

The results of the preceding analyses support the notion that the data of Lapid et al. (2008) are contaminated by response bias but, before this issue is further considered in the next section, it is useful to summarize performance differences across reminder and 2AFC tasks as they emerge from our reanalysis of the data from Experiments 5 and 6. Figure 9 shows a summary of estimated PSEs (Figure 9A) and DLs (Figure 9B) in each of the four conditions within each experiment. Recall that these summaries reflect estimates obtained by fitting Equation A1 in all cases, and that unpatterned data sets were excluded from all analyses.

Consider first the case of PSEs in Figure 9A. Whether in Experiment 5 or 6, the average PSE in the R condition (for aggregated data across the two versions of the reminder task) lies in between those for the R1 condition (the "regular" reminder task, in which the standard is presented first) and the R2 condition (the "irregular" reminder task, in which the standard is presented second). Also, the size of the error bars indicates that the variability of PSE estimates from the R condition is smaller than that of PSE estimates from either the R1 or the R2 conditions.¹¹ This pattern is indicative of response bias, as illustrated in Figure 5. It is also remarkable that the average PSE is in all conditions above the POE of 500 msec. The difference is statistically significant in the R and FC (for 2AFC) conditions of Experiments 5 and 6. Whether statistically significant or not, this bias in the estimation of the PSE is expected when psychometric functions are fitted to scarce data sets such as those of Lapid et al. (2008; see García-Pérez & Alcalá-Quintana, 2005, Figure 14).

Figure 8. Relationship between difference limens (DLs) estimated with the 2AFC task (ordinate) and DLs estimated with the reminder task when the standard is presented first (A), when the standard is presented second (B), and for aggregated data across presentation orders (C). Data come from Lapid, Ulrich, and Rammsayer's (2008) Experiments 5 (solid symbols) and 6 (open symbols).

As for DLs displayed in Figure 9B, the large differences across experiments are inconsequential because they are presumably related to the auditory (Experiment 5) versus visual (Experiment 6) modalities involved. Other than that, recall from the preceding section that the average DL with the reminder task in Experiment 1 (in which the standard was presented first) was 30.99 ± 1.17 msec, whereas the average DL with the 2AFC task in Experiment 2 was 39.55 ± 2.05 msec, yielding the significant difference reported earlier and implying DLs that are about 28% larger in the 2AFC task. It should be kept in mind that those figures came from an across-subjects study, in which each task involved a different group of observers who, in turn, may have different response biases that contaminate the comparison. In contrast, a single group of observers performed both tasks in Experiment 5, which yielded an average DL of 40.90 ± 3.83 msec in the R1 condition (the

"regular" reminder task, as in Experiment 1) and an average DL of 45.80 ± 4.59 msec in the FC condition (a 2AFC task like that in Experiment 2). These averages do not differ significantly by a paired-samples *t* test, and the average DL in the 2AFC task is only about 12% larger than in the reminder task. Yet another different group of observers performed both tasks in Experiment 6, which yielded an average DL of 60.80 ± 6.17 msec in the R1 condition and an average DL of 83.91 ± 9.81 msec in the FC condition. These averages are significantly different by a paired-samples *t* test [$t(14) = -2.70, p < .05$, two-tailed], and the average DL in the 2AFC task is about 38% larger than in the reminder task. Interestingly, then, the size of the difference between DLs estimated from the reminder and the 2AFC tasks seems to vary greatly across groups of participants, a characteristic in itself understandable, given the small sample sizes and the large standard er-

Figure 9. Mean point of subjective equality (PSE) (A) and difference limen (DL) (B) across conditions in Lapid, Ulrich, and Rammsayer's (2008) Experiments 5 and 6. Mean values are given by the height of each bar, and error bars represent standard errors of the means. R1, reminder task with the standard presented first; R2, reminder task with the standard presented second; R, aggregated data for both versions of the reminder task; FC, 2AFC task.

rors of individual DL estimates arising from psychometric functions fitted to scarce data (see García-Pérez & Alcalá-Quintana, 2005).¹² Differences in the magnitude and sign of response bias within and across groups can alter the pattern of differences between DLs estimated with the reminder or the 2AFC task. Then, whether DLs estimated with the reminder and the 2AFC tasks differ is heavily dependent on variations in response bias within and across the groups of observers performing each task. This is best appreciated by noting that DL estimates in the reminder task of Experiment 1 averaged 30.99 msec under our analysis, whereas DL estimates in the identical reminder task of Experiment 5 (which differed only in that it was performed by a different group of observers) averaged 40.90 msec, also under our analysis. Thus, a comparison of the outcomes of the reminder task in different groups of observers yields a difference of 9.90 msec that is statistically significant [$t(27.28) = 2.42, p < .05$, two-tailed]. Recall also that the average DL from the 2AFC task in Experiment 2 was 39.55 msec. Had this figure been compared with the results of the reminder task in Experiment 5, the conclusion would have been that reminder and 2AFC tasks yield DL estimates that are not significantly different. Thus, great caution is required in interpreting data from different groups of observers, particularly when the results are based on very few observations per data point and when the contaminating effects of extraneous factors (such as differences in response bias across groups) have not been eliminated.

On another front, the smallest average DL across the conditions in Experiments 5 and 6 occurs for R1, whereas the three remaining conditions yield analogous DLs (see Figure 9B). These results seem to support Lapid et al.'s (2008) conclusion that discrimination performance improves under the reminder task when the standard stimulus is presented in the first interval, although this superiority only reaches statistical significance in Experiment 5 [$t(21) = -2.28, p < .05$, two-tailed]. Finally, our illustration of the effects of response bias in visual contrast discrimination with a sufficiently large number of trials (Figure 5) suggests that the estimated DL in the R condition should be larger than those in the R1 and R2 conditions, but Figure 9B does not confirm this expectation. The reason for this lies in a subtlety that should not be overlooked. As the two bottom rows in Figure 6 reveal, data from the experiments of Lapid et al. were often too noisy as a result of the small number of trials that were administered; hence, the fitted psychometric functions did not display the clean pattern that was reported in Figure 5 for data involving large numbers of trials. In these conditions, large sampling errors may make the fitted psychometric functions in the R1 or R2 conditions so broad that their aggregation to define the R condition cannot be broader (this is actually the case for the observers represented in the two bottom rows of Figure 6, a characteristic shared by the data of many other observers). Hence, the expectation that DLs in the R condition should be larger than those in the R1 or R2 conditions does not necessarily hold when parameter estimates are affected by large sampling errors.

Given that the trace of response bias has been found in the data of Lapid et al. (2008), an interesting question is whether this bias can be eliminated to obtain uncontaminated PSE and DL estimates from reminder and 2AFC tasks. Unfortunately, this cannot be done in post hoc analyses unless preparatory measures for this purpose were taken during data collection. What these measures should be is described in the next section, where we present a model of response bias that indeed suggests how it can be eliminated.

A MODEL OF RESPONSE BIAS, AND HOW TO ELIMINATE IT

Conventional signal detection theory analyses assume that the sensory effect of a stimulus is a random variable whose mean $\mu(x)$ depends on the magnitude x of the stimulus, and whose variance σ^2 is independent of this magnitude.¹³ On each trial of the 2AFC or reminder tasks, the observer compares the sensory effects drawn in each interval and reports the interval in which it was larger. Assuming a normal distribution for the sensory effects S_1 and S_2 drawn during the first and second intervals by stimuli with magnitudes x_1 and x_2 , respectively, this response rule subsumes a normally distributed decision variable $D = S_2 - S_1$ with mean $\mu(x_2) - \mu(x_1)$ and variance $2\sigma^2$, such that the observer chooses the first interval when $D < 0$ and chooses the second interval when $D > 0$, yielding the so-called *standard difference model*. But this model cannot account for response bias, and the alternative that is usually considered to this effect is the so-called *difference model with bias*, which relies on the same distributional assumptions as the standard difference model does, but replaces the decision rule so that the observer chooses the first interval when $D < c$ and chooses the second interval when $D > c$, with some fixed criterion $c \in \mathbb{R}$.

The standard difference model and the difference model with bias both assume that (1) observers can on each trial collect conclusive sensory information on which to base their responses and (2) that they are never undecided. In contrast, participants in psychophysical experiments spontaneously report that they have to guess at random on some trials because both intervals appeared to display stimuli with the same magnitudes, which reveals that sometimes sensory evidence is inconclusive. Fechner (1860/1966) himself discussed this issue at some length, and substantial empirical evidence of indecision was collected by a number of authors who also devised methods to deal with it at the beginning of the 20th century (for an excellent review of this research, see Part I of Link, 1992). The questions arise, then, as to how observers behave when guessing at random because of indecision and how a signal detection model can incorporate this characteristic.

Alcalá-Quintana and García-Pérez (2009; see also García-Pérez & Alcalá-Quintana, 2009a) presented and tested experimentally a variant of the standard difference model that incorporates a guessing rule. This *difference model with guessing* assumes an indifference zone, such that the observer chooses the first interval when $D < -\delta$,

chooses the second interval when $D > \delta$, and guesses at random when $-\delta < D < \delta$. In the latter case, the reasons that observers choose the first or second interval are unrelated to sensory processes, but spurious factors such as handedness or the design of the interface used to collect responses generally make them show a preference toward one of the intervals when guessing at random (García-Pérez & Alcalá-Quintana, 2009a; Jäkel & Wichmann, 2006). For our purposes in this section, the difference model with guessing and the difference model with bias are both equally effective at accounting for response bias, but we will adopt the difference model with guessing for our illustration, because Alcalá-Quintana and García-Pérez (2009) and García-Pérez and Alcalá-Quintana (2009a) have shown that some predictions of the difference model with bias are not borne by data which, at the same time, support the predictions of the difference model with guessing.

Figure 10 illustrates the difference model with guessing and shows the psychometric functions predicted for 2AFC discrimination tasks in various conditions. Without loss of generality, all illustrations assume $\delta = 0.75$, $\sigma^2 = 1$, a standard stimulus whose magnitude is $s = 500$ msec, and sensory effects whose mean depends on stimulus magnitude as $\mu(x) = x/40$.¹⁴ The left panel in Figure 10A shows the distribution of D in a trial in which the standard is presented in the first interval and the magnitude of the comparison is $x = 460$ msec, so that the mean of D is $460/40 - 500/40 = -1$; the right panel shows an analogous trial in which the standard is instead presented in the second interval so that the mean of D is now $500/40 - 460/40 = 1$. Values of D within the regions separated by the vertical lines in each panel make the observer give the response indicated at the top, and the probability of that response is given by the numeral at the bottom. Following forced choice instructions, observers must give a response even when they are undecided (which occurs with probability .322 in the illustrations of Figure 10A) and the response that they give in these cases is completely unrelated to sensory processes, although their preference toward hitting one or the other designated response keys has dramatic consequences. Ideally, observers would distribute their guesses evenly and hit the designated keys for the first and second intervals with the same probability, but this even distribution is unlikely to occur in actual practice. To illustrate the effects of a failure of this ideal behavior, suppose that the observer is strongly biased and always hits the designated key for the first interval when guessing; then, the probability of reporting the comparison to have a larger magnitude when the standard is presented first (left panel in Figure 10A) will be .108, because bias toward the first interval turns all guesses into the observers' selection of the standard as the stimulus that has a larger magnitude. In contrast, when the standard is presented second, this probability will be $.108 + .322 = .430$, because all guesses now turn into selections of the comparison stimulus. Each of the two probabilities just mentioned is indicated by a solid circle at the appropriate abscissa in the applicable panel of Figure 10B. As the magnitude of the comparison stimulus varies to cover the range from well below the magnitude s of the standard

to well above it, the distribution of D in each panel of Figure 10A moves accordingly over the fixed vertical lines that determine the responses. Given the above assumption as to how the mean of D varies with the magnitude of the comparison stimulus, the model yields the psychometric functions shown in Figure 10B, each of which has the PSE away from the POE to the same extent (30 msec) but in a different direction, and they imply the same DL of 38 msec. It can easily be noted that the two curves would be swapped across panels if the observer's response bias had the form of a complete preference toward hitting the designated key for the second interval on guessing.

When the order of presentation of standard and comparison stimuli is randomized across trials, the observer's bias will have each of these opposite effects in the subset of applicable trials (i.e., one when the standard is presented first and the other when the standard is presented second). As a result, and if any given comparison stimulus is presented the same number of times in each interval, the psychometric function for randomized presentations is the average of the two functions separately displayed in Figure 10B,¹⁵ yielding the result in Figure 10C. This function now has the PSE right on the POE, but it is broader than either of the functions shown in Figure 10B, and carries a DL of 43.67 msec, which is about 15% higher than that carried by the functions in Figure 10B. The difference model with guessing can thus reproduce the overall results reported by Lapid et al. (2008) and also the details that have been uncovered by our analysis in the preceding sections: The reminder task shows evidence of response bias in the form of narrow psychometric functions whose PSE is far from the POE (conditions similar to those in Figure 10B) and thus renders smaller estimates of the DL than those obtained from the broader psychometric functions that arise when presentation order is randomized (as in Lapid et al.'s 2AFC task; conditions similar to those in Figure 10C).

One way around this response bias would be to have observers distribute their guesses evenly across "interval 1" and "interval 2" responses, but this is unlikely to succeed, if only because this even distribution would have to occur separately at each magnitude of the comparison stimulus. Next, we describe an alternative strategy successfully used for this purpose on numerous occasions (e.g., Alcalá-Quintana & García-Pérez, 2007, 2009; García-Pérez & Alcalá-Quintana, 2007, 2009a; García-Pérez, Alcalá-Quintana, Woods, & Peli, 2010) and that takes advantage of a response format that has also been used for other purposes (García-Pérez, 2010; Hellström, 2003; Kaernbach, 2001; Poggel, Strasburger, & MacKeben, 2007; van Vleet & Robertson, 2006; von Waldkirch, Lukowicz, & Tröster, 2004, 2005). This strategy consists of three steps. The first step is to enable a third response key for observers to use when both stimuli in a trial seem to be identical, and to record these "guess-key" hits differently in the data file.¹⁶ The second step (in post hoc analyses) is to transform the recorded guesses at each magnitude of the comparison stimulus into equal numbers of correct and incorrect responses, just as the unbiased observer would ideally have done. The final step is to fit the psychometric function in Equation A1 to the data. Formally, this is analogous to dis-

Figure 10. Illustration of the difference model with guessing and its predictions. (A) Distribution of the decision variable D when the standard is presented in the first interval (left panel) or in the second interval (right panel). Two vertical lines partition the horizontal axis into three regions, each of which yields the response indicated at the top, and whose probabilities are given by the numerals near the bottom. (B) Psychometric function arising in the conditions indicated at the top when the observer's response bias has the form of a preference toward hitting the designated key for the first interval upon guessing. (C) Psychometric function arising when the order in which standard and comparison stimuli are presented is randomized across trials. (D) Psychometric functions arising when response bias is eliminated (solid curves) in conditions that are otherwise identical to those in panels B. The dashed curves are predictions from the standard difference model in which there is no interleaved indifference zone—the central region in panels A has degenerated to null width—and, hence, observers do not guess.

tributing the probability within the “guess” region in the panels of Figure 10A evenly across “interval 1” and “interval 2” responses, and yields the psychometric functions shown as solid curves in the panels of Figure 10D. Whether the standard is presented first or second (left and right panels of Figure 10D, respectively), the psychometric function has exactly the same shape; hence, the psychometric function when order of presentation is randomized will have the same shape, by virtue of being the average of two identical functions. In all cases, the PSE lies on the POE, and the DL is the same as it is in the functions of Figure 10B.

Although it is obvious that the above strategy eliminates response bias, there is still the question of whether the resultant functions in Figure 10D actually carry the “true” PSE and DL. The answer depends only on what is considered to be the true PSE and DL. The standard difference model arising from signal detection theory analyses is conventionally used to model the psychometric function, which then results from assumptions of lack of bias and no guessing and represents the performance of an ideal observer who always collects and uses reliable sensory information. Compared with the difference model with guessing, illustrated in Figure 10A, the standard difference model implies $\delta = 0$; thus, no guessing. Then, if only for reference to such an ideal observer, the true PSE and DL can be considered to be those embedded in the psychometric function arising from the standard difference model. This function is shown as a dashed curve in the panels of Figure 10D, and was derived with the same approach as the remaining psychometric functions in Figure 10; but for the case that $\delta = 0$, so that $D < 0$ yields an “interval 1” response, $D > 0$ yields an “interval 2” response, and guessing never takes place. It can easily be seen that the psychometric functions arising from the difference model with guessing are virtually identical to this “true” one when response bias has been eliminated by separating out all guesses and distributing them evenly into “interval 1” and “interval 2” responses at each applicable magnitude of the comparison stimulus. The minor discrepancy between the solid and dashed curves in Figure 10D affects only the DL, because the solid psychometric functions (from the difference model with guessing) are only slightly broader than the dashed functions (from the standard difference model), whereas solid and dashed functions have the PSE at the same location. The magnitude of the discrepancy varies only slightly with the width 2δ of the indifference zone, and can actually be eliminated by estimating parameter δ (García-Pérez & Alcalá-Quintana, 2009a).

THE STRATEGY OF ULRICH AND VORBERG (2009)

An article by Ulrich and Vorberg (2009) discussing what form the psychometric function fitted to 2AFC discrimination data should have was published after the first draft of this article had passed the first review round. The material presented by Ulrich and Vorberg has obvious bearings on the issues addressed in this article, and it is worth commenting on them here.

The Appropriate Psychometric Function for 2AFC Data

Ulrich and Vorberg (2009) started their theoretical analysis with the acknowledgment that there may be constant errors (in the form of shifts in the location of the PSE, changes in the slope of the psychometric function, or both) according to whether the standard or the comparison stimulus is presented first, although they did not propose a model to illustrate why these constant errors occur. Then, they assumed that the psychometric function Ψ_1 , when the standard is presented first, has the form of Equation 1 with parameters a_1 and b_1 , whereas the psychometric function Ψ_2 , when the standard is presented second, also has the form of Equation 1, but now with parameters a_2 and b_2 . Next, they claimed that, despite these constant errors of unknown origin and contingent on presentation order, the PSE must necessarily lie at the POE when presentation order is randomized (i.e., in what they called a 2AFC task) and when the data represent responses to equal numbers of trials with the standard presented first or second at each comparison level. Thus, the psychometric function Ψ_{2AFC} for 2AFC data must be the average of Ψ_1 and Ψ_2 —that is,

$$\Psi_{2AFC}(x) = \frac{\Psi_1(x; a_1, b_1) + \Psi_2(x; a_2, b_2)}{2}. \quad (3)$$

The functions Ψ_1 and Ψ_2 in Equation 3 should then be jointly estimated under the additional constraint that Ψ_{2AFC} has its PSE at the POE, which is easily seen to imply

$$a_2 = s + \frac{b_2}{b_1}(s - a_1), \quad (4)$$

where s is the magnitude of the standard stimulus.¹⁷ In other words, a_2 is not a free parameter in the fit. Note also that the psychometric function for 2AFC data is not really obtained by fitting some mathematical function to 2AFC data, but is rather predicted to have the form given by Equation 3 when Ψ_1 and Ψ_2 are jointly fitted under the constraint of Equation 4 to separate subsets of 2AFC data: Ψ_1 to data from trials in which the standard was presented first, and Ψ_2 to data from trials in which the standard was presented second. Ulrich and Vorberg also reported data from a single observer (one of the authors) illustrating their approach, and showing that 2AFC data are well described by Equation 3.

Naturally, Equation 3 cannot be expressed back in the form of Equation 1, which raises the issue as to whether the purportedly corrected estimates of the DL in 2AFC tasks that we reported earlier are in error because they come from Equation 1 (actually, from Equation A1, which extends Equation 1 with a lapsing-rate parameter) and not from the proper Equation 3 (or its extension to include a lapsing-rate parameter). It is useful to keep in mind that the relevant topic in the present discussion is whether Equation A1 yields a DL meaningfully different from that arising from Equation 3, and not whether the parameters of Equation A1 are interpretable in some other respect. The answer to this question is given in Figure 11, which shows that Equation 3 with the parameters estimated by Ulrich and Vorberg

Figure 11. Psychometric functions of different functional form fitted to 2AFC data. The dashed curves are the components of Equation 3 fitted by Ulrich and Vorberg (2009; their Figure 4) to their data, and the curve created by the string of open circles is their average as given by Equation 3. The solid curve is the best-fitting psychometric function with the mathematical form of Equation A1, which gives an excellent approximation and yields a point of subjective equality (PSE) and a difference limen (DL) that are almost identical to those provided by Equation 3.

(2009; see their Figure 4) from their data can be extremely accurately approximated by Equation A1 with appropriate parameters. In other words, if the data conform to Equation 3, they will also conform to Equation A1. As a result, in the illustration of Figure 11, the DL from Equation A1 is 64.44 msec, whereas the DL from Equation 3 is 60.51 msec. Thus, for this observer, fitting Equation A1 to 2AFC data does not yield a DL estimate meaningfully different from that obtained by fitting Equation 3. There is also the question of whether the mathematical form of the components of Equation 3 is actually logistic: Cumulative Gaussians or Weibull functions are also possible candidates used in many contexts, and there is no theoretical or empirical argument justifying the use of a specific mathematical form.¹⁸ Then, there is no strong reason to expect that the psychometric function for 2AFC data should have the mathematical form of the average of two logistic functions. What matters then is not whether the function fitted to 2AFC data has a precise mathematical form but rather, as illustrated in Figure 11, that the fitted function ranges from 0 to 1 (leaving lapses aside) and describes the data adequately.

In any case, for the single observer in Figure 11, Equation 3 actually yields a smaller DL estimate than Equation A1, which suggests that the slight difference between reminder and 2AFC estimates of the DL in Lapid et al.'s (2008) experiments could be even smaller than we have reported above, if Equation 3 were fitted to their data. Then, the question is whether our earlier demonstration that Lapid et al.'s data do not support their conclusion that reminder and 2AFC tasks yield different DL estimates (which we arrived at by fitting Equation A1 to their 2AFC data) will still hold when the proper Equation 3 is fitted instead to 2AFC data. This seems likely, in view of Fig-

ure 11, but we nevertheless addressed this issue, again using data from Lapid et al.'s Experiments 5 and 6.

Because lapses should be taken into account, we did not fit Equation 3 but rather an extension including a lapsing-rate parameter—namely,

$$\Psi_{2AFC}(x) = \frac{\Psi_1(x; a_1, b_1, \lambda_1) + \Psi_2(x; a_2, b_2, \lambda_2)}{2}, \quad (5)$$

where Ψ_1 and Ψ_2 have the mathematical form of Equation A1. With these extra parameters, the constraint that Ψ_{2AFC} has its PSE at the POE is now easily seen to imply

$$a_2 = s + b_2 \ln \left[\frac{(\lambda_1 - \lambda_2) \exp(a_1/b_1) + (1 - \lambda_1 - \lambda_2) \exp(s/b_1)}{(\lambda_1 - \lambda_2) \exp(s/b_1) + (1 - \lambda_1 - \lambda_2) \exp(a_1/b_1)} \right]. \quad (6)$$

We then fitted jointly and under the constraint in Equation 6 the components Ψ_1 and Ψ_2 of Equation 5 respectively to data from the regular and irregular reminder tasks in Experiments 5 and 6 of Lapid et al. (2008). Representative results are plotted in Figure 12 for the same observers whose data and separately fitted psychometric functions with the form of Equation A1 were shown in Figure 6. It is important to stress that the psychometric functions plotted in the first and second columns of Figure 12 actually represent the functions Ψ_1 and Ψ_2 that were fitted under the constraint of Equation 6, and that the psychometric functions plotted in the two other columns are identical and represent the average given by Equation 5, which has not been directly fitted to the aggregate data displayed in the third column or to the independent, one-sided 2AFC data displayed in the right column. Note also that the PSEs reported in the panels of

Figure 12. Data (symbols) and jointly fitted psychometric functions (curves) with the mathematical form of Equation 5 for the same observers whose data (and separately fitted psychometric functions with the form of Equation A1) were plotted in Figure 6. The curves in the first and second columns represent the components Ψ_1 and Ψ_2 of Equation 5, fitted jointly to the applicable data sets under the constraint of Equation 6. The curves in the third and fourth columns (which are identical) are simply the average of Ψ_1 and Ψ_2 as given by Equation 5, displayed along with combined reminder data (third column) or with independent 2AFC data (fourth column) to which the curve has not been directly fitted. These curves, then, represent the psychometric function that should describe 2AFC data according to the theoretical analysis of Ulrich and Vorberg (2009).

the two columns on the right (which match the POE) merely reflect a characteristic forced into the fitted function by the constraint in Equation 6. Also, a comparison of the DLs reported across the panels of Figure 12 with those reported for the same observers and data in Figure 6 does not reveal major or consistent discrepancies.

The overall picture is given in Figure 13A, which shows DL estimates from Equation 5 (such as those reported in the panels on the right column of Figure 12) against DL

estimates from Equation A1 fitted to 2AFC data (such as those reported in the panels on the right column of Figure 6) for each of the participants in Lapid et al.'s (2008) Experiments 5 (solid symbols) and 6 (open symbols). Excluding again DL estimates that are above 100 msec, DLs estimated from Equations 5 and A1 respectively average 49.00 ± 4.03 and 46.69 ± 4.92 in Experiment 5 and 60.10 ± 5.19 and 59.28 ± 6.56 in Experiment 6. The differences are not significant in either of the two experiments. Then, our DL

Figure 13. (A) Relationship between difference limens (DLs) estimated for 2AFC data from Equation 5 (ordinate) and those estimated from Equation A1 (abscissa). Data come from Lapid, Ulrich, and Rammsayer's (2008) Experiments 5 (solid symbols) and 6 (open symbols). (B) Scatterplot of values for the Akaike information criterion (AIC) in the fit of Equation 5 (ordinate) against values for the AIC in the fit of Equation A1 (abscissa) producing the DL estimates reported in panel A.

estimates for 2AFC data based on Equation A1 are not meaningfully different from those that arise from Equation 5. Figure 13B further shows that the fit of Equation A1 to 2AFC data (as illustrated in the right column of Figure 6) is better than the fit of Equation 5 (which represents a prediction) to the same 2AFC data (as illustrated in the right column of Figure 12), as measured through the Akaike information criterion (AIC), which provides directly comparable measures of fit (see Liu & Smith, 2009).

The final issue that we will consider in our evaluation of Lapid et al.'s (2008) conclusion under the approach of Ulrich and Vorberg (2009) is how DL estimates for 2AFC tasks from Equation 5 compare with DL estimates for reminder tasks as given by the two components of the numerator of Equation 5. Figures 14A and 14B, respec-

tively, show the PSE and DL estimates from Equation 5 in each of the conditions of Lapid et al.'s Experiments 5 and 6. Figure 14A merely illustrates the consequences of the constraint in Equation 6, namely, the PSE for 2AFC data (labeled FC in the abscissa) is invariably constrained to be at the standard duration of 500 msec, which forces the estimated PSE in conditions R1 and R2 (the reminder tasks with the standard presented first or second, respectively) away from the POE in opposite directions. Nevertheless, average PSEs in the R1 and R2 conditions are not significantly different from the standard level of 500 msec in any of the two experiments, and average PSEs in the R1 and R2 conditions are not significantly different from one another in any of the two experiments either. Figure 14B shows that DLs for 2AFC tasks arising from Equation 5 generally lie in

Figure 14. Mean point of subjective equality (PSE) (A) and difference limen (DL) (B) across conditions in Lapid, Ulrich, and Rammsayer's (2008) Experiments 5 and 6 when psychometric functions are fitted with the approach of Ulrich and Vorberg (2009). Graphical conventions are as in Figure 9. The apparent lack of error bars for PSEs in the FC conditions of Experiments 5 and 6 only reflects that all individual estimates of the PSE are constrained by Ulrich and Vorberg's approach to lie at the point of objective equality (POE) of 500 msec, and, thus, there is no variability in the estimates.

between those in the R1 and R2 tasks in both experiments, but the differences are only statistically significant between the R1 task and the two other tasks in Experiment 5 [for the difference between R1 and R2, $t(19) = -2.34$, $p < .05$, two-tailed; for the difference between R1 and FC, $t(21) = -4.35$, $p < .0005$, two-tailed] and between the R1 and FC tasks in Experiment 6 [$t(12) = -3.27$, $p < .001$, two-tailed]. Significant as they are, these differences are indeed very small (e.g., in Experiment 5, the average DL is 43.89 ± 4.25 msec in the R1 task and 49.00 ± 4.03 msec in the FC task), yielding DLs that are less than 12% larger in the 2AFC than in the reminder task.

A comparison of Figure 14B (means obtained by fitting Equation 5 as advocated by Ulrich & Vorberg, 2009) and Figure 9B (means obtained by fitting Equation A1) further shows that fitting one or the other equation does not change our conclusion: DLs in the 2AFC task are only minimally (and not always significantly) different from those in the reminder task, which opposes the conclusion of Lapid et al. (2008) regarding DLs that are significantly different and 84% larger in 2AFC tasks. But there is also the question, which we take up next, of whether the theoretical analysis of Ulrich and Vorberg, and the ensuing data-fitting strategy, are generally valid.

Validity of the Constraint PSE = POE

Our analysis in the preceding subsection has not questioned the validity of Ulrich and Vorberg's (2009) constraint that the PSE must equal the POE in 2AFC tasks, because our goal there was merely to show that their approach does not produce DL estimates meaningfully different from those obtained by fitting Equation A1. But the validity of their constraint must be considered carefully.

There are at least two conditions in which the surmise that the PSE equals the POE is incorrect. One of them is restricted to the nonetheless frequent cases in which comparison and standard stimuli differ in more dimensions than that in which their magnitudes are compared. This is always the case in the study of visual illusions, which are defined by the fact that stimuli physically identical in some respect (e.g., the length of two lines in the horizontal-vertical or the Müller-Lyer illusions; the brightness or color of two patches in all sorts of simultaneous contrast illusions, etc.) appear different because of the context in which they are embedded. Research on visual illusions (see, e.g., Armstrong & Marks, 1997; Chacón, 2004; Foley & McCourt, 1985; Jordan & Schiano, 1986) has often used discrimination tasks to provide objective data (i.e., beyond the initial consensus of introspections), indicating that the illusion actually occurs when the PSE is located at the point at which standard and comparison stimuli have the same magnitude along the dimension in which they are compared. Similar evidence that the PSE and the POE do not generally coincide comes also from other common studies involving nonillusory standard and comparison stimuli that differ along noncompared dimensions, like windowed sinusoidal gratings of different orientation, mean luminance, direction of motion, spatial frequency, size, or spatial phase whose contrast is to be compared (see, e.g., Brady & Field, 1995; García-Pérez

& Peli, 2001; Georgeson & Sullivan, 1975; Peli, 1995, 1997). Standard and comparison stimuli in these cases are processed by different networks that may also have different operational characteristics and, then, whether or not the PSE and the POE match cannot be established until data are collected and carefully analyzed without external constraints being imposed.

The second condition in which the assumption that PSE and POE match can be incorrect arises even when standard and comparison stimuli are identical in all other respects. This condition stems from a ubiquitous threat to all types of discrimination experiment. In the words of Fechner (1860/1966, p. 73), "the experiment should therefore be arranged, if possible, so that although the observations follow each other closely, each is uninfluenced by the other." It has indeed been shown that, with comparison and standard stimuli identical in all other respects, too-short intertrial intervals (ITIs) or interstimulus intervals (ISIs) produce a shift of the PSE away from the POE that is eliminated only when ITI and ISI have appropriate durations (Alcalá-Quintana & García-Pérez, 2009). This effect appears to be caused by a temporary sensory impairment reflecting that the processing network has not had time to reset back to baseline conditions during the ISI or the ITI (Buonomano et al., 2009). It should also be noted that the problem cannot be eliminated with ITIs and ISIs that are too long, because memory issues may come into play to produce the same net effect (Wickelgren, 1969). With experience and a careful analysis of data, appropriate ISI and ITI durations can be found for the stimuli and sensory modality under study, but the threat of extraneous factors shifting the location of the PSE in discrimination experiments is always there.

In sum, the validity of the theoretical constraint that the psychometric function for 2AFC discrimination data must have its 50% point when standard and comparison stimuli have the same magnitude—that is, $\Psi_{2AFC}(s) = .5$ —cannot be taken for granted. Then, Ulrich and Vorberg's (2009) approach to fitting 2AFC data under this constraint has very limited applicability.

The final argument against the approach of Ulrich and Vorberg (2009) is purely empirical. Consider the data and fitted psychometric functions in the bottom row of Figure 6. The data in the first three panels in that row clearly indicate that the PSE is above the POE (i.e., whether the standard is always presented first or second, and for the aggregate of those two data sets), and psychometric functions fitted independently to each data set reflect this characteristic. On the other hand, one-sided data from the 2AFC condition (rightmost panel) are too noisy to indicate the location of the PSE, but a psychometric function fitted to these data has its PSE only minimally above the POE. Whether these characteristics are authentic or are only a consequence of sampling error is unclear, but patterns of this type are not uncommon in the data from Lapid et al.'s (2008) Experiments 5 and 6. The consequences of fitting those data with the joint and constrained approach of Ulrich and Vorberg are displayed in the bottom row of Figure 12. Because the predicted psychometric function for 2AFC data (third and fourth panels in the bottom row

of Figure 12) is forced to have its PSE at the POE, and because this function is the average of the two functions jointly fitted to the data in the first and second panels in that row, the data in the two panels on the left are somehow “ignored” to ensure that the fitted psychometric functions have their PSEs on opposite sides of the POE (and not on the same side, as the data actually call for). Maximum-likelihood methods naturally succeed in finding parameters for the fitted psychometric functions, but the agreement with the data is poor: Note that the curves in the first and second panels along the bottom row of Figure 12 fall above almost all of the data points instead of running through them (compare with the bottom row in Figure 6). It is true that the data were scarce (only 10 observations at each of 11 points) and that sampling error is then substantial, but fitting a model by fiat, and against what the data indicate, does not seem the best way to seek empirical evidence on any issue.

A word seems in order about how Ulrich and Vorberg’s (2009) theoretical demonstration that $\Psi_{2AFC}(s) = .5$ can be reconciled with the foregoing discussion, arguing that in most cases this equality does not hold. To avoid a lengthy presentation, we must assume here that the reader is familiar with the basics of Ulrich and Vorberg’s demonstration in their Equations 1–9, although we will use the notation introduced earlier in this section. The discrepancy can be resolved by noting that Ulrich and Vorberg’s demonstration relies on an assumption that does not hold in the cases discussed earlier in this section. In particular, they assumed that $\Psi_1(x) + \Psi_2(x) = 1$ when $x = s$. In plain words, this assumption means that the sensory effects of test and standard stimuli are only determined by their particular magnitude in the dimension along which they are compared. Thus, when test and standard stimuli have the same magnitude, their sensory effects are identical and, in a signal-detection theoretical interpretation, they are indistinguishable. In all of the cases discussed earlier in this section, identity of test and standard in the dimension along which they are compared nevertheless renders different sensory effects; consider the case of a luminance patch, whose perceived brightness depends on the luminance of its surround, so that two patches of the same physical luminance are not perceived to have the same brightness when they are embedded in surrounds of different luminance. Suppose, then, that the average sensory effect of the standard stimulus is $\mu_S(s)$, whereas the average sensory effect of the test stimulus is $\mu_T(x)$, and consider the general case in which the functions μ_S and μ_T depend on dimensions other than that according to which test and standard stimuli are compared. Thus, $x = s$ does not imply $\mu_T(x) = \mu_S(s)$, because mere equality of test and standard stimuli in the dimension in which they are compared (e.g., the luminance of two uniform patches) does not ensure equality of their sensory effects (i.e., their perceived brightness when surrounded by backgrounds of different luminance). In consequence, the assumption that Ulrich and Vorberg should have used in their derivation is that $\Psi_1(x) + \Psi_2(x) = 1$, when $\mu_T(x) = \mu_S(s)$, and not just when $x = s$. Unfortunately, we do not

know whether the functions μ_S and μ_T are identical even when test and standard stimuli are nominally identical in all respects (see our earlier discussion of temporary sensory impairment). In these conditions, the only valid conclusion from Ulrich and Vorberg’s analysis is that $\Psi_{2AFC}(x) = .5$ at some point x , which may or may not be the point $x = s$; then, the PSE may or may not lie at the POE.

Comparison With the Approach Based on the Difference Model With Guessing

It should finally be noted that the approach presented by Ulrich and Vorberg (2009) represents a simplified version of the more general difference model with guessing presented above. Indeed, Ulrich and Vorberg claimed that the psychometric functions Ψ_1 and Ψ_2 , describing discrimination performance when the standard is presented first or second, may have different parameters. This is actually the situation illustrated for the difference model with guessing in Figure 10B, with the additional characteristic that this model actually describes how and why those two psychometric functions differ.¹⁹ Next, under the difference model with guessing, the psychometric function when presentation order is randomized is given by the average shown in Figure 10C. This is also what Ulrich and Vorberg claim it should be. Other than these superficial similarities, the difference model with guessing does not impose the constraint that the PSE must equal the POE (which, as discussed above, will rarely hold) upon fitting a psychometric function, and it indicates how the factors causing the differences between Ψ_1 and Ψ_2 can be eliminated. In other words, the difference model with guessing allows an experimental elimination of the factors responsible for the differences between Ψ_1 and Ψ_2 . The approach of Ulrich and Vorberg, on the other hand, can only aim to eliminate those influences statistically, something that can only be done on the rare occasions (which are also impossible to predict in advance) when the constraint that PSE and POE coincide holds empirically.

CONCLUSION

We have reanalyzed the data of Lapid et al. (2008), mostly with the purpose of fixing an error in their estimation of the DL with the 2AFC task, an error that arises from fitting an incorrect equation to 2AFC data. We have also fixed a more general error that affects estimates from either task and that arises from disregarding the lapsing-rate parameter that should be included in the fitted psychometric function. Our reanalysis has thus consisted of fitting Equation A1 to data from the reminder and the 2AFC tasks, with the result that differences between DL estimates from either task are minimal and do not always reach statistical significance across the set of experiments that we have reanalyzed. These results and conclusion hold also when parameters are estimated by fitting instead a constrained equation proposed by Ulrich and Vorberg (2009), which they claimed is the proper way to analyze 2AFC data (although we have also argued against such claims).

We have also shown that Lapid et al.'s (2008) data are contaminated by response bias, which has the nagging consequence of masquerading as variations in estimated PSEs and DLs across tasks. In a discrimination experiment, the difference model with guessing described in this article shows that response bias has different manifestations when the standard is presented first and when it is presented second, and that has yet another manifestation when the order of presentation of standard and comparison is randomized across trials. These different manifestations actually produce the effects on estimated DLs reported by Lapid et al. and that remained (although substantially weakened) under the reanalysis presented in this article. But these effects should not be attributed to the nominally different tasks (reminder vs. 2AFC), but rather to the presence of response bias. Indeed, the difference model with guessing shows that elimination of response bias renders psychometric functions whose shape (and, hence, whose associated PSEs and DLs) is the same across presentation modalities (fixed or randomized) and, hence, across nominal tasks. Unfortunately, elimination of response bias requires knowledge of the trials in which each observer guessed, information not recorded in the experiments of Lapid et al., so rendering their data hopelessly contaminated.²⁰ In these conditions, any conclusion about differences between DLs estimated with reminder and 2AFC tasks must be questioned. It is interesting to note in this respect that differences between DLs across tasks come out as significant, even under our reanalysis for data from Experiments 1 and 2, each of which tested one of the tasks using a different group of observers. Then, the differences could equally reasonably be attributed to the tasks themselves (as Lapid et al. did), or to different response biases in each group of observers. Various sources of evidence support the tenability of the latter explanation. First, the average DL in the reminder task of Experiment 1 differed significantly from that obtained by a different group of observers under the identical reminder task in Experiment 5 who, in turn, obtained an average DL almost identical to that of yet another different group of observers in the 2AFC task of Experiment 2. Second, Experiment 5 tested both tasks (referred to as R1 and FC in Figure 9) in the same group of observers and revealed no significant differences between DLs estimated with either task. It should be stressed, however, that none of these conclusions is solid when the effects of response bias are present in the data. And the same is true for the superior discrimination performance that observers showed in the R1 condition compared with the R2 condition of Experiment 6 under our reanalysis, a superiority not corroborated in Experiment 5 (also under our reanalysis).

Finally, we should stress that the difference model with guessing can be used to eliminate the confounding effects of response bias from any study that seeks a fair comparison of the relative efficacy of alternative psychophysical methods. Also, to obtain dependable estimates of the parameters of the fitted psychometric functions in this or other cases and, hence, of the PSE and the DL, at least 200 trials should be administered with a carefully designed sampling plan (see García-Pérez & Alcalá-Quintana, 2005).

AUTHOR NOTE

We thank Einat Lapid for sharing the data from her study, and Rolf Ulrich and two anonymous reviewers for their comments on an earlier draft of this article. This research was supported by Grant PSI2009-08800 from Ministerio de Ciencia e Innovación (Spain). Correspondence concerning this article should be sent to M. A. García-Pérez, Departamento de Metodología, Facultad de Psicología, Universidad Complutense, Campus de Somosaguas, 28223 Madrid, Spain (e-mail: miguel@psi.ucm.es).

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.
- ALCALÁ-QUINTANA, R., & GARCÍA-PÉREZ, M. A. (2009). *A model for the time-order error in contrast discrimination*. Manuscript submitted for publication.
- ARMSTRONG, L., & MARKS, L. E. (1997). Differential effects of stimulus context on perceived length: Implications for the horizontal-vertical illusion. *Perception & Psychophysics*, **59**, 1200-1213.
- BIRD, C. M., HENNING, G. B., & WICHMANN, F. A. (2002). Contrast discrimination with sinusoidal gratings of different spatial frequency. *Journal of the Optical Society of America A*, **19**, 1267-1273. doi:10.1364/JOSAA.19.001267
- BRADY, N., & FIELD, D. J. (1995). What's constant in contrast constancy? The effects of scaling on the perceived contrast of band-pass patterns. *Vision Research*, **35**, 739-756. doi:10.1016/0042-6989(94)00172-1
- BUONOMANO, D. V., BRAMEN, J., & KHODADADIFAR, M. (2009). Influence of the interstimulus interval on temporal processing and learning: Testing the state-dependent network model. *Philosophical Transactions of the Royal Society B*, **364**, 1865-1873. doi:10.1098/rstb.2009.0019
- CHACÓN, J. (2004). Perceived contrast explains asymmetries in visual-search tasks with shaded stimuli. *Perception*, **33**, 1499-1509. doi:10.1068/p5286
- FECHNER, G. T. (1966). *Elements of psychophysics: Vol. 1*. New York: Holt, Rinehart & Winston. (Original work published 1860)
- FOLEY, J. M., & MCCOURT, M. E. (1985). Visual grating induction. *Journal of the Optical Society of America A*, **2**, 1220-1230. doi:10.1364/JOSAA.2.001220
- FOLEY, J. M., & SCHWARZ, W. (1998). Spatial attention: Effect of position uncertainty and number of distractor patterns on the threshold-versus-contrast function for contrast discrimination. *Journal of the Optical Society of America A*, **15**, 1036-1047. doi:10.1364/JOSAA.15.001036
- GARCÍA-PÉREZ, M. A. (2010). Denoising forced-choice detection data. *British Journal of Mathematical & Statistical Psychology*, **63**, 75-100. doi:10.1348/000711009X424057
- GARCÍA-PÉREZ, M. A., & ALCALÁ-QUINTANA, R. (2005). Sampling plans for fitting the psychometric function. *Spanish Journal of Psychology*, **8**, 256-289. Available at www.ucm.es/sjp.
- GARCÍA-PÉREZ, M. A., & ALCALÁ-QUINTANA, R. (2007). The transducer model for contrast detection and discrimination: Formal relations, implications, and an empirical test. *Spatial Vision*, **20**, 5-43. doi:10.1163/156856807779369724
- GARCÍA-PÉREZ, M. A., & ALCALÁ-QUINTANA, R. (2009a). *The difference model with guessing explains interval bias in two-alternative forced-choice detection procedures*. Manuscript submitted for publication.
- GARCÍA-PÉREZ, M. A., & ALCALÁ-QUINTANA, R. (2009b). Fixed vs. variable noise in 2AFC contrast discrimination: Lessons from psychometric functions. *Spatial Vision*, **22**, 273-300. doi:10.1163/156856809788746309
- GARCÍA-PÉREZ, M. A., ALCALÁ-QUINTANA, R., WOODS, R. L., & PELI, E. (2010). *Psychometric functions for detection and discrimination with and without flankers*. Manuscript submitted for publication.
- GARCÍA-PÉREZ, M. A., & PELI, E. (2001). Intrascopic perception. *Journal of Neuroscience*, **21**, 7313-7322.
- GEORGESON, M. A., & MEESE, T. S. (2006). Fixed or variable noise in contrast discrimination? The jury's still out. . . . *Vision Research*, **46**, 4294-4303. doi:10.1016/j.visres.2005.08.024

- GEORGESON, M. A., & SULLIVAN, G. D. (1975). Contrast constancy: Deblurring in human vision by spatial frequency channels. *Journal of Physiology*, **252**, 627-656.
- HAYES, A. F., & CAI, L. (2007). Further evaluating the conditional decision rule for comparing two independent means. *British Journal of Mathematical & Statistical Psychology*, **60**, 217-244. doi:10.1348/000711005X62576
- HELLSTRÖM, Å. (1979). Time errors and differential sensation weighting. *Journal of Experimental Psychology: Human Perception & Performance*, **5**, 460-477. doi:10.1037/0096-1523.5.3.460
- HELLSTRÖM, Å. (1985). The time-order error and its relatives: Mirrors of cognitive processes in comparing. *Psychological Bulletin*, **97**, 35-61. doi:10.1037/0033-2909.97.1.35
- HELLSTRÖM, Å. (2003). Comparison is not just subtraction: Effects of time- and space-order on subjective stimulus difference. *Perception & Psychophysics*, **65**, 1161-1177.
- HENNING, G. B., BIRD, C. M., & WICHMANN, F. A. (2002). Contrast discrimination with pulse trains in pink noise. *Journal of the Optical Society of America A*, **19**, 1259-1266. doi:10.1364/JOSAA.19.001259
- JÄKEL, F., & WICHMANN, F. A. (2006). Spatial four-alternative forced-choice method is the preferred psychophysical method for naïve observers. *Journal of Vision*, **6**, 1307-1322. doi:10.1167/6.11.13
- JORDAN, K., & SCHIANO, D. J. (1986). Serial processing and the parallel-lines illusion: Length contrast through relative spatial separation of contours. *Perception & Psychophysics*, **40**, 384-390.
- KAERNBACH, C. (2001). Adaptive threshold estimation with unforced-choice tasks. *Perception & Psychophysics*, **63**, 1377-1388.
- KATKOV, M., TSODYKS, M., & SAGI, D. (2006a). Analysis of a two-alternative forced-choice signal detection theory model. *Journal of Mathematical Psychology*, **50**, 411-420. doi:10.1016/j.jmp.2005.11.002
- KATKOV, M., TSODYKS, M., & SAGI, D. (2006b). Singularities in the inverse modeling of 2AFC contrast discrimination data. *Vision Research*, **46**, 259-266. doi:10.1016/j.visres.2005.09.022
- KLEIN, S. A. (2006). Separating transducer non-linearities and multiplicative noise in contrast discrimination. *Vision Research*, **46**, 4279-4293. doi:10.1016/j.visres.2006.03.032
- KONTSEVICH, L. L., CHEN, C.-C., & TYLER, C. W. (2002). Separating the effects of response nonlinearity and internal noise psychophysically. *Vision Research*, **42**, 1771-1784. doi:10.1016/S0042-6989(02)00091-3
- LAPID, E., ULRICH, R., & RAMMSAYER, T. (2008). On estimating the difference limen in duration discrimination tasks: A comparison of the 2AFC and the reminder task. *Perception & Psychophysics*, **70**, 291-305. doi:10.3758/PP.70.2.291
- LINK, S. W. (1992). *The wave theory of difference and similarity*. Hillsdale, NJ: Erlbaum.
- LIU, C. C., & SMITH, P. L. (2009). Comparing time-accuracy curves: Beyond goodness-of-fit measures. *Psychonomic Bulletin & Review*, **16**, 190-203. doi:10.3758/PBR.16.1.190
- MEESE, T. S. (1995). Using the standard staircase to measure the point of subjective equality: A guide based on computer simulations. *Perception & Psychophysics*, **57**, 267-281.
- MILLER, J., & ULRICH, R. (2001). On the analysis of psychometric functions: The Spearman-Kärber method. *Perception & Psychophysics*, **63**, 1399-1420.
- NUMERICAL ALGORITHMS GROUP (1999). *NAG Fortran library manual, mark 19*. Oxford: Author.
- PELI, E. (1995). Suprathreshold contrast perception across differences in mean luminance: Effects of stimulus size, dichoptic presentation, and length of adaptation. *Journal of the Optical Society of America A*, **12**, 817-823. doi:10.1364/JOSAA.12.000817
- PELI, E. (1997). In search of a contrast metric: Matching the perceived contrast of Gabor patches at different phases and bandwidths. *Vision Research*, **37**, 3217-3224. doi:10.1016/S0042-6989(96)00262-3
- POGGEL, D. A., STRASBURGER, H., & MACKEBEN, M. (2007). Cueing attention by relative motion in the periphery of the visual field. *Perception*, **36**, 955-970. doi:10.1068/p5752
- POKORNY, R. A. (1998). Using maximum-likelihood adaptive methods to estimate difference thresholds and points of subjective equality. *Behavior Research Methods, Instruments, & Computers*, **30**, 624-636.
- SUMMERS, R. J., & MEESE, T. S. (2009). The influence of fixation points on contrast detection and discrimination of patches of grating: Masking and facilitation. *Vision Research*, **49**, 1894-1900. doi:10.1016/j.visres.2009.04.027
- ULRICH, R., & MILLER, J. (2004). Threshold estimation in two-alternative forced-choice (2AFC) tasks: The Spearman-Kärber method. *Perception & Psychophysics*, **66**, 517-533.
- ULRICH, R., & VORBERG, D. (2009). Estimating the difference limen in 2AFC tasks: Pitfalls and improved estimators. *Attention, Perception, & Psychophysics*, **71**, 1219-1227. doi:10.3758/APP.71.6.1219
- VAN VLEET, T. M., & ROBERTSON, L. C. (2006). Cross-modal interactions in time and space: Auditory influence on visual attention in hemispatial neglect. *Journal of Cognitive Neuroscience*, **18**, 1368-1379. doi:10.1162/jocn.2006.18.8.1368
- VON WALDKIRCH, M., LUKOWICZ, P., & TRÖSTER, G. (2004). Multiple imaging technique for extending depth of focus in retinal displays. *Optics Express*, **12**, 6350-6355. doi:10.1364/OPEX.12.006350
- VON WALDKIRCH, M., LUKOWICZ, P., & TRÖSTER, G. (2005). Oscillating fluid lens in coherent retinal projection displays for extending depth of focus. *Optics Communications*, **253**, 407-418. doi:10.1016/j.optcom.2005.05.006
- WERKHOVEN, P., & SNIPPE, H. P. (1996). An efficient adaptive procedure for psychophysical discrimination experiments. *Behavior Research Methods, Instruments, & Computers*, **28**, 556-562.
- WESTHEIMER, G. (1999). Discrimination of short time intervals by the human observer. *Experimental Brain Research*, **129**, 121-126. doi:10.1007/s002210050942
- WICHMANN, F. A., & HILL, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, **63**, 1293-1313.
- WICKELGREN, W. A. (1969). Associative strength theory of recognition memory for pitch. *Journal of Mathematical Psychology*, **6**, 13-61. doi:10.1016/0022-2496(69)90028-5
- ZIMMERMAN, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical & Statistical Psychology*, **57**, 173-181. doi:10.1348/000711004849222

NOTES

1. Because of the response requested from the observer, the reminder task could be categorized as a yes-no task with a preceding (or following) standard. Note, however that a 2AFC task could also request the same response from the observer by asking the exact same question, but it is unclear that a 2AFC task would thus be regarded as a yes-no task with a random order of presentation of standard and comparison. This route to categorization is likely to create confusion. It is perhaps clearer to use a classification system that leaves aside the question that is asked to the observer. In this system (which is in common use in visual psychophysics), a yes-no task comprises a single presentation interval and involves no comparison, whereas a 2AFC task consists of two presentation intervals and requires a comparison. From this point of view, the reminder task is only a 2AFC task with the peculiarity that the order of presentation of standard and comparison is fixed.

2. It should be incidentally noted that this second characteristic is not truly a requirement of the 2AFC task. Although this treatment should not deteriorate the 2AFC task itself, it might certainly reduce the accuracy with which the parameters of a psychometric function can be estimated, because the psychometric function is fitted to data covering only half its domain.

3. This is not always the case. In Lapid et al.'s (2008) Experiments 5 and 6 (and presumably also in their Experiments 2, 3, and 4) the question that the observers had to respond to on every trial was exactly the same in 2AFC and reminder tasks.

4. Recall that Lapid et al. (2008) fitted Equation 2 with x defined actually as the difference $x - s$, so that parameter a becomes the DL (i.e., the 75% point), although in their Figure 5 they plotted Equation 2 as a function of the magnitude x of the comparison stimulus and not as a function of the difference $x - s$. For consistency with Equation 1, we will continue to regard Equation 2 as a function of the magnitude x of the comparison stimulus. The only consequence of this decision is that the DL is then defined as $a - s$.

5. Lapid et al. (2008) defined the DL in these cases as half the distance between the 25% and the 75% points, but this is indeed the distance between the 50% and the 75% points (or between the 25% and the 50% points), given the symmetry of the logistic function fitted to the data.

6. At first glance, this statement may be regarded as incorrect in the light of actual practices followed in other articles. In particular, there is some tradition in the vision research literature to measure 2AFC contrast discrimination performance using magnitudes for the comparison stimulus that are always above that of the standard stimulus, hence yielding also one-sided data. In these cases, a psychometric function like that in Equation 2 has occasionally been fitted to this type of data (e.g., Bird, Henning, & Wichmann, 2002; Henning, Bird, & Wichmann, 2002; Summers & Meese, 2009), but it should be realized that in those occasions the psychometric function was not considered a function of the log contrast x of the comparison stimulus, but rather as a function of the logarithm of the increment contrast—that is, as a function of $\log(10^x - 10^s)$. Elsewhere, we have shown (see García-Pérez & Alcalá-Quintana, 2007, particularly their Figure 1) that transforming the independent variable in this way is formally equivalent to taking the right half (i.e., for $x \geq a$) of the psychometric function in Equation 1, labeling the horizontal axis for the resultant half function with values at or above zero (i.e., defining the variable on the horizontal axis as $x - s$), and then replotting the function on a logarithmic horizontal axis, which then yields the shape of Equation 2. The final logarithmic transformation that justifies this approach was not applied by Lapid et al. (2008), who still plotted their data and fitted Equation 2 on the original horizontal axis representing the difference in magnitude between the comparison and the standard stimuli.

7. Ulrich and Vorberg (2009) have recently developed a theoretical argument whereby 2AFC discrimination data should not be fitted by a function similar to that in Equation 1. A detailed discussion of this issue will be deferred to a later section of this article, where we will show that Equation 1 is indeed appropriate for 2AFC discrimination data, even considering the theoretical argument of Ulrich and Vorberg.

8. Rounded to the nearest integers, the estimates reported for the same data in Figure 4A are 537 msec (for the PSE) and 20 msec (for the DL). Differences from the values reported by Lapid et al. (2008) are surely due to the different software and algorithm used to obtain these maximum-likelihood estimates.

9. That the number of trials used to fit each individual psychometric function by Lapid et al. (2008) was too small shows in the test–retest correlations reported in their Table 1. These correlations ranged between .62 and .76, which, although significantly different from zero, reflect that parameter estimates obtained from only 100 trials (in the 2AFC task) or 110 trials (in the reminder task) are subject to large sampling variability (see García-Pérez & Alcalá-Quintana, 2005). And there is the additional issue that one-sided data must provide parameter estimates that are even more affected by sampling variability.

10. These sample sizes result from discarding unpatterned data sets with the same criterion described for Figure 2 above.

11. Part of this effect is due to the fact that estimates from the R condition are based on twice as many trials as those used in the R1 and R2 conditions.

12. The fact that differences between DLs estimated with the reminder and the 2AFC tasks are not significant in Experiment 5 but significant in Experiment 6 might also be caused by the different sensory modalities (auditory and visual) involved in each case. This would also imply that the differences cannot be attributed solely to the different tasks.

13. Whether the variance of sensory effects is constant or varies with stimulus level does not change the results of our subsequent analysis, which is why we adopt this typical simplification. In any case, it has been shown in the context of visual contrast perception that discrimination data can be equally accounted for, whether variance is constant or not (see García-Pérez & Alcalá-Quintana, 2009b). But this result is not limited to visual contrast perception: A formal analysis by Katkov, Tsodyks, and Sagi (2006a, 2006b) has demonstrated a singularity that makes equal-variance and unequal-variance models functionally identical when it comes to predicting the shape of the psychometric function—which is actually our only goal here.

14. This linear function is perhaps unrealistic, but it is simple enough for the present illustration. A more realistic expression will alter only the final mathematical form of the psychometric function, but the way in which response bias affects PSEs and DLs will remain unchanged.

15. It is important to stress that equal overall numbers of trials in which the comparison stimulus is presented first or second do not suffice. It is also required that this equality hold at every individual magnitude that the comparison stimulus had. If only, say, six trials are administered with the comparison stimulus at some particular magnitude, and the standard stimulus happened to be presented first in all of them, the resultant data point will still be contaminated by response bias.

16. When used in conjunction with adaptive methods (in which the magnitude of the comparison stimulus in the next trial depends on the response given by the observer in the current trial), further action is needed, consisting simply of deciding whether the response of the observer would have been correct or incorrect had an actual random guess been produced. A simulated coin-toss process is sufficient for this purpose, whose outcome thus allows setting the magnitude of the comparison stimulus for the next trial.

17. Although Equation 3 is monotone increasing, in their Figures 2 and 3 Ulrich and Vorberg (2009) plotted $1 - \Psi_{2AFC}(x)$ for $x < s$ and $\Psi_{2AFC}(x)$ for $x > s$, which renders a nonmonotonic function. The reason is that they understand the ordinate of the psychometric function as reflecting probability correct in the strict physical sense that the observer reports the standard when the standard has longer duration and reports the comparison when it is the comparison that has a longer duration. The typical approach at least in vision and auditory research is to regard the psychometric function for 2AFC discrimination tasks as indicating the probability that the comparison is reported to have a larger magnitude, which is what the monotonic psychometric function in Equation 3 expresses. A nonmonotonic function like that produced by Ulrich and Vorberg is overly inconvenient for a number of reasons, including the impossibility of using adaptive methods with comparison levels that can be both above and below the standard level (see Meese, 1995; Pokorný, 1998; Werkhoven & Snippe, 1996).

18. Even under the usual assumptions of normally distributed sensory effects with equal variance under signal detection theory, the psychometric function has the form of a cumulative Gaussian only if the transducer function is linear, as in the simple illustration of Figure 10. But transducer functions are typically nonlinear, and they always render psychometric functions with a complex mathematical expression (see García-Pérez & Alcalá-Quintana, 2007).

19. Recall that the illustration in Figure 10 presented a simple version of the difference model with guessing for clarity. The model also incorporates lapsing-rate parameters and sensory effects potentially governed by different functions μ_T and μ_S (for test and standard stimuli). With all of these, psychometric functions for different presentation orders also have different asymptotes and slopes and, hence, they do not differ only in location as in the simplified illustration of Figure 10, and they do not necessarily have the PSE at the POE either (see Alcalá-Quintana & García-Pérez, 2009).

20. This is true for the difference model with guessing that we have considered in this article. If the difference model with bias were adopted, the data of Lapid et al. (2008) could be analyzed to estimate the criterion parameter c that best accounts for each observer's response bias. However, this is easier said than done, because off-the-shelf functional forms for the psychometric function (such as that in Equation A1) do not include c among their parameters. Then, the functional form of the psychometric function arising from the difference model with bias would have to be derived by making explicit assumptions about the distribution of sensory effects and, more importantly, about the relation between stimulus level and the mean of that distribution (for a detailed example in the case of the difference model with guessing, see Alcalá-Quintana & García-Pérez, 2009).

APPENDIX

Maximum-Likelihood Estimation

Psychometric functions were fitted by seeking maximum-likelihood estimates for their parameters through NAG subroutine E04JYF (Numerical Algorithms Group, 1999), which allows constrained optimization. Generally, all data (whether coming from reminder or 2AFC tasks) were fitted to an amended version of Equation 1 that incorporates a lapsing-rate parameter λ representing the probability that an observer hits the wrong key by mistake upon responding. It has been shown that lapses can have dramatic effects on the width of the fitted psychometric function (see Figure 1 of Wichmann & Hill, 2001) if they are not compensated for by way of this extra parameter. Hence, omitting this parameter may occasionally produce spuriously large estimates of the DL because the DL is indeed a measure of the width of the psychometric function. The form of the fitted psychometric function was then

$$\Psi(x; a, b, \lambda) = \lambda + \frac{1 - 2\lambda}{1 + \exp[-(x - a) / b]}, \quad (\text{A1})$$

and it can easily be seen that $\lambda = 0$ turns Equation A1 into Equation 1, whereas $\lambda \neq 0$ renders a scaled psychometric function that ranges from λ to $1 - \lambda$. The mathematical form in Equation A1 arises on consideration that observers incur lapses with identical probability at the low and high ends of x . In the general case, the psychometric function will range from a lower asymptote at γ (which stands for the probability of a lapse when $x \ll s$) to an upper asymptote at $1 - \lambda$ (which stands for the probability of *not* lapsing when $x \gg s$), but we will only consider the simpler case of Equation A1. Whether for data from reminder or 2AFC tasks, x in Equation A1 represents the magnitude of the comparison stimulus (as illustrated in Figure 1). Parameters were estimated under the permissive constraints $100 \leq \hat{a} \leq 900$, $4 \leq \hat{b} \leq 150$, and $0 \leq \hat{\lambda} \leq .24$; then in all cases the estimated PSE is simply $\hat{a}$ and the estimated DL is $\hat{b} \log[(.75 - \hat{\lambda}) / (.25 - \hat{\lambda})]$, which is easily seen to reduce to the expression used by Lapid et al. (2008) with Equation 1 when $\hat{\lambda} = 0$ and to compensate for the effect of lapses when $\hat{\lambda} \neq 0$.

We occasionally aimed to reproduce the analyses of Lapid et al. (2008), and in these cases we fitted instead Equation 1 to data from the reminder task and Equation 2 to data from the 2AFC task, although in the latter case we continued to consider x to be the magnitude of the comparison stimulus (and not the difference between this magnitude and that of the standard), which is inconsequential. An example of the latter approach to fitting Equation 2 to 2AFC data is shown in Figure 1; an example of the approach to fitting Equation 1 to data from the reminder task is shown in Figure 4. In these cases, the PSE and the DL were computed from parameter estimates, as was done by Lapid et al.