Supplemental Material: Optimizing the rapid measurement of detection thresholds in infants

Pete R. Jones, Sarah Kalwarowsky, Oliver J. Braddick, Janette Atkinson, and Marko Nardini

1. Asymmetries in threshold sampling distributions

In the main manuscript, it was shown that thresholds estimated far from the mid-point of the psychometric function become progressively inaccurate and imprecise (see Figure 4). For experimenters looking to average data across multiple observers, an important further consideration is that the mid-point of the psychometric function may vary between individuals.

Such a situation is shown for a simulated cohort of hypothetical observers in **Figure S1** (main panel). For example, consider a staircase targeting a threshold of 50% performance (e.g., a 1-up, 1-down staircase; bottom subpanel of Figure S1). For a simulated infant with a lapse/guess rate equal to the group-mean, 50% performance corresponded to a stimulus of $\sim$7 cycles-per-degree. However, because of the differences in lapse/guess rate parameters, some simulated infants exhibited higher (better) thresholds, while some infants exhibited lower (poorer) thresholds. Notably, this distribution of thresholds is not Gaussian (although lapse and guess rates were), and exhibits a slight leftward (negative) skew [$\text{Skew}_{50\%} = -0.86$]. Consequently, the group-mean threshold is pulled downwards (mean $<$ median). This meant in practice that the mean threshold (green vertical line) was an underestimate of the group's average sensitivity.

Given the particular psychometric functions simulated in **Figure S1**, the skew at 50% was relatively slight. However, consider now a staircase targeting a threshold of 75% correct performance (e.g., a weighted up-2 down-1 staircase; top subpanel of Figure S1). For an observer with a lapse/guess rate equal to the group-mean, this corresponded to a stimulus of $\sim$4 cycles-per-degree. However, again there was inter-individual variability, and again this led to a negatively skewed distribution of threshold estimates [$\text{Skew}_{75\%} = -0.46$]. Moreover, in simulated observers with lapse rates of 0.25 or greater, no stimulus magnitude was sufficient to attain 75% correct performance. In these cases the adaptive staircase tended downwards towards floor (zero). This resulted in a multimodal distribution of thresholds, with one negatively skewed main component, and a substantial near-zero component. Together, both of these components serve to pull the group-mean threshold downwards (mean $<<$ median), meaning that the mean threshold (green vertical line) was a profound underestimate of the group's average sensitivity.

In short, if one is averaging over results obtained from a number of individuals, and if these individuals differ in their underlying psychometric function, then more extreme thresholds will produce increasingly asymmetric sampling-distributions of threshold. A practical consequence of this is that the arithmetic mean of the threshold data will no longer capture the central tendency of the observers. Failure to take account of this could lead, for example, for an author to erroneously report a difference in mean sensitivity between two groups of participants, when they in fact differ only in terms of inter-individual variability in lapse rates.

These deviations from normality also mean, as discussed previously by other authors, that "confidence limits cannot readily be calculated from commonly used

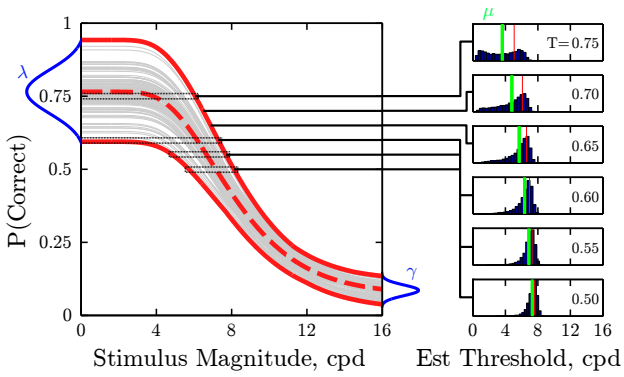

Fig S1. Threshold sampling distribution for individuals with varying guess-/lapse-rate parameters. An adaptive staircase algorithm was used to estimate the thresholds of 10,000 simulated observers. As shown in the main plot, each simulated observer had an expected lapse rate, λ, of 0.25, and an expected guess rate, γ, of 0.07. However, these values were randomly jittered between individuals, using values drawn from a zero-mean Gaussian distribution with a standard deviation 0.075 and 0.02, respectively (blue curves). Six threshold estimates were made within each observer, one at each of six target threshold levels (Percent correct: $50\%, 55\%, \ldots, 75\%$). Each threshold was estimated by geometrically averaging the last 64 reversals of a 256-reversal, weighted staircase. The resultant sampling distributions of threshold estimates are shown in the individual subpanels. Threshold estimates exhibited an increasingly negative skew as the threshold target lay further from the mid-points of the underlying psychometric functions. Because of this asymmetry in threshold estimates, the arithmetic mean of the data (green vertical lines) was often a poor summary statistic of the typical threshold of the group.

(symmetrical) formulas such as ± 1.96 Standard Error" (McKee et al., 1985), and that comparisons of confidence intervals computed in this way are liable to be highly misleading. Thus, one recommended practice is to always use numerical bootstrapping methods to compute confidence intervals are threshold estimates (DiCiccio and Efron, 1996).

1.1. Averaging threshold data generated from a single psychometric function

It might be expected that when averaging data across one or more individuals whose psychometric functions are identical, then a skewed threshold sampling distribution would be less of a concern (**Figure S2**). However, if targeting thresholds at the very extremities of the psychometric function (**Figure S2**, $T = 0.75$), then a proportion of staircases may nonetheless fail to converge on threshold, again causing the expected mean of multiple threshold estimates to become misleading. In such instances, when combining thresholds across multiple estimates, it may be best to transform the data first, or to simply take the highest threshold estimates, as discussed in the main manuscript.

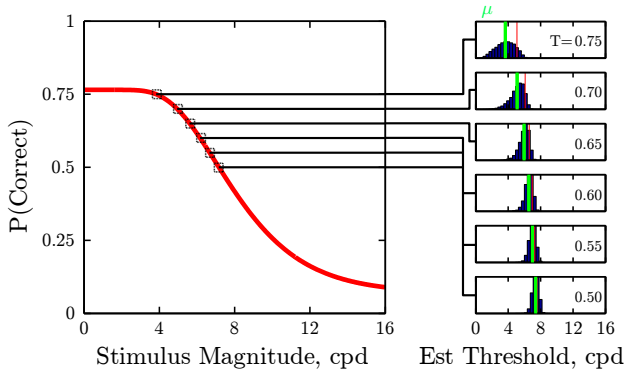

Fig S2. Threshold sampling distribution for individuals with invariant guess-/lapse-rate parameters. Same format as **Figure S1**.

1.2. Averaging threshold data derived from smaller numbers of trials

It is important to note that, in the foregoing, a very large number of trials were used to estimate thresholds, with each adaptive staircase lasting many hundreds or thousands of trials. As shown in **Figure S3**, when fewer trials are used (as per most real experiments), threshold estimates become more volatile, and asymmetries in the sampling distribution are accentuated. With finite trials it is therefore even more important to avoid targeting extreme levels of percent correct performance to derive unbiased estimates of group-mean sensitivity.

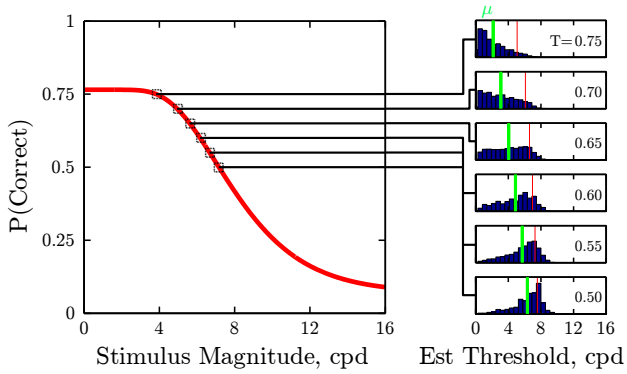

Fig S3. Threshold sampling distribution for individuals with invariant guess-/lapse-rate parameters, using short adaptive trial sequences. Data were generated as per **Figure S2**, except that thresholds were computed by averaging over the second four of an eight reversal adaptive staircase (i.e., rather than the last 64 reversals of a 256-reversal staircase).

Supplemental References

DiCiccio, T. J. and Efron, B. (**1996**). Bootstrap confidence intervals. *Statistical Science*, 11(3):189–212.

McKee, S. P., Klein, S. A., and Teller, D. Y. (**1985**). Statistical properties of forced-choice psychometric functions: Implications of probit analysis. *Perception & Psychophysics*, 37(4):286–298.