LETTER **Communicated by Sergei Gepshtein**

Model Based or Model Free? Comparing Adaptive Methods for Estimating Thresholds in Neuroscience

Julien Audiffren
julien.audiffren@unifr.ch
Jean-Pierre Bresciani
jean-pierre.bresciani@unifr.ch
Department of Neuroscience, Fribourg University, 1700 Fribourg, Switzerland

The quantification of human perception through the study of psychometric functions Ψ is one of the pillars of experimental psychophysics. In particular, the evaluation of the threshold is at the heart of many neuroscience and cognitive psychology studies, and a wide range of adaptive procedures has been developed to improve its estimation. However, these procedures are often implicitly based on different mathematical assumptions on the psychometric function, and unfortunately, these assumptions cannot always be validated prior to data collection. This raises questions about the accuracy of the estimator produced using the different procedures. In the study we examine in this letter, we compare five adaptive procedures commonly used in psychophysics to estimate the threshold: Dichotomous Optimistic Search (DOS), Staircase, PsiMethod, Gaussian Processes, and QuestPlus. These procedures range from model-based methods, such as the PsiMethod, which relies on strong assumptions regarding the shape of Ψ, to model-free methods, such as DOS, for which assumptions are minimal. The comparisons are performed using simulations of multiple experiments, with psychometric functions of various complexity. The results show that while model-based methods perform well when Ψ is an ideal psychometric function, model-free methods rapidly outshine them when Ψ deviates from this model, as, for instance, when Ψ is a beta cumulative distribution function. Our results highlight the importance of carefully choosing the most appropriate method depending on the context.

1 Introduction

Psychophysics methods are used to investigate the relationship between physical stimuli and the subjective percepts or responses they elicit. Aside from underlying basic research in neuroscience and human perception, psychophysics methods have widespread applications, spanning from the study of attention (Scheuerman et al., 2017), to the conception of allowing systems to compress audio signals without altering perceived signal

Neural Computation **34, 338–359** (2022)
https://doi.org/10.1162/neco_a_01461

quality (see Dreschler & Verschuure, 1996; Zwicker, 2000), or to the evaluation of treatments for pain relief (Nir et al., 2011). Psychophysics experiments to study human perception usually consist of presenting the observer with a sequence of stimuli of varying intensity (e.g., loudness, frequency, brightness, contrast) and measuring the response associated with each stimulus. The overall performance of the observer is then summarized using a psychometric function Ψ, which encodes the proportion of stimuli detected by the observer as a function of the intensity.

Frequently, the psychometric function is derived from the experimental results using a Bayesian approach. In this case, Ψ is assumed to be a two-parameter function of a given family, such as the Weibull cumulative distribution function (c.d.f.), and its parameters are fitted to the collected data, using, for instance, the maximum likelihood principle (Kontsevich & Tyler, 1999). In some situations, however, the experimenters are interested in determining a specific landmark of Ψ rather than in estimating the entire function. A particularly important landmark is the threshold, noted s_*, where the stimulus is just noticeable (Benson, Hutt, & Brown, 1989). Estimating this threshold is arguably one of the most common targets of psychophysics experiments, notably because this value can be informative about the underlying sensory and perceptual processes. While it is possible to obtain this threshold by first estimating the entire psychometric function and then computing the inverse of that function for the desired performance level, this approach is demanding and tends to require a considerable amount of experimental data (Song, Garnett, & Barbour, 2017). Consequently, multiple procedures have been proposed to estimate directly the threshold without estimating Ψ. Among these procedures, the most commonly used is arguably the Staircase method (Levitt, 1971).

Another key aspect of psychophysics experiments lies in the choice of a sequence of stimuli to present to the observer. Note that ideally, this sequence should be as short as possible in order to limit fatigue-evoked bias (Wichmann & Hill, 2001a). One commonly used technique is the method of constant stimuli (Wichmann & Hill, 2001a), which consists of presenting the observer with a sequence of stimuli, spanning from imperceptible to consistently perceptible. However, this approach suffers from many limitations, particularly when the objective is to estimate the threshold, notably due to the fact that the sequence of stimuli is independent of the observer's responses. First, the sequence may include points that might be irrelevant to the estimation of the threshold. Second, the fixed sequence does not account for the differences of individual observers. This is a key problem, as the threshold can vary by a factor of ten between individuals (Benson et al., 1989). Consequently, there has been an increased interest in using adaptive algorithms consisting of dynamically adapting stimulus intensity to the observer's responses (Leek, 2001). In particular, multiple adaptive psychophysics procedures have been proposed, such as the Staircase (Levitt, 1971) and the PsiMethod (Kontsevich & Tyler, 1999). These procedures

include an adaptive algorithm that determines the sequence of stimuli to be presented to the observer. This algorithm takes into account both the parameters of the experiments and the past responses of the observer. Adaptive procedures also include a method to compute an estimator of s_* at the end of the experiment.

While adaptive procedures share the same objectives, the scope of their assumptions regarding Ψ varies broadly. Importantly, the general setting of the psychophysics experiment yields almost no assumption regarding Ψ. Specifically, the only commonly accepted hypotheses are that Ψ is always nondecreasing (the stronger the stimulus is, the easier it is to perceive) and continuous (Leek, 2001). But these hypotheses are insufficient to develop a principled algorithm to estimate the threshold.[1] Thus, each adaptive procedure yields its own set of additional hypotheses on Ψ, which unfortunately are not always clearly stated and cannot be verified. In particular, the shape and properties of Ψ associated with a particular task are never directly accessible to the experimenter and can therefore only be estimated (Zchaluk & Foster, 2009). Moreover, even after the experiment has been performed and data have been collected, identifying the best parametric model for Ψ or evaluating the relevance of a given choice is deeply challenging (Strasburger, 2001). This hinders the reliability of Bayesian methods because adaptive methods that require prior knowledge on the shape of the psychometric function tend to produce significantly worse estimations when the assumptions they rely on are false (García-Pérez & Alcalá-Quintana, 2007; Hatzfeld, Hoang, & Kupnik, 2016). Therefore, the choice of the adaptive procedure is of paramount importance to accurately evaluate a behavioral or perceptive task. However, to the best of our knowledge, there is no clear consensus on which method to use for any given experiment, in particular when little is known about the psychometric function.

While it would be impossible to pinpoint the best procedure for each possible behavioral or perceptive task, in this work, we aimed at studying multiple settings to develop general guidelines. We focused on the threshold estimation problem for one-dimensional psychometric functions, and we paid particular attention to the link between the strength of the hypotheses underpinning adaptive procedures and their performance, both when their assumptions are met and when they are not. We looked at five adaptive procedures that we regrouped in two general classes: two model-based methods (i.e., methods that yield strong assumptions on the global shape of Ψ), namely, the PsiMethod (Kontsevich & Tyler, 1999) and the QuestPlus (Watson, 2017), and three model-free methods (where assumptions are made on local properties of Ψ), namely, the Staircase (Levitt, 1971), the Gaussian Processes (GP; Song et al., 2017), and the Dichotomous Optimistic Search (DOS;

[1] For a more in-depth discussion of the minimal set of mathematical assumptions necessary to solve this problem, we refer readers to the global black box optimization literature, and in particular to Grill, Valko, and Munos (2015).

Audiffren, 2021a, 2021b). These procedures were compared using multiple simulations and a wide range of psychometric functions, ranging from commonly studied Ψ, such as Weibull c.d.f. (García-Pérez & Alcalá-Quintana, 2007), to unusual Ψ functions, such as beta c.d.f. For every psychometric function, multiple different hyperparameters were tested, reflecting both steep and flat slopes, in two different settings: the yes/no and the 2-AFC (alternative forced choice) framework (see Wichmann & Hill, 2001b). Based on the results of these experiments, we discuss possible guidelines regarding the choice of the most appropriate adaptive procedure depending on the setting of the task at hand.

It should be noted that previous work has provided valuable reviews on the estimation of psychometric functions, such as Wichmann and Hill (2001a), which discusses the estimation of the entire psychometric function, or Klein (2001), which proposes a global approach to the problem. However, compared to existing studies, our work follows a different objective: to determine how well one can estimate the threshold given the data of an adaptive procedure (i.e., a generally significantly lower number of data points that are dependent on the observer's answer). This is different from their objective, which was to estimate the threshold, slope, and goodness of fit of different nonadaptive procedures, with significantly larger stimuli budgets. Moreover, this work provides a thorough comparison of new, highly efficient adaptive procedures that have been introduced in the past few years, including, notably, QuestPlus (a powerful generalization of Quest and GP and DOS, two robust nonparametric methods) (Watson, 2017; Song et al., 2017; Audiffren, 2021a). The methods are compared in multiple settings, using a significantly wider range of possible psychometric functions, in particular unusual psychometric functions that differ significantly from logistic/gaussian c.d.f. Furthermore, our experiments analyze the degradation of performance of the adaptive procedures when their assumptions regarding Ψ are no longer correct (i.e., the model of Ψ), which allows us to pinpoint the respective benefits and drawbacks of both model-free and model-based procedures. Finally, we report an additional metrics in our experiments, the regret, which provides additional information regarding the weaknesses of the methods' results (see previous work on the estimation of quantiles, such as Chaudhuri & Kalyanakrishnan, 2017).

2 Method

2.1 Problem Setting. We start by introducing a formal setting of the threshold estimation problem, which was used to compare the different adaptive procedures. Note that this setting is built on the one introduced by Audiffren (2021b).

2.1.1 Notation. In the following, we use the notation from Audiffren (2021a). Let T denote the time horizon (i.e., the maximum number of stimuli

presented to the observer during one experiment), $\mathbb{I} \subset \mathbb{R}$ the (closed) interval of possible stimuli, $\Psi : \mathbb{I} \mapsto [0, 1]$ the (unknown) psychometric function, $\mu_* \in [0, 1]$ the target probability, and $s_* \doteq \Psi^{-1}(\mu_*)$ the threshold. Finally, let $\gamma = \inf_{s\in\mathbb{I}} \Psi(s)$ denote the guess rate—the chance of the observer to make a correct guess independent of the stimulus—and $\lambda = 1 - \sup_{s\in\mathbb{I}} \Psi(s)$ the lapse rate. Note that these notations are identical to Wichmann and Hill (2001a), except for s_* and μ_*, which had no equivalent.

2.1.2 Psychophysics Experiment. In the threshold estimation problem, the objective of a psychophysics experiment is to find an estimator $\hat{s}$ of the threshold s_* with at most T stimuli. $\mathbb{I}$, T, and μ_* are known to the experimenter, but Ψ and s_* are not. The process unfolds as follows. For each round $t \in [1, \dots, T]$, the experimenter chooses an intensity $s \in \mathbb{I}$, and the observer detects it with probability $\Psi(s)$—more precisely, the detection process follows a Bernoulli law of mean $\Psi(s)$, and all samplings are assumed independent. The observer then communicates the result to the experimenter. At the end of the experiment ($t = T$), the experimenter computes the best guess for the target stimulus s_*, noted $\hat{s}$. Importantly, this setting can model most psychometric experiments by choosing the correct values of μ_*, γ, λ, T, or Ψ. For instance, the Yes/No setting can be obtained by setting $\mu^* = 0.5$ and $\gamma = 0$ (Wichmann & Hill, 2001a), while the N-AFC setting results from choosing $\gamma = 1/N$ and $\mu^* = 0.707$ (Lengyel & Fiser, 2019). In the following, we assume without any loss of generality that $\mathbb{I} = [0, 1]$. This is because any stimuli interval can easily be mapped to $[0, 1]$ using either linear or logarithmic invertible transformations. We also assume that the target probability is strictly reachable, that is, $\gamma < \mu_* < 1 - \lambda$, which is always the case in well-posed psychometric experiments. Finally, and due to the nature of the task (i.e., detecting stimuli of various intensity), the psychometric function is assumed to be continuous and strictly increasing (see Leek, 2001).

2.1.3 Evaluation. Multiple metrics have been proposed to assess the estimator $\hat{s}$ produced by a psychometric procedure, the most common being arguably the Accuracy and the Sweat Factor (Hatzfeld et al., 2016). These two metrics evaluate the distribution of $\hat{s}$ with respect to s_* over multiple simulation runs. Formally, let f denote a psychometric procedure and s_* be the solution to the threshold estimation problem with Ψ and μ_*. Given $\hat{s}_1^f, \dots, \hat{s}_N^f$ N estimations produced by an adaptive method f over N simulations, the Accuracy of an estimator is defined as

$$\text{Accuracy}(f) = \frac{1}{N} \sum_{n=1}^{N} (\hat{s}_n^f - s_*). \tag{2.1}$$

Note that equation 2.1 is also called bias and that it measures the difference between the mean of $\hat{s}^f$ over multiple experiments and the real threshold s_*,

representing the empirical evaluation of the bias of the estimator $\hat{s}^f$. Note that since s_* is not known during a real experiment, the bias can only be computed using simulations. The other metric is the Sweat Factor, which is defined as

$$\text{Sweat Factor}(f) = \sqrt{\frac{1}{N-1}\left[\sum_{n=1}^{N}\left(\hat{s}_n^f\right)^2 - \left(\sum_{n=1}^{N}\hat{s}_n^f\right)^2\right]}. \tag{2.2}$$

The Sweat Factor is used to evaluate the precision of the estimator, that is, the dispersion of $\hat{s}^f$ over N runs. This metric can also be seen as the standard deviation of the distance between $\hat{s}^f$ and s_*. More recently, a new metric has been introduced to evaluate the distribution of $\Psi(\hat{s}^f)$ with respect to μ_* (Audiffren, 2021b):

$$\text{Regret}(f) = \frac{1}{N}\sum_{n=1}^{N}|\mu_* - \Psi(\hat{s}_n^f)|. \tag{2.3}$$

Equation 2.3 measures the mean (over multiple experiments) of the distance between μ_* and $\Psi(\hat{s}^f)$. While there is a significant relation between Regret and Accuracy, it is important to note that Regret encodes the distance between probabilities rather than the distance between stimuli. In other words, while Accuracy represents the distance between the predicted stimulus intensity and the target landmark, Regret quantifies how representative this estimator is. It is interesting to note that this metric takes the shape of Ψ into account and thus automatically adapts to the difficulty of the threshold estimation problem: a small Regret indicates an estimated stimulus whose detection probability $\Psi(\hat{s})$ is very close to the desired value μ_*. Similar to Accuracy, Regret can only be computed using simulations, as it requires knowing Ψ.

2.2 Adaptive Procedures. For our simulations, we studied five adaptive procedures that we regrouped in two broad classes: model-based methods, which are Bayesian methods with strong assumptions on the global shape of Ψ, and model-free methods, which only rely on assumptions regarding the local properties of Ψ.

2.2.1 Model-Based Procedures. We considered two Bayesian procedures, the PsiMethod and the QuestPlus.

The PsiMethod (Kontsevich & Tyler, 1999) aims at estimating the parameters of Ψ (i.e., location and slope) by choosing at each step the stimulus that maximizes the expected reduction of uncertainty in the posterior distribution of the parameters. We used the parameters recommended by

Kontsevich and Tyler (1999), a modified version of the implementation provided by Peirce et al. (2019), with gaussian c.d.f. as a prior.

The QuestPlus (Watson, 2017) is a generalization of the Quest method (Watson & Pelli, 1983), which estimates the location and slope parameters, as well as the lapse rate λ. Similar to the PsiMethod, the QuestPlus relies on the minimum entropy principle to choose the adaptive sequence of stimuli to present to the observer and on the maximum likelihood principle to estimate the final values of the parameters. We used the parameters recommended by Watson (2017), and a modified version of the implementation provided by Peirce et al. (2019), with Weibull c.d.f. as the shape assumption for Ψ.

Note that both methods use a multidimensional grid of possible values for their parameters (location, slope, lapse), as well as the stimuli values. While high-resolution grids might provide better approximations of the Bayesian method and thus better results, they also lead to a significant increase in the computational complexity of the method. A careful consideration of the trade-off is generally important for the optimal use of these methods.

2.2.2 Model-Free Procedures. We also considered three model-free methods: two non-Bayesian procedures (Staircase and DOS) and Gaussian Processes (GP).

The GP method has been proposed by Gardner, Malkomes et al. (2015), Gardner, Song, Weinberger, Barbour, and Cunningham (2015), and Song et al. (2017). It is inspired by a classical tool for nonlinear regression (MacKay, 1998). While intrinsically a Bayesian method, GP's prior is significantly weaker than the previously mentioned Bayesian methods, as Ψ is only assumed to belong to a collection of reproducing kernel Hilbert spaces, a very rich functional space (see Kadri et al., 2016). GP methods have been shown to significantly outperform other methods for some applications such as audiograms (see Gardner, Malkomes et al., 2015). The use of GP procedures requires the choice of a kernel family (such as linear kernels; Song et al., 2017), a family of mean functions, a regularization (also called noise variance), and two grids of possible parameters (one for the kernel and one for the mean functions). GP methods tend to perform best when these functions are handcrafted for the problem at hand (Gardner, Malkomes et al., 2015), and their performance may worsen when the kernel is ill suited (see Schulz, Speekenbrink, & Krause, 2018). It is important to note that we chose to class GP as a model-free family since the choice of the kernel function influences the local properties of the Ψ function rather than its global behavior. We used GP with the parameters used in Song et al. (2017).

The Staircase (and its iterations) (Cornsweet, 1962; Levitt, 1971; Wichmann & Jäkel, 2018) is arguably the most commonly used adaptive method. This procedure can be seen as performing an asymmetric random walk

on the space of stimuli. The main parameters of the Staircase are the UP/DOWN ratio, which conditions the target probability toward which the Staircase will converge (e.g., $\mu_* = 0.5$ for UP = DOWN = 1 and $\mu_* = 0.707$ for UP = 1 and DOWN = 2 (see Brown, 1996), and the step sizes. Levitt (1971) have shown that if Ψ is the c.d.f. of a gaussian distribution, STAIR has a guaranteed convergence for the appropriate choice of step sizes. We used the Staircase with the parameters recommended by Lengyel and Fiser (2019).

Finally, the DOS (Dichotomous Optimistic Search; Audiffren, 2021a) is a more recent procedure inspired by hierarchical bandits and black box optimization. Compared to the other procedures we have noted, the DOS is model and parameter free. Specifically, the DOS does not require any knowledge regarding the shape of the psychometric function. It only assumes that Ψ is mildly smooth around s_* and has no parameter besides μ_* and T. Finally, it should be noted that DOS provides strong guarantees regarding its convergence speed (Audiffren, 2021a).

2.3 Psychometric Functions. Four different shapes of psychometric functions were studied in our simulations: two commonly used functions—logistic and Weibull c.d.f.—and two functions seldomly used in psychophyics—Beta c.d.f. and a Hölder-continuous function (Hölder for short). While both the logistic and Weibull c.d.f. functions have been specifically studied (see Hatzfeld et al., 2016; García-Pérez & Alcalá-Quintana, 2007), the Beta c.d.f. and Hölder functions were chosen because recent work (e.g., Zchaluk & Foster, 2009) has suggested that commonly used c.d.f. may not be able to capture the psychometric functions for all psychophysical tasks. Therefore, we chose to use Beta c.d.f. and Hölder Ψ to assess the robustness of the different adaptive procedures when the psychometric functions have an unusual shape. Importantly, each function satisfies a minimal set of assumptions: Ψ is nondecreasing (the stronger the stimulus is, the easier it is to perceive) and continuous (Leek, 2001). Every psychometric function listed below was tested with multiple sets of parameters (slope, location), reflecting both steep and flat slopes, similar to the experimental design of Hatzfeld et al. (2016). Moreover, two different settings were used: the Yes/No ($\gamma = 0$, $\lambda = 0.02$, $\mu = 0.5$) and the 2-AFC ($\gamma = 0.5$, $\lambda = 0.02$, $\mu = 0.707$) framework. Figure 1 shows the behavior of each steep function in the Yes/No setting.

2.3.1 Logistic Function. See Hatzfeld et al. (2016). We used $\alpha = 0.4, \beta = 10$ for the flat setting and $\alpha = 0.7, \beta = 24$ for the steep setting:

$$\Psi(s|\alpha, \beta) = \gamma + (1 - \gamma - \lambda)\frac{1}{1 + e^{-\beta(s-\alpha)}}.$$

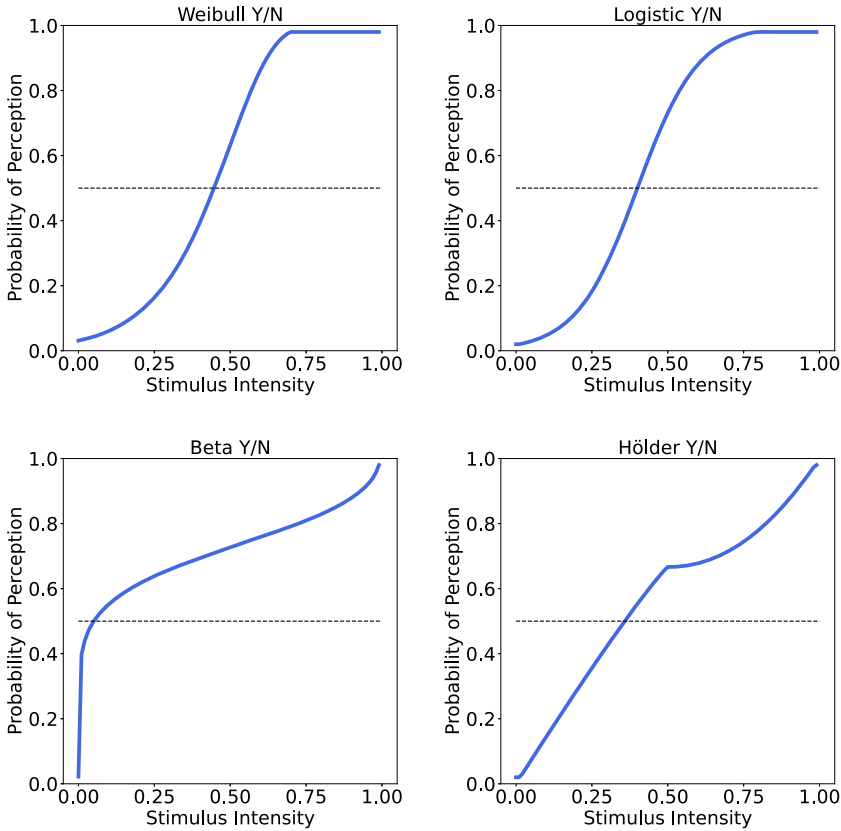

Figure 1: Illustration of the four psychometric functions that we studied, for the steep set of parameters in the Yes/No setting. The dashed black line represents the 50% perception level. Top left: Weibull c.d.f. Top right: Logistic c.d.f. Bottom left: Beta c.d.f. Bottom right: Hölder function. Note that the 0.5 threshold is associated with different stimuli for each function. In particular the threshold of the beta function is ≈ 0.09 and is therefore very close to the lower end of the stimuli intensity interval.

2.3.2 Weibull. See García-Pérez & Alcalá-Quintana (2007). We used $\alpha = 0.5$, $\beta = 3$ for the flat setting and $\alpha = 0.7$, $\beta = 9$ for the steep setting:

$$\Psi(s|\alpha, \beta) = 1 - \lambda - (1 - \gamma - \lambda) \exp\left(-10^{\beta(s-\alpha)}\right).$$

2.3.3 Beta. This is another type of Ψ that follows a c.d.f. model. While Beta's behavior is similar to Logistic and Weibull in specific landmarks, its

general behavior is different. We used $\alpha = 0.14$, $\beta = 0.27$ for the flat setting and $\alpha = 0.40$, $\beta = 0.74$ for the steep setting:

$$\Psi(s|\alpha, \beta) = \gamma + (1 - \gamma - \lambda)\frac{\int_0^s x^{\alpha-1}(1-x)^{\beta-1}dx}{\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx}.$$

2.3.4 Hölder. This function satisfies the basic hypotheses on Ψ, but contrary to previous Ψ, it is not a classical c.d.f. and therefore grants the possibility of testing the robustness of the different adaptive procedures even further. We used $\alpha = 1.1$, $\beta = 2.1$, $s' = 0.5$, and $m' = 0.25$ for the flat setting and $\alpha = 0.7$, $\beta = 2$, $s' = 0.4$, $m' = 0.6$ for the steep setting:

$$\Psi(s|\alpha, \beta, s', m') = \max\left(\min\left(m' + 1_{s>s'}|s - s'|^{\alpha} - 1_{s<s'}|s - s'|^{\beta}, 1 - \lambda\right), \gamma\right).$$

2.4 Simulation Parameters. In our simulations, for each of the psychometric functions described in section 2.3, the settings used for the 2-AFC and Yes/No frameworks are described above in section 2.1. All adaptive procedures detailed in section 2.2 were evaluated with three different stimuli budgets: 50, 100, and 200. This allowed us to reproduce different experimental constraints. Importantly, all chosen values were relatively small in order to limit the influence of fatigue and of learning effects that could interfere with the experiment (Wichmann & Hill, 2001a).[2] Each simulation was run 100 times. When reported, *p*-values were obtained using the Mann-Whitney U-test, and the 95% confidence intervals were obtained using the 1.96 standard deviation half width.

3 Results

3.1 Detailed Comparison. We begin by comparing the performance of each adaptive procedure for each setting. Figures 2 and 3 (resp. 4 and 5, 6 and 7, and 8 and 9) show the distribution of Accuracy ($\hat{s} - s_*$) and Regret ($|\Psi(\hat{s}) - \Psi(s_*)|$) over 100 runs for the Weibull (resp. Logistic, Beta, and Hölder) psychometric functions for both frameworks (Yes/No, 2-AFC) with the three stimuli budgets T = 50, 100, and 200. An interactive tool to help the visualization of these results can be found at comparall.herokuapp.com.

3.2 Yes/No Experiment.

3.2.1 Weibull Function. DOS, GP, and Staircase achieved comparable Accuracy and Sweat Factor performance for all values of T (see Figure 2).

[2] Note that Audiffren (2021b) has performed an analysis for larger values of T ($T > 1000$) to assess the empirical convergence rate of different methods. This study has shown that the DOS performs significantly better than its counterparts.

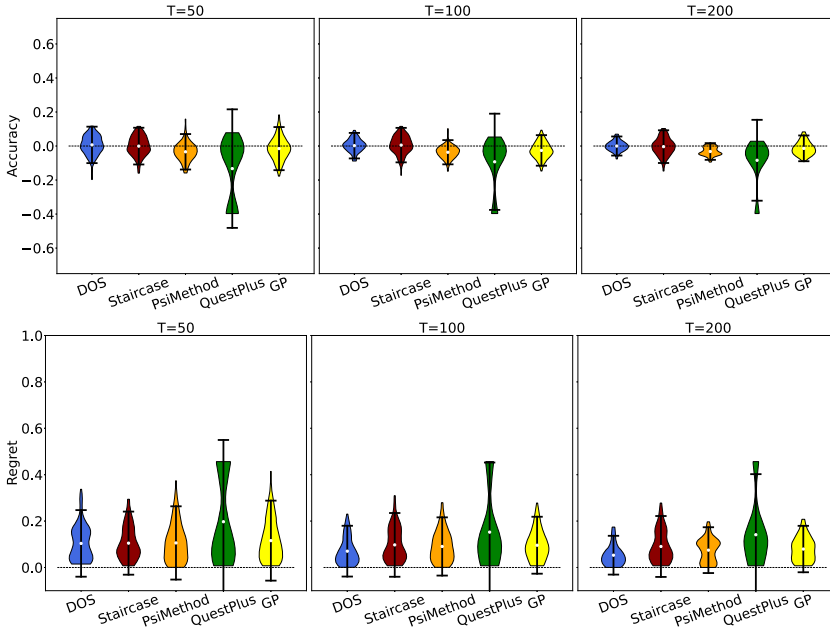

Figure 2: Distribution of Accuracy and Regret over 100 runs in the Weibull Yes/No setting. White dots (resp. whiskers) represent the average values (resp. 1.96 standard deviation). The ideal value for Accuracy and Regret is 0, represented by a dashed line.

The Accuracy of the QuestPlus and PsiMethod was slightly worse ($p < 0.01$), while the Sweat Factor of the QuestPlus was significantly worse ($p < 0.01$). Interestingly, the distribution of Accuracy for QuestPlus was bimodal, and each mode corresponded to a set of parameters (flat and steep). The Accuracy of QuestPlus was comparable to that of other methods for the steep function and mildly worse for the flat function ($p < 0.01$). While most observations still hold for the Regret distribution, the Regret of the PsiMethod was comparable to that of model-free methods, while QuestPlus had a significantly higher Regret than all other procedures ($p < 0.01$). This illustrates that QuestPlus has a more complex grid of hyperparameter to optimize, which translates into a slower convergence rate.

3.2.2 Logistic Function. Interestingly, all three model-free methods (DOS, Staircase, and GP) achieved comparable Accuracy and Sweat Factor performance for $T = 50$ and $T = 100$, while the Staircase was slightly worse for $T = 200$ ($p < 0.01$) (see Figure 3). The QuestPlus procedure had a significantly worse Sweat Factor in all cases ($p < 0.01$) but similar Accuracy for $T = 100$ and $T = 200$ ($p < 0.01$). The Accuracy and Sweat Factor of the

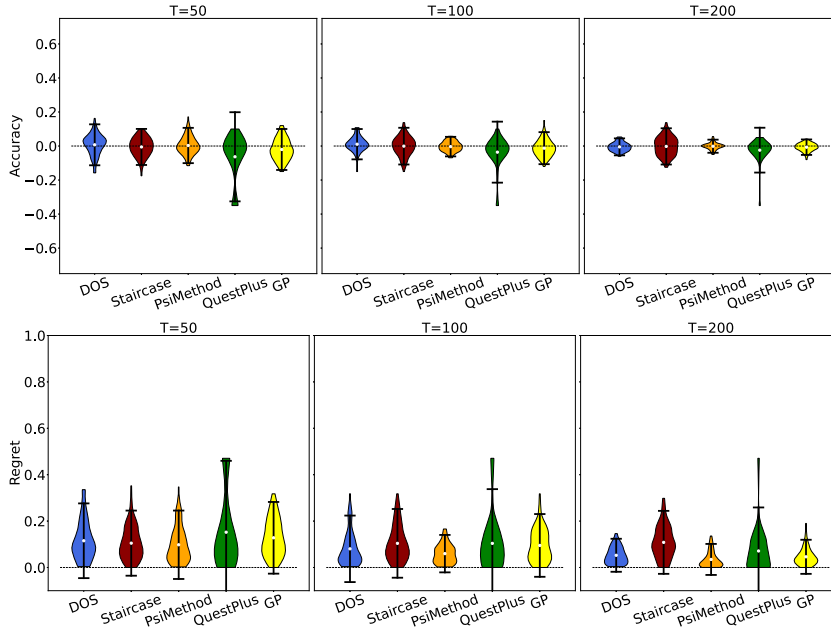

Figure 3: Distribution of Accuracy and Regret over 100 runs in the Logistic Yes/No setting. White dots (resp. whiskers) represent the average values (resp. 1.96 standard deviation). The ideal value for Accuracy and Regret is 0, represented by a dashed line.

PsiMethod were comparable to those of DOS and GP for all values of T. These differences are even more apparent in the Regret distributions: while DOS, PsiMethod, and GP had comparable Regret, QuestPlus and Staircase were significantly worse for $T = 100$ and $T = 200$ ($p < 0.01$). Overall, the poor performance of QuestPlus can be explained by the fact that its Weibull prior is not accurate in these simulations (Logistic functions), as well as by its slower convergence rate. Interestingly, while the PsiMethod also has a Weibull prior, it achieved one of the best results for this setting.

3.2.3 Beta Function. All model-free methods (DOS, GP, and Staircase), as well as QuestPlus, achieved comparable Accuracy and Sweat Factor performance for all stimuli budgets ($T = 50$, 100, and 200), while the PsiMethod showed significantly worse Accuracy ($p < 0.01$) (see Figure 4). The pattern of performance was similar regarding the Regret metric. Specifically, DOS had a better Regret for $T = 50$ ($p < 0.01$) and a similar Regret for $T = 100$ and $T = 200$, while the PsiMethod had a worse Regret with the same settings ($p < 0.01$).

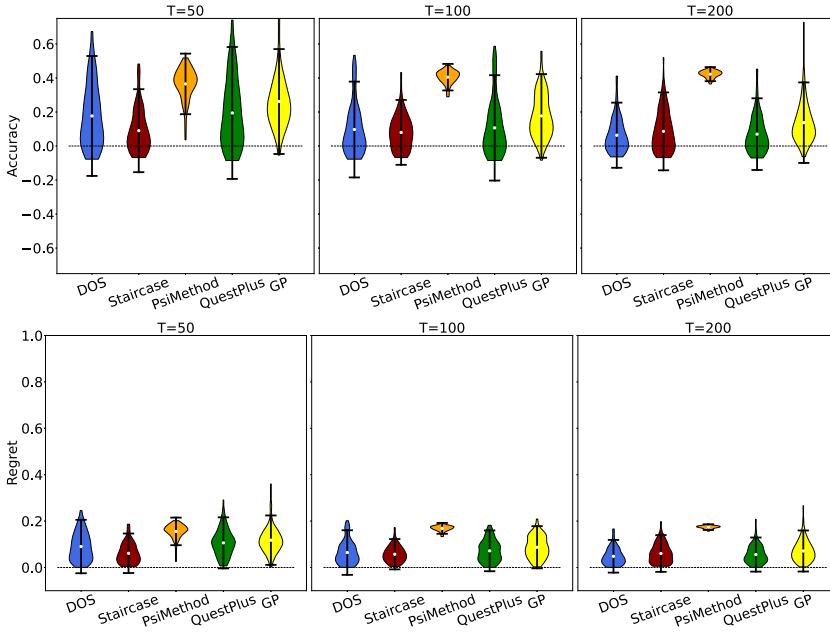

Figure 4: Distribution of Accuracy and Regret over 100 runs in the Beta Yes/No setting. White dots (resp. whiskers) represent the average values (resp. 1.96 standard deviation). The ideal value for Accuracy and Regret is 0, represented by a dashed line.

3.2.4 Hölder Function. Similar to the Beta function, all methods but the PsiMethod achieved comparable Accuracy for $T = 50, 100$, and 200 (see Figure 5). The Accuracy and Regret of the PsiMethod were significantly worse for $T = 100$ and $T = 200$ ($p < 0.01$). In all cases, QuestPlus had a significantly worse Sweat Factor ($p < 0.01$) and a significantly worse Regret for $T = 50$ ($p < 0.01$).

3.3 2-AFC Experiment.

3.3.1 Weibull Function. In line with the results obtained in the Yes/No setting, DOS and Staircase achieved comparable Accuracy and Regret (see Figure 6). Interestingly, they achieved better Regret than in the Yes/No experiments ($p < 0.01$) despite a mildly worse Accuracy. This is due to the fact that (1) Ψ is flatter around $\mu_* = 0.707$ compared to $\mu_* = 0.5$, and (2) μ_* is closer to 1. Thus, overestimating s_* leads to a lower Regret. As opposed to that, GP achieved worse Accuracy and Regret ($p < 0.01$). Moreover, the Accuracy of the QuestPlus method was significantly better than in the Yes/No setting and better than that of the other methods ($p < 0.01$).

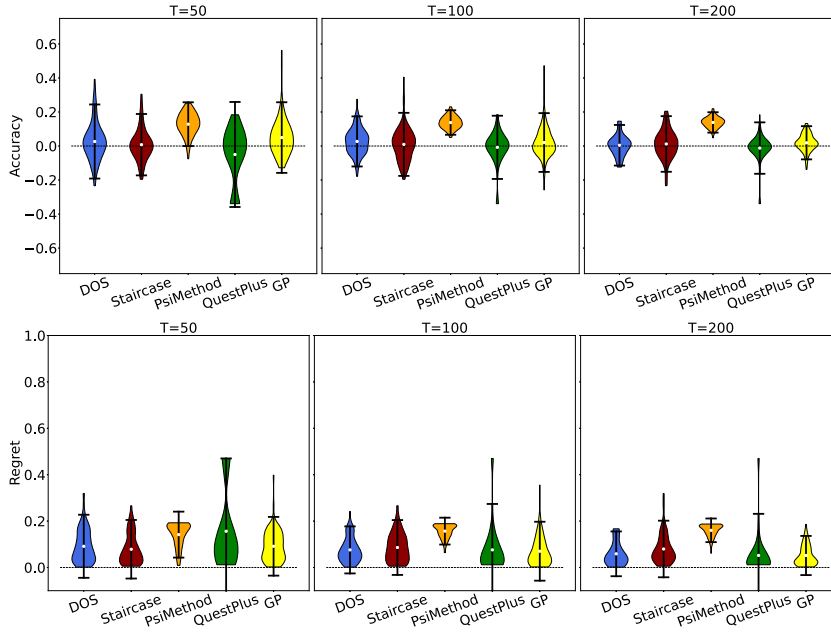

Figure 5: Distribution of Accuracy and Regret over 100 runs in the Hölder Yes/No setting. White dots (resp. whiskers) represent the average values (resp. 1.96 standard deviation). The ideal value for Accuracy and Regret is 0, represented by a dashed line.

3.3.2 Logistic Function. The pattern of results was very similar to that observed for the Weibull function. In particular, QuestPlus achieved better Accuracy than the other methods, while having similar Regret (see Figure 7). Regarding the model-free methods, DOS had a better Accuracy than its counterparts while having a similar Regret. GP had significantly worse Accuracy and Sweat Factor for all values of T ($p < 0.01$), but no statistically significant difference was found for its Regret. The PsiMethod achieved worse Accuracy and Regret than DOS and QuestPlus for $T = 50$ and $T = 100$, and comparable performance for $T = 200$. Finally, and similar to the Weibull setting, all methods achieved a lower Regret compared to the Yes/No setting. This highlights the fact that Ψ is significantly flatter in the 2-AFC setting than in the Yes/ No setting, as $\gamma = 0.5$.

3.3.3 Beta Function. Here, DOS, Staircase, and PsiMethod achieved comparable Accuracy and Regret, while QuestPlus and GP achieved worse Accuracy for T = 50 and T = 100 ($p < 0.01$). Regret was comparable among all methods for $T = 200$ (see Figure 8).

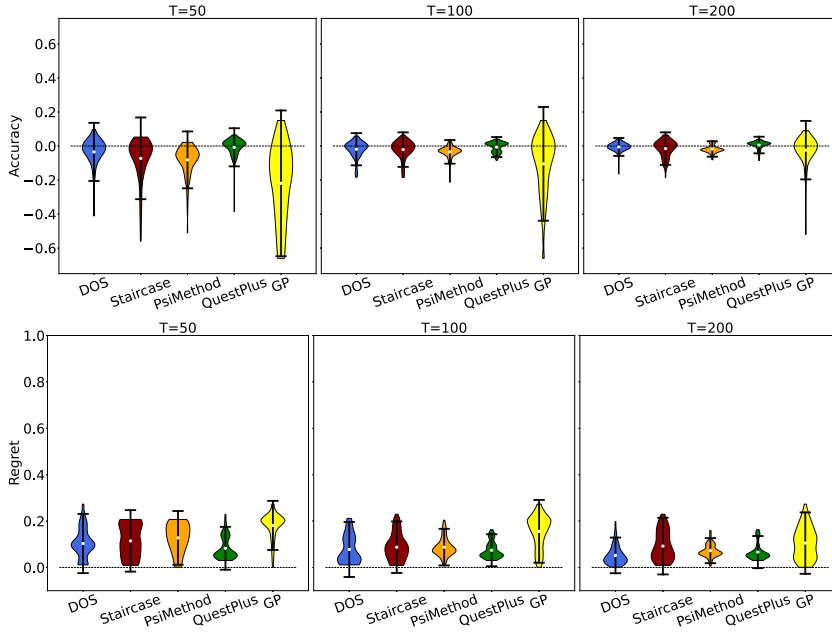

Figure 6: Distribution of Accuracy and Regret over 100 runs in the Weibull 2-AFC setting. White dots (resp. whiskers) represent the average values (resp. 1.96 standard deviation). The ideal value for Accuracy and Regret is 0, represented by a dashed line.

3.3.4 Hölder Function. In this 2-AFC experiment, all three model-free methods achieved comparable Accuracy and Regret for all values of T (see Figure 9). PsiMethod and QuestPlus had worse Accuracy and Regret for $T = 50$ and $T = 100$ but similar performance compared to their model-free counterparts for $T = 200$.

3.4 Global Analysis. The previous results show that DOS and Staircase generally achieved comparable results on all tested settings regarding Accuracy and Regret, while GP frequently had slightly worse results than its model-free counterparts. This may been explained by the fact that GP aims at estimating the entire Ψ function instead of the threshold. It therefore requires more stimuli and converges more slowly. As opposed to that, the performance of the PsiMethod varied significantly between settings. Specifically, while the PsiMethod achieved good results, for example, the Logistic Yes/No setting, its performance drastically worsened in the Beta and Hölder Yes/No framework (when its Bayesian model is widely incorrect). Interestingly, QuestPlus exhibited different behavior: its performance was significantly less variable than that of the PsiMethod while being

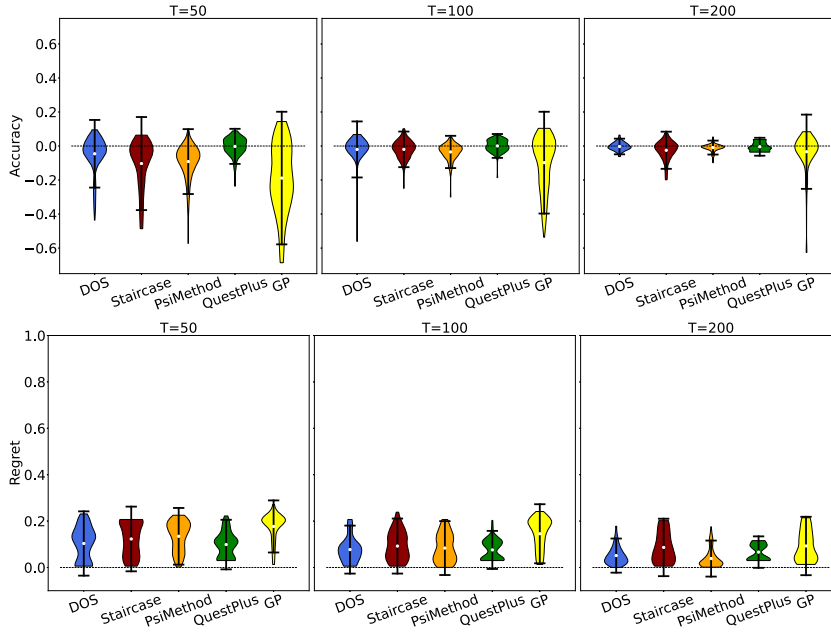

Figure 7: Distribution of Accuracy and Regret over 100 runs in the Logistic 2-AFC setting. White dots (resp. whiskers) represents the average values (resp. 1.96 standard deviation). The ideal values for Accuracy and Regret is 0, represented by a dashed line.

generally slightly worse. The difference observed between the QuestPlus and the PsiMethod might be explained by the fact that QuestPlus has significantly more parameters to estimate than the PsiMethod. It therefore requires more observations before reaching an accurate estimation of the threshold. But this also provides more robustness to QuestPlus than to the PsiMethod, and it allows it to perform significantly better in a wider range of setting.

Table 1 displays the average Accuracy (unsigned), Regret, and Sweat Factor, aggregated over all combinations of parameters, experimental settings, and simulation runs. This global analysis confirms the trend observed with each psychometric function. In particular, model-free methods had better average Accuracy and Regret than model-based methods, as model-based procedures performed poorly in some settings. This was especially true for the PsiMethod, whose average Accuracy was particularly poor in the Beta and Hölder Yes/No setting. Note that while the PsiMethod had the best Sweat Factor in the Yes/No setting, its Accuracy and Regret were significantly worse than the model-free methods. Overall, DOS and Staircase methods had comparable values of absolute Accuracy, with a slight

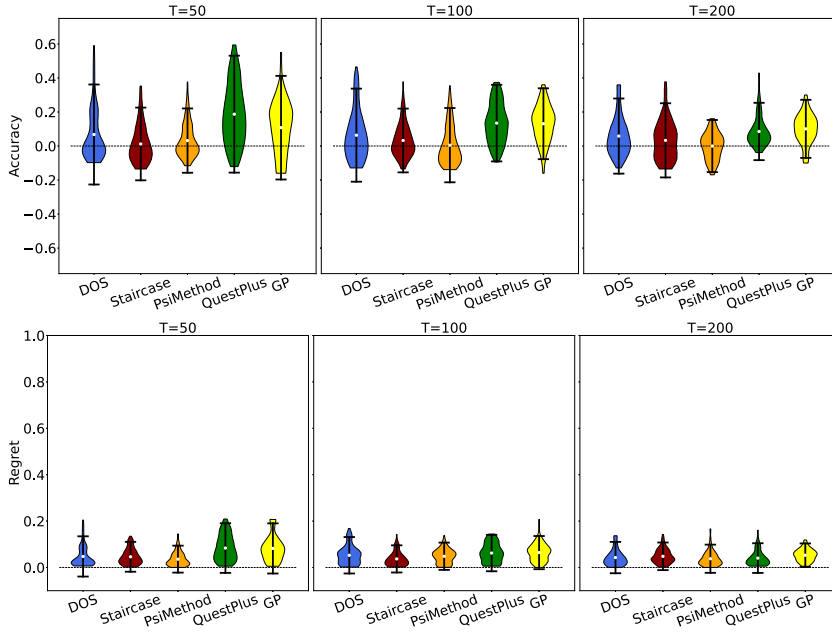

Figure 8: Distribution of Accuracy and Regret over 100 runs in the Beta 2-AFC setting. White dots (resp. whiskers) represent the average values (resp. 1.96 standard deviation). The ideal value for Accuracy and Regret is 0, represented by a dashed line.

advantage for DOS, while DOS strongly outperformed its competitors with respect to the Regret metric in almost all cases (the only exception being Yes/No $T = 50$). Finally, GP performance was slightly worse than its model-free counterparts, which can be explained by the fact that contrary to what has been done in previous work with very specific settings such as for audiograms (Gardner, Malkomes et al., 2015), GP kernel and parameterization were not finely tuned to Ψ. This was done to illustrate the most general case in which the experimenter lacks prior knowledge regarding the Ψ function.

4 Discussion

4.1 Methods Robustness. In our experiments, the model-free procedures (DOS, Staircase, and GP) achieved consistent performance across all tested psychometric functions and settings. The results obtained with the QuestPlus and PsiMethod were more variable, particularly in unusual psychometric functions such as the Beta c.d.f. This lower performance is in line with the results of previous studies that the use of an incorrect prior

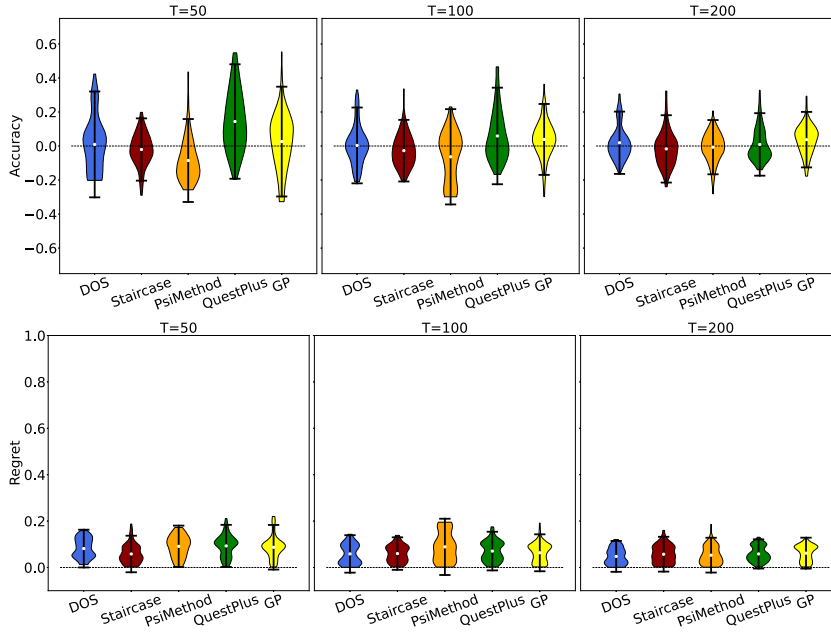

Figure 9: Distribution of Accuracy and Regret over 100 runs in the Hölder 2-AFC setting. White dots (resp. whiskers) represent the average values (resp. 1.96 standard deviation). The ideal value for Accuracy and Regret is 0, represented by a dashed line.

may lead to significantly worse estimates (García-Pérez & Alcalá-Quintana, 2007; Hatzfeld et al., 2016). However, when their prior was correct, the Bayesian methods tended to perform on par with model-free methods. Because model-free methods do not rely on any prior regarding the shape of Ψ, on average, they performed better across multiple functions.

4.2 Method Parameters. The QuestPlus, PsiMethods, and GPmethods, require setting multiple parameters that may have a significant impact on their performance. For instance, it may be argued that a different sampling strategy among the sweet points of each method may lead to better results (Shen & Richards, 2012) or that a better-tuned grid of parameters may improve the performance of the procedures (Song et al., 2017). However, there is no clear consensus on which one may be optimal, as it has been shown to depend on the experimental settings and objectives, as well as on the underlying psychometric function (Shen & Richards, 2012; Strasburger, 2001). In our simulations, we attempted to replicate an arbitrary psychophysical task, for which the underlying Ψ function might not have been well studied and the optimal set of parameters is unknown—a common occurrence in psychophysics research (see Schütz, Braun, Kerzel, & Gegenfurtner, 2008).

Table 1: Average Absolute Accuracy, Regret and Sweat Factor over All Simulations, for All Settings, for the Yes/No (Top) and 2-AFC (Bottom) Experiments.

Setting	T	Avg. Metric	DOS	Staircase	PsiMethod	QuestPlus	GP
Yes/No	50	Abs. Accuracy	0.034	**0.026**	0.118	0.078	0.087
		Sweat Factor	0.102	0.081	**0.080**	0.150	0.097
		Regret	0.95	**0.087**	0.152	0.158	0.113
	100	Abs. Accuracy	**0.024**	**0.024**	0.110	0.051	0.059
		Sweat Factor	0.075	0.075	**0.042**	0.091	0.077
		Regret	**0.073**	0.086	0.149	0.106	0.088
	200	Abs. Accuracy	**0.018**	0.026	0.119	0.047	0.044
		Sweat Factor	0.053	0.075	**0.029**	0.073	0.058
		Regret	**0.053**	0.085	0.141	0.087	0.065
2-AFC	50	Abs. Accuracy	**0.039**	0.052	0.083	0.075	0.135
		Sweat Factor	0.124	**0.116**	0.120	0.145	0.184
		Regret	**0.084**	0.085	0.109	0.106	0.132
	100	Abs. Accuracy	**0.025**	0.030	0.070	0.057	0.093
		Sweat Factor	0.096	**0.073**	0.107	0.104	0.133
		Regret	**0.066**	0.069	0.089	0.082	0.107
	200	Abs. Accuracy	**0.018**	0.022	0.068	0.055	0.049
		Sweat Factor	**0.064**	0.079	0.065	0.074	0.092
		Regret	**0.048**	0.070	0.058	0.066	0.078

Note: The best value for each metric and stimulus budget is highlighted in bold.

In this setting, the advantage of methods that required a small number of parameters (such as the Staircase) or no parameters at all (such as DOS) is nonnegligible.

4.3 Bayesian Priors. While methods that use Bayesian priors and a large number of parameters might yield some disadvantages in the settings studied in this work, they can be very effective under a different set of assumptions. Recently, Gardner, Malkomes et al. (2015) showed that with the correct model, GP can be used to diagnose noise-induced hearing loss (NIHL) with as little as 30 stimuli, a very small fraction of the hundreds of stimuli typically used in regular audiometric tests. Note, however, that this is notably because the NIHL setting has been extensively studied in the past, so that many properties of its psychometric functions are agreed on and can be used to improve the results. This, unfortunately, is not the case for all psychophysics settings.

4.4 Dos and Staircase. The two procedures that displayed the best global performance in our simulations are the Staircase (Levitt, 1971) and DOS (Audiffren, 2021b). The Staircase is one of the most commonly used adaptive procedures in psychophysics, and the good accuracy observed with our simulations is in line with the results reported in previous studies (Hatzfeld et al., 2016). However, the Staircase requires choosing an

appropriate sequence of step sizes to perform optimally, and it is limited to a small number of target probabilities, such as $\mu_* = 0.5$ or 0.707 (Brown, 1996). DOS is completely parameter free, it can converge toward any arbitrary target μ_*, and it is significantly better than the Staircase in the 2-AFC setting, when $\mu_* \neq 0.5$. Moreover, Audiffren (2021b) has proved strong theoretical guarantees for DOS, including convergence speed. Importantly, these guarantees hold under minimal local assumptions on Ψ, whereas to the best of our knowledge, no similar guarantees exist for other procedures. Overall, if no strong prior knowledge about Ψ is available, the Staircase and DOS appear to be the procedures of choice to best estimate the threshold in psychophysical tasks.

4.5 Psychometric Fields. Note that all previous conclusions only apply to threshold estimation. For other applications, the DOS and Staircase do not supersede model-based procedures such as QuestPlus. This is notably because Bayesian methods are designed to estimate the entire Ψ function, whereas the DOS and Staircase procedures are targeted toward estimating s_*. The use of the Staircase to estimate the entire psychometric function suffers from biases and limitations (Kaernbach, 2001), while the DOS has not been studied in that context and for that purpose. Similarly, Bayesian methods can be used on multidimensional psychometric functions (called psychometric fields) to estimate the level sets of the functions. This is, for instance, the case with audio signals, when the stimulus can be characterized using volume and frequency. Model-free methods such as the DOS and Staircase cannot easily be extended to this type of setting.

References

Audiffren, J. (2021a). Dichotomous optimistic search to quantify human perception. In *Proceedings of the 38th International Conference on Machine Learning*. New York: ACM.

Audiffren, J. (2021b). Quantifying human perception with multi-armed bandits. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems*. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Benson, A. J., Hutt, E. C., & Brown, S. F. (1989). Thresholds for the perception of whole body angular movement about a vertical axis. *Aviation, Space, and Environmental Medicine*, *60*(3), 205–213. 2712798

Brown, L. G. (1996). Additional rules for the transformed up-down method in psychophysics. *Perception and Psychophysics*, *58*(6), 959–962. 10.3758/BF03205497, PubMed: 8768190

Chaudhuri, A. R., & Kalyanakrishnan, S. (2017). PAC identification of a bandit arm relative to a reward quantile. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. Palo Alto, CA: AAAI Press.

Cornsweet, T. N. (1962). The staircase-method in psychophysics. *American Journal of Psychology*, *75*(3), 485.

Dreschler, W. A., & Verschuure, J. (1996). *Psychophysical evaluation of fast compression systems*. Singapore: World Scientific.

García-Pérez, M. A., & Alcalá-Quintana, R. (2007). Bayesian adaptive estimation of arbitrary points on a psychometric function. *British Journal of Mathematical and Statistical Psychology*, *60*(1), 147–174.

Gardner, J., Malkomes, G., Garnett, R., Weinberger, K. Q., Barbour, D., & Cunningham, J. P. (2015). Bayesian active model selection with an application to automated audiometry. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems*, *28* (pp. 2386–2394). Red Hook, NY: Curran.

Gardner, J. R., Song, X., Weinberger, K. Q., Barbour, D. L., & Cunningham, J. P. (2015). Psychophysical detection testing with Bayesian active learning. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence* (pp. 286–295). San Mateo, CA: Morgan Kaufmann.

Grill, J.-B., Valko, M., & Munos, R. (2015). Black-box optimization of noisy functions with unknown smoothness. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems*, *28* (pp. 667–675). Red Hook, NY: Curran.

Hatzfeld, C., Hoang, V. Q., & Kupnik, M. (2016). It's all about the subject-options to improve psychometric procedure performance. In *Proceedings of International Conference on Human Haptic Sensing and Touch Enabled Computer Applications* (pp. 394–403). Berlin: Springer.

Kadri, H., Duflos, E., Preux, P., Canu, S., Rakotomamonjy, A., & Audiffren, J. (2016). Operator-valued kernels for learning from functional response data. *Journal of Machine Learning Research*, *17*(20), 1–54.

Kaernbach, C. (2001). Slope bias of psychometric functions derived from adaptive data. *Perception and Psychophysics*, *63*(8), 1389–1398. 10.3758/BF03194550, PubMed: 11800464

Klein, S. A. (2001). Measuring, estimating, and understanding the psychometric function: A commentary. *Perception and Psychophysics*, *63*(8), 1421–1455. 10.3758/BF03194552, PubMed: 11800466

Kontsevich, L. L., & Tyler, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*, *39*(16), 2729–2737. 10.1016/S0042-6989(98)00285-5, PubMed: 10492833

Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Perception and Psychophysics*, *63*(8), 1279–1292. 10.3758/BF03194543, PubMed: 11800457

Lengyel, G., & Fiser, J. (2019). The relationship between initial threshold, learning, and generalization in perceptual learning. *Journal of Vision*, *19*(4), 28. 10.1167/19.4.28, PubMed: 31022729

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America*, *49*(2B), 467–477. 10.1121/1.1912375

MacKay, D. J. (1998). Introduction to gaussian processes. *NATO ASI Series F Computer and Systems Sciences*, *168*, 133–166.

Nir, R.-R., Granovsky, Y., Yarnitsky, D., Sprecher, E., & Granot, M. (2011). A psychophysical study of endogenous analgesia: The role of the conditioning pain in the induction and magnitude of conditioned pain modulation. *European Journal of Pain*, *15*(5), 491–497. 10.1016/j.ejpain.2010.10.001, PubMed: 21035364

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., . . . Lindeløv, J. K., (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. 10.3758/s13428-018-01193-y, PubMed: 30734206

Scheuerman, J., Venable, K. B., Anderson, M. T., & Golob, E. J. (2017). Modeling spatial auditory attention: Handling equiprobable attended locations. In *International Workshop on Cognition and Artificial Intelligence for Human-Centred Design at the International Joint Conference on Artificial Intelligence* (pp. 1–7).

Schulz, E., Speekenbrink, M., & Krause, A. (2018). A tutorial on gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, *85*, 1–16. 10.1016/j.jmp.2018.03.001

Schütz, A. C., Braun, D. I., Kerzel, D., & Gegenfurtner, K. R. (2008). Improved visual sensitivity during smooth pursuit eye movements. *Nature Neuroscience*, *11*(10), 1211–1216.

Shen, Y., & Richards, V. M. (2012). A maximum-likelihood procedure for estimating psychometric functions: Thresholds, slopes, and lapses of attention. *Journal of the Acoustical Society of America*, *132*(2), 957–967. 10.1121/1.4733540

Song, X. D., Garnett, R., & Barbour, D. L. (2017). Psychometric function estimation by probabilistic classification. *Journal of the Acoustical Society of America*, *141*(4), 2513–2525. 10.1121/1.4979594

Strasburger, H. (2001). Converting between measures of slope of the psychometric function. *Perception and Psychophysics*, *63*(8), 1348–1355. 10.3758/BF03194547, PubMed: 11800461

Watson, A. B. (2017). Quest+: A general multidimensional Bayesian adaptive psychometric method. *Journal of Vision*, *17*(3), 10–10. 10.1167/17.3.10, PubMed: 28355623

Watson, A. B., & Pelli, D. G. (1983). Quest: A Bayesian adaptive psychometric method. *Perception and Psychophysics*, *33*(2), 113–120. 10.3758/BF03202828, PubMed: 6844102

Wichmann, F. A., & Hill, N. J. (2001a). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception and Psychophysics*, *63*(8), 1293–1313. 10.3758/BF03194544

Wichmann, F. A., & Hill, N. J. (2001b). The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Perception and Psychophysics*, *63*(8), 1314–1329. 10.3758/BF03194545

Wichmann, F. A., & Jäkel, F. (2018). Methods in psychophysics. In E. J. Wagenmakers & J. T. Wixted (Eds.), *Stevens' handbook of experimental psychology and cognitive neuroscience* (5th ed.) (pp. 1–42). Hoboken, NJ: Wiley.

Zchaluk, K., & Foster, D. H. (2009). Model-free estimation of the psychometric function. *Attention, Perception, and Psychophysics*, *71*(6), 1414–1425. 10.3758/APP.71.6.1414

Zwicker, T. (2000). Psychoacoustics as the basis for modern audio signal data compression. *Journal of the Acoustical Society of America*, *107*(5), 2875–2875. 10.1121/1.428677

Received April 13, 2021; accepted August 2, 2021.