The Spanish Journal of Psychology
2005, Vol. 8, No. 2, 256-289

ISSN 1138-7416

Sampling Plans for Fitting the Psychometric Function

Miguel A. García-Pérez and Rocío Alcalá-Quintana
Universidad Complutense de Madrid

Research on estimation of a psychometric function Ψ has usually focused on comparing alternative algorithms to apply to the data, rarely addressing how best to gather the data themselves (i.e., what sampling plan best deploys the affordable number of trials). Simulation methods were used here to assess the performance of several sampling plans in yes–no and forced-choice tasks, including the QUEST method and several variants of up–down staircases and of the method of constant stimuli (MOCS). We also assessed the efficacy of four parameter estimation methods. Performance comparisons were based on analyses of usability (i.e., the percentage of times that a plan yields usable data for the estimation of all the parameters of Ψ) and of the resultant distributions of parameter estimates. Maximum likelihood turned out to be the best parameter estimation method. As for sampling plans, QUEST never exceeded 80% usability even when 1000 trials were administered and rendered accurate estimates of threshold but misestimated the remaining parameters. MOCS and up–down staircases yielded similar and acceptable usability (above 95% with 400–500 trials) and, although neither type of plan allowed estimating all parameters with optimal precision, each type appeared well suited to estimating a distinct subset of parameters. An analysis of the causes of this differential suitability allowed designing alternative sampling plans (all based on up–down staircases) for yes–no and forced-choice tasks. These alternative plans rendered near optimal distributions of estimates for all parameters. The results just described apply when the fitted Ψ has the same mathematical form as the actual Ψ generating the data; in case of form mismatch, all parameters except threshold were generally misestimated but the relative performance of all the sampling plans remained identical. Detailed practical recommendations are given.

Keywords: psychometric function, psychophysical methods, least squares, maximum likelihood, simulation

Los estudios sobre estimación de la función psicométrica Ψ se han centrado tradicionalmente en comparar los algoritmos que se pueden aplicar a los datos, dejando al margen el problema de cómo recoger los propios datos (es decir, qué esquema de muestreo despliega de mejor forma los ensayos disponibles). Aquí se utilizan técnicas de simulación para evaluar el rendimiento de varios esquemas de muestreo en tareas de sí–no y de elección forzada, incluyendo QUEST y distintas variantes de escaleras de paso fijo y del método de los estímulos constantes. También se evalúa la eficacia de cuatro métodos de estimación de parámetros. Las comparaciones se basan en análisis de usabilidad (es decir, del porcentaje de veces que un esquema proporciona datos válidos para estimar todos los parámetros de Ψ) y de las distribuciones de las estimaciones. El mejor método de estimación resultó ser el de máxima verosimilitud. En cuanto a esquemas de muestreo, QUEST no llegó a rendir una usabilidad del 80% ni siquiera cuando se administraron 1000 ensayos y, aunque proporcionó buenas estimaciones del umbral, estimó erróneamente el resto de los parámetros. El método de los estímulos constantes y las escaleras de paso fijo rindieron una usabilidad similar (superior al 95% con 400–500 ensayos) y, aunque ninguno de estos esquemas permitió estimar con precisión óptima todos los parámetros, cada tipo de esquema se mostró adecuado para estimar un subconjunto distinto de parámetros. El análisis de las causas de estas diferencias permitió diseñar esquemas alternativos (todos ellos basados en escaleras de paso fijo) para tareas de sí–no y de elección forzada. Estos esquemas alternativos proporcionaron estimaciones con distribuciones casi óptimas. Los resultados descritos son válidos cuando la función cuyos parámetros se estiman tiene la misma forma analítica que la función psicométrica que ha generado los datos; cuando esas funciones difieren en forma, todos los parámetros excepto el umbral resultan estimados erróneamente, aunque la eficacia relativa de los distintos esquemas de muestreo no varía. Se ofrecen recomendaciones prácticas basadas en estos resultados.

Palabras clave: función psicométrica, métodos psicofísicos, mínimos cuadrados, máxima verosimilitud, simulación

Acknowledgments: This research was supported by Grant AP2001–0759 from Ministerio de Educación, Cultura y Deporte and by Grant BSO2001–1685 from Ministerio de Ciencia y Tecnología.

Address correspondence to: Miguel A. García-Pérez, Departamento de Metodología, Facultad de Psicología, Universidad Complutense, Campus de Somosaguas, 28223 Madrid (Spain). Phone: +34 913 943 061. Fax: +34 913 943 189.

E-mail: miguel@psi.ucm.es

1. Introduction

When the necessity to estimate the psychometric function Ψ arises in psychophysics, the method of constant stimuli (MOCS) appears to be the only available option. The tight association between estimating Ψ and using MOCS perhaps developed due to lack of alternative methods, but the link currently continues to be reinforced by the fact that even the most recent explicit attempts to determine how best to estimate Ψ either have only compared variants of MOCS or have only proposed or evaluated statistical criteria or algorithms to deal with MOCS data (Foster & Bischof, 1991; Lam, Mills, & Dubno, 1996; Maloney, 1990; Miller & Ulrich, 2001; O'Regan & Humbert, 1989; Treutwein & Strasburger, 1999; Wichmann & Hill, 2001a, 2001b).

The above statement notwithstanding, estimation of Ψ using adaptive methods has been described (Brand & Kollmeier, 2002; Hall, 1981; Kaernbach, 2001; Leek, Hanna, & Marshall, 1992; Serrano-Pedraza & Sierra-Vázquez, 2003; Swanson & Birch, 1992; Treutwein & Strasburger, 1999; Werkhoven & Snippe, 1996). Yet, none of those papers undertook an analysis that could help to determine an optimal sampling plan. Either the performance of the methods was barely or not at all compared with that of MOCS for all cases of interest or only estimation of a subset of the parameters of Ψ was considered. In practice, adaptive methods designed to place trials near threshold have been used to estimate all the parameters of Ψ (e.g., Strasburger, 2001; Treutwein & Strasburger, 1999; Watson & Turano, 1995). All things considered, the question is yet unsolved as to what psychophysical method places its allowance of trials at the stimulus levels that turn out to be most useful for an accurate estimation of Ψ. In other words, if a psychophysicist can only afford, say, 400 trials, what psychophysical method should he/she use to ensure an optimal deployment? This study looked directly into this question for arbitrary numbers of affordable trials.

The main focus of this paper is the estimation of all the parameters of Ψ, including its asymptotes. Quite often researchers are only interested in estimating the slope or threshold parameters, but it should be noted that Wichmann and Hill (2001a) have shown that estimates of these parameters can be substantially inaccurate if the asymptotes of Ψ are not estimated concurrently. Therefore, the issues addressed in this paper are also relevant when only accurate estimates of threshold and slope are required.

This work compares the performance of several MOCS sampling plans with that of plans arising from adaptive methods whose suitability for the estimation of Ψ has never been studied in depth. The comparison uses lack of bias and efficiency in parameter estimation as criteria. Our ultimate goal is to provide practicing psychophysicists with instructions as to how to configure the psychophysical method that renders the best sampling plan for the estimation of Ψ.

Issues of goodness of fit will not be addressed here because our goal is to compare alternative sampling plans as to the accuracy and precision of the parameter estimates that they provide. Thus, our focus is in the comparison of estimated and true parameters, whereas goodness-of-fit tests compare estimated parameters with the data. Wichmann and Hill (2001a, 2001b) have shown how to test goodness of fit and obtain confidence intervals bypassing the non-dependable features of off-the-shelf statistics, and the bootstrap method that they analyzed can also be applied to data arising from the sampling plans that we consider here.

This research addressed two experimental situations each of which has its own peculiarities, poses its own difficulties, and, as it turned out, requires a different sampling plan for the estimation of Ψ. These are known as "yes–no tasks" (where the lower asymptote of Ψ is a free parameter representing a false positive rate that is close to zero) and "m-alternative forced-choice tasks" (or mAFC with $m \geq 2$, where the lower asymptote is constant at $1/m$). Note that the link between mAFC tasks and a fixed lower asymptote is only justifiable in *mAFC detection tasks* where one of the intervals presents a stimulus and the other presents a blank, although it also applies to *mAFC identification tasks* (where one of the intervals presents a target and the others present equally attractive distractors, or when there is a single presentation of one of m equally confusable stimuli and the subject's task is to indicate which one it was). We will tag our references as mAFC *detection tasks* (implicitly referring to identification tasks too) to distinguish them from the *2AFC discrimination tasks* described in Section 6.1, where the lower asymptote becomes a free parameter with characteristics that make Ψ very similar to that in yes–no tasks.

The paper is organized as follows. Section 2 describes the sampling plans under study. Section 3 describes the remaining factors in the study and other details of our method. Sections 4–6 present our results and highlight the strong and weak points of each sampling plan. Section 7 proposes new plans based on combinations that ensure the concurrence of strong points across several adaptive strategies, and the improved performance of these new plans is also documented there. Finally, Section 8 summarizes our results and gives practical recommendations.

2. Sampling Plans

Our sampling plans arise from psychophysical methods that are fairly well known to practicing researchers. Some of these were designed for yes–no tasks and cannot be used with forced-choice tasks; others were designed to be used with the latter and are unlikely to be useful with the former; and others can be used in either case. Next we will describe them briefly to stress the characteristics of the ensuing sampling plans whose comparison is the goal of this study.

2.1. Conventional MOCS

Conventional MOCS with a fixed number of trials can be set up in innumerable forms by varying the number and spacing of stimulus levels. O'Regan and Humbert (1989; see also Brand & Kollmeier, 2002) carried out a theoretical analysis elucidating the locations where three levels should be placed to minimize the asymptotic variance of maximum likelihood estimates of threshold and slope in the simplest case. Wichmann and Hill (2001a, 2001b) used simulations to evaluate seven MOCS plans that differed as to the spacing of six stimulus levels and their location relative to that of Ψ, showing that some of these plans are more prone to bias or imprecise estimation. In laboratory practice, five to ten stimulus levels are normally used and they are often equally spaced, but it is much harder to set a precise location for these levels with respect to Ψ, or their spacing relative to the support σ of Ψ (a look at Figure 2 below will be useful at this point to find out what we mean by σ; references to the support σ of Ψ are frequent in this section).

Figure 1a illustrates conventional MOCS with five levels and a lattice that spans the region of support of Ψ and is centered with it, although number of levels and positioning varied in our study (see Section 3.4). The panel on the right of Figure 1a shows that MOCS allows the agreement between binned data (circles) and estimated function (solid curve) to be judged.

2.2. Single-Presentation MOCS

Single-presentation MOCS is a degenerate variant in which only one trial is placed at each level. Interest in this plan originated in bioassay, where high cost makes observations scarce. An initial evaluation (Ramsey, 1972) revealed the merits of 6-level single-presentation MOCS compared to conventional MOCS with the same number of observations, and subsequent analyses in that context confirmed the conclusion (Müller & Schmitt, 1990). Treutwein and Strasburger (1999) evaluated single-presentation MOCS with 100 equally-spaced levels over the range of interest. Figure 1b shows an illustration, and note in the right panel that the unappealing aspect of the data leaves our eyes unable to judge the fit of the estimated function (solid curve).

2.3. QUEST

QUEST is a parametric, adaptive Bayesian threshold estimation method that was proposed by Watson and Pelli (1983) after it had been developed and tested in the fields of bioassay (Freeman, 1970; Marks, 1962; Ramsey, 1972) and educational and psychological testing (Owen, 1975). Some groups (e.g., Santoro, Burr, & Morrone, 2002; Simmers, Bex, Smith, & Wilkins, 2001; Snowden & Hammett, 1998; Solomon & Morgan, 2000; Watson & Turano, 1995) fit Ψ to QUEST data even though QUEST was never shown to provide dependable estimates of Ψ.[1]

Alcalá-Quintana and García-Pérez (2002, 2004a) showed that the setups of Watson and Pelli (1983) and King-Smith, Grigsby, Vingrys, Benes, and Supowit (1994) are suboptimal, and they identified a dependable setup. Here we gave QUEST a further advantage and set it up in ideal conditions: The model function built into QUEST matches the actual Ψ producing the data. Thus, QUEST performs here at its best, and its expected performance in practice is inferior. QUEST is illustrated in Figure 1c, showing that it places trials unevenly across the available range of stimulus levels. The highest density of trials occurs around the target point and, as a result of the ever changing step size, most levels are tested merely once but a few levels get tested 2–4 times.

2.4. Adaptive Staircase With Up–Down Transformed Rules (UDTRS)

Also with the goal of estimating threshold but using a non-parametric method, Wetherill and Levitt (1965; see also Brown, 1996) modified the up–down method of Dixon and Mood (1948) so that the procedure targets a point other than that at which the probability of success is 0.5. This is accomplished by varying what is called the *up–down rule*, which refers to the number of consecutive correct (alternatively, incorrect) responses that are required at the current stimulus level to bring it down (alternatively, up) by one step for the next trial. For use with *m*AFC detection tasks we will consider 1–2, 1–3, and 1–4 rules, where *u*–*d* stands for "*u* consecutive wrong responses take the stimulus level one step up and *d* consecutive correct responses take the stimulus level one step down." Wetherill and Levitt (1965) referred to these as *transformed rules* because they differ from the original 1–1 rule of Dixon and Mood (1948). Although the conventional estimator based on the average of the stimulus levels at the reversal points has been shown to be biased and non-dependable under realistic conditions

[1] Quite on the contrary, a number of authors have claimed that QUEST is not well suited to estimating the slope of Ψ and they have hence modified its placement rule so as to allow the simultaneous estimation of threshold and slope (see King-Smith & Rose, 1997; Kontsevich & Tyler, 1999; Snoeren & Puts, 1997). The reason that we are including QUEST in our study and not any of these alternative methods is that QUEST has been extensively improved through simulation studies and that it has also been extensively used in empirical research as a method both for estimating threshold and for fitting a psychometric function to the resultant data. On the other hand, the alternative methods have not been assessed by simulation or developed beyond the small-scale studies in the original papers, and a cited-reference search carried out on December 13, 2004 in the Science Citation Index only returned five papers in which some of these three methods had been used in experimental psychophysics, compared to more than 150 papers that have used QUEST or its variants for analogous purposes, including cases in which it was used to estimate Ψ.

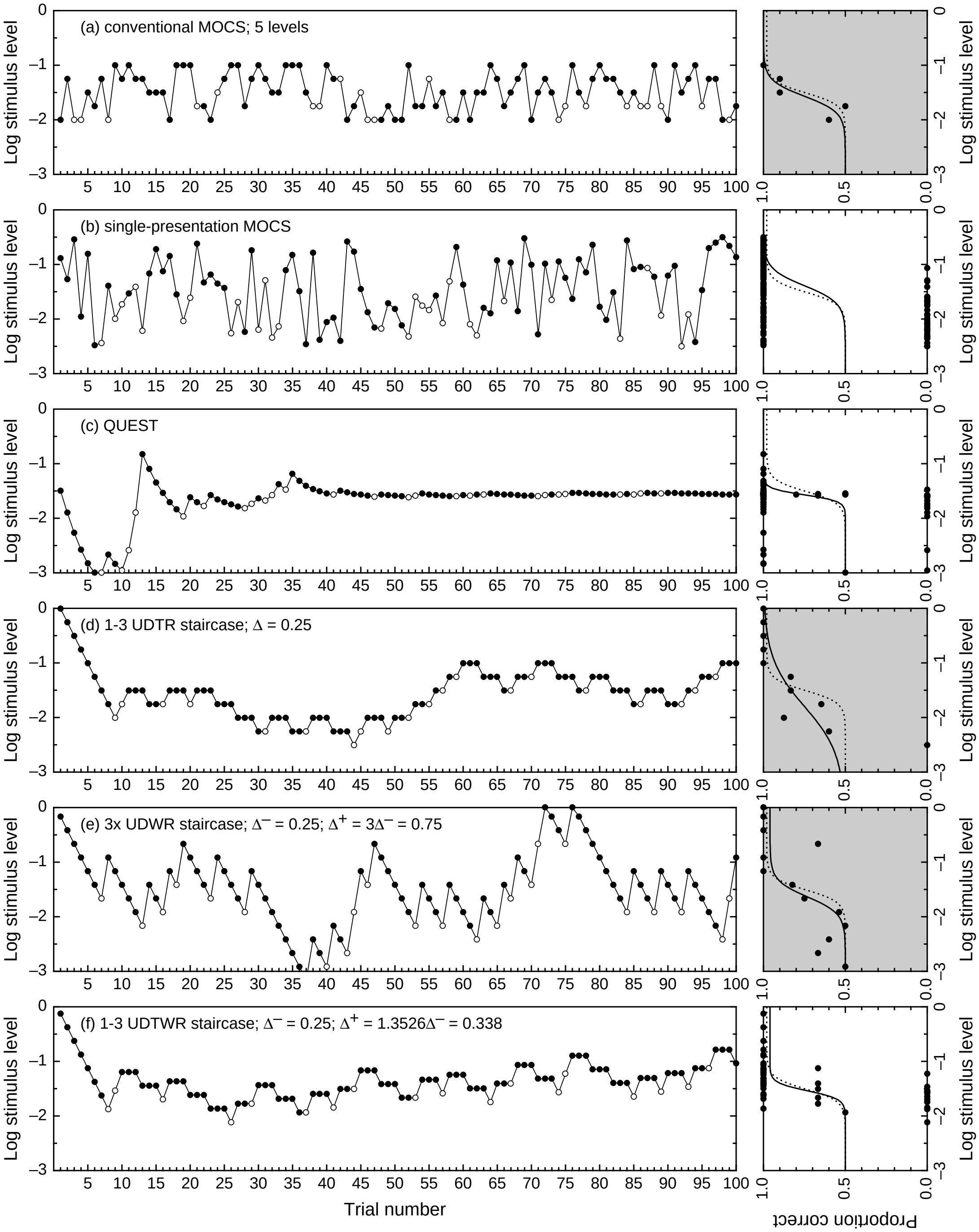

Figure 1. Sample tracks (left column) and results (right column) of the application of six of the sampling plans analyzed in this study. These incidental results do not necessarily reflect the average performance of each plan, which is shown in Figure 4. All psychophysical methods ran 100 trials of a putative 2AFC detection task, where the lower asymptote of Ψ is fixed at 0.5. Each circle in the track on the left column indicates the level (ordinate) at which the trial (abscissa) was given and also indicates its outcome (open: wrong response; solid: correct response). The actual Ψ of the simulated subject, shown as a dotted curve in the panels on the right column, was identical in all cases: a logistic function with $\gamma = 0.5$, $\lambda = 0.02$, $\theta = -1.5$, and $\sigma = 1$ (see Equations 1–3 in Section 3.1). The solid curve on the right panels is the best-fitting logistic function with maximum likelihood estimates of its parameters (except γ, which is fixed here), in each case using binned data represented on the right panels as solid circles. The three grayed panels on the right column mark alternative sampling plans that give rise to lattices with identical spacing. (a) Conventional MOCS with 5 equidistant levels that cover the entire support Ψ and are centered within that region. (b) Single-presentation MOCS covering the central two log units of the stimulus range. (c) QUEST. (d) 1–3 UDTR staircase with steps up and down whose size is $\Delta = 1/4$. (e) 3× UDWR staircase with steps down of size $\Delta^- = 1/4$ too. (f) 1–3 UDTWR staircase (thus using $\Delta^+ = 1.3526\Delta^-$) with $\Delta^- = 1/4$ too.

(García-Pérez, 1998), UDTR staircases place trials over and over at a small number of stimulus levels in a given lattice and, then, the resultant data can be used to estimate Ψ.

The spacing between levels (what is called the *step size* Δ) as well as the relative location of the lattice (determined by what is called the *starting point*) must be chosen unaidedly by the practitioner. Our study used a number of variants for these characteristics (see Section 3.4). The illustration in Figure 1d involves a 1–3 rule with $\Delta = 0.25$, showing also other features that will be described in Sections 3.4 and 3.6. Note that this step size yields the same spacing between levels as was used with conventional MOCS in Figure 1a.

With yes–no tasks we also used 1–1, 2–1, 3–1, and 4–1 rules. The reason for not using these with *m*AFC detection tasks lies in the workings of UDTRs. In general, 1–*d*, 1–1, and *u*–1 rules (with u, $d > 1$) respectively place trials around points where the probability of success is above, at, and below 0.5. Only 1–*d* rules are then useful when Ψ has a lower asymptote at 0.5. Figure 16c will illustrate a 3–1 UDTR staircase.

2.5. *Adaptive Staircase With Up–Down Weighted Rules (UDWRS)*

Also to allow up–down methods to estimate arbitrary points on Ψ, Kaernbach (1991) proposed retaining the 1–1 rule but letting the size Δ^+ of a step up be an integer multiple of the size Δ^- of a step down, yielding what he called *weighted rules*. We will refer to these as $k\times$ UDWR staircases, where $k \geq 2$ is the integer just mentioned. The average of reversals estimator is also errant under $k\times$ UDWR staircases (García-Pérez, 1998), but UDWR staircases again provide an alternative sampling plan that may be useful for estimating Ψ. UDWR staircases with $k = 2$, 3, and 4 were implemented for use with *m*AFC detection tasks. Figure 1e illustrates a $3\times$ UDWR staircase with $\Delta^- = 0.25$ and, hence, $\Delta^+ = 0.75$. Note that this step size yields again the same spacing as was used with conventional MOCS in Figure 1a.

For use with yes–no tasks, we also considered $\frac{1}{k}\times$ UDWR staircases (with $k = 2$, 3, and 4), implying cases in which Δ^- is an integer multiple of Δ^+. The reason lies again in the extended range of Ψ in yes–no tasks. Figure 16b will illustrate a $\frac{1}{3}\times$ UDWR staircase.

2.6. *Adaptive Staircase With Up–Down Transformed and Weighted Rules (UDTWRS)*

García-Pérez (1998) showed that the average of reversals estimator is dependable only when the ratio of Δ^+ to Δ^- has a specific value that covaries with up–down rule. This leads to transformed and weighted rules, but not in just any form. Specifically, the ratio of Δ^+ to Δ^- should take the values 3.5149, 1.8222, 1.3526, and 1.1884 respectively for use with the 1–1, 1–2, 1–3, and 1–4 rules in 2AFC detection tasks. Because UDTWR staircases yield yet another type of sampling plan, their capability to provide useful data for the estimation of Ψ was evaluated here.

Our study with 2AFC detection tasks included 1–1, 1–2, 1–3, and 1–4 UDTWR staircases. Figure 1f illustrates the 1–3 case with $\Delta^- = 0.25$ and, hence, $\Delta^+ = 0.338$. UDTWR staircases were not used with general *m*AFC detection tasks because the ratios of Δ^+ to Δ^- mentioned in the previous paragraph place trials around a stable point in the upper half of the region of support of Ψ only when $m = 2$ and unpublished results show that stability breaks down in 3AFC or 4AFC. And they were not used with yes–no tasks because results in García-Pérez (2001) show that no ratios exist that place trials around a stable point in the lower half of the region of support of Ψ.

3. Method

Our results are based on Monte Carlo simulations using custom software. Ten thousand replicates were run per condition, defined as a unique combination of form and parameters for Ψ, overall number N of trials, and sampling plan. The latter comprises a total of six variants of conventional MOCS, one variant of single-presentation MOCS, one of QUEST, 27 variants of UDTR staircases arising from the factorial combination of three rules and nine step sizes (63 variants in the case of yes–no tasks, because of the use of four additional rules), 27 variants of UDWR staircases (54 for yes–no tasks), and 36 variants of UDTWR staircases (none for yes–no tasks). Sections 3.1–3.5 describe all factors included in the design except sampling plan, which was described in Section 2. Sections 3.6–3.8 give details on implementation and data analysis.

3.1. *Psychometric Functions and Ranges for Their Parameters*

Using visual contrast detection tasks as a referent, and without loss of generality, the domain of Ψ was defined to be the negative real line as if the relevant physical variable x were log contrast.[2] Thus, the upper bound $x_u = 0$ of the domain of Ψ is unsurpassable. For methods that require a lower bound (single-presentation MOCS and QUEST), a bounded range was defined as the interval $[x_l, x_u] = [-3, 0]$, which is the range spanned by the vertical axis in Figure 1.

[2] In general, x can be thought of as the variable representing the stimulus dimension that is relevant to the task, where x is measured in whichever units are appropriate.

The psychometric function Ψ is a four-parameter function whose mathematical form could be logistic or Weibull in our study. A logistic Ψ is given by

$$\Psi(x) = \gamma + \frac{1 - \lambda - \gamma}{1 + \exp[-b(x - T)]}, \quad (1)$$

whose parameters are described next (see also Figure 2). Parameter γ sets a lower asymptote reflecting either a *false positive* rate (in yes–no tasks) or a *guessing* rate (in *m*AFC detection tasks). In the former case, γ is a free parameter with values near zero; in the latter, γ is a fixed constant supposed to be valued at $1/m$. Parameter λ is always free and represents the *false negative* or *lapsing* rate, its value is also near zero, and it sets an upper asymptote at $1 - \lambda$. Parameter T determines the location of Ψ by labeling the point satisfying $\Psi(T) = (1 - \lambda + \gamma)/2$, that is, the point at which the probability of success is halfway between γ and $1 - \lambda$. Finally, b is a slope parameter in that the slope of Ψ at $x = T$ equals $b(1 - \lambda - \gamma)/4$.

As discussed in García-Pérez (1998, 2001) and Alcalá-Quintana and García-Pérez (2004a), a convenient reparameterization of Ψ replaces the slope parameter b with a support parameter σ that describes the effective width of Ψ, and also replaces the location parameter T with a threshold parameter θ that labels the point satisfying $\Psi(\theta) = \pi$ for arbitrary $\pi \in (\gamma, 1 - \lambda)$ instead of the midpoint of the range (see Figure 2). This reparameterization uses the transformations

$$\theta = T + \frac{1}{b} \ln\left[\frac{\pi - \gamma}{1 - \lambda - \pi}\right], \quad (2)$$

$$\sigma = \frac{2}{b} \ln\left[\frac{1 - \lambda - \gamma - \delta}{\delta}\right], \quad (3)$$

where δ is an auxiliary parameter defined here as $\delta = (1 - \lambda - \gamma)/100$. With this δ, Equation 3 becomes $\sigma = 9.19/b$ so that σ is inversely related to b and represents the width of the region over which Ψ spans the central 98% of its range, whichever this range is as determined by the values of γ and λ. Formally, given some δ satisfying $0 < \delta < (1 - \lambda - \gamma)/2$, we define $\sigma = \Psi^{-1}(1 - \lambda - \delta) - \Psi^{-1}(\gamma + \delta)$, that is, the distance between the point at which Ψ evaluates to $\gamma + \delta$ (slightly above the lower asymptote) and that at which Ψ evaluates to $1 - \lambda - \delta$ (slightly below the upper asymptote). With reference to Figure 2, the *region of support* of Ψ is the range of stimulus levels within the two vertical lines

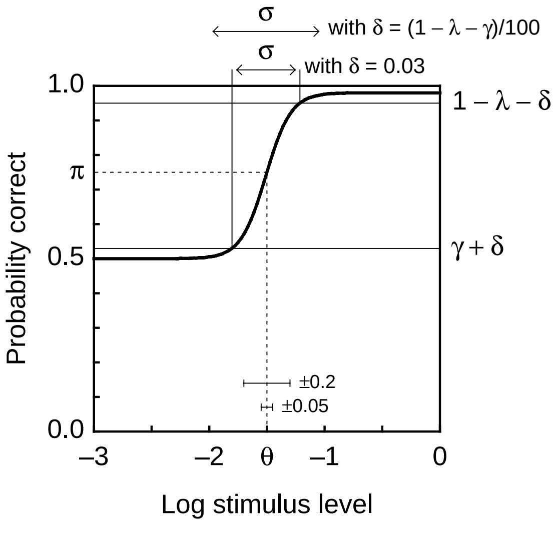

Figure 2. Meaning of the parameters of a logistic Ψ. Parameter γ is here assumed to be a guessing rate fixed at 0.5 (as in a 2AFC detection task). Parameter λ is set here at 0.02 so that the upper horizontal asymptote lies at $1 - \lambda = 0.98$. Parameter T is not marked here. Under our reparameterization, a probability π is chosen ($\pi = 0.75$ here) and parameter θ ($\theta = -1.5$ here) is the point at which Ψ evaluates to π (see the dashed line reflecting the horizontal location of θ onto the vertical location of π). Parameter b is devoid of meaning, but our replacement parameter σ indicates the support of Ψ defined as the width of the horizontal region required for Ψ to span some central percentage of its total range, that is, the width of the central region where Ψ shows non-asymptotic behavior. The particular percentage chosen determines the value of an auxiliary parameter δ, which is added to the lower asymptote and subtracted from the upper asymptote to draw the two solid horizontal lines across the graph. Each of these lines crosses Ψ at one point, and the horizontal distance between those two points represents the support of Ψ ($\sigma = 0.589$ here). In this illustration we set $\delta = 0.03$ (yielding the width of the central 87.5% of the range of Ψ) so that the relevant lines and crossings are visible; our study used $\delta = (1 - \lambda - \gamma)/100$, so that σ measures the central 98% of the range of Ψ. Relative to this latter value for δ the curve plotted here has indeed $\sigma = 1$ and is thus the same as was used in Figure 1. For later reference in Section 4, the two small horizontal segments near the bottom of the plot span ±0.2 and ±0.05 units around θ.

used to determine the support of Ψ and describes the region where the probability of a correct response varies with stimulus level. Above (alternatively, below) the region of support, the probability of a correct response is independent of stimulus level and is only determined by the false negative (alternatively, false positive or guessing) rate.

A Weibull Ψ, on the other hand, is given by

$$\Psi(x) = 1 - \lambda - (1 - \lambda - \gamma)\exp\left[-10^{\beta(x-T)}\right], \qquad (4)$$

where all parameters have the same meaning but T now satisfies $\Psi(T) = (1 - \lambda) - (1 - \lambda - \gamma)/\mathrm{e}$ and β (which replaces b) makes the slope of Ψ at $x = T$ equal to $\beta(1 - \lambda - \gamma)\ln(10)/\mathrm{e}$. The alternative threshold and support parameters in a Weibull Ψ are given by

$$\theta = T + \frac{1}{b}\log\left[\ln(1 - \lambda - \gamma) - \ln(1 - \lambda - \pi)\right], \qquad (5)$$

$$\sigma = \frac{1}{\beta}\log\left[\frac{\ln[\delta/(1 - \lambda - \gamma)]}{\ln[1 - \delta/(1 - \lambda - \gamma)]}\right], \qquad (6)$$

with π and δ as above, the latter reducing Equation 6 to $\sigma = 2.66/\beta$. The factor 2.66 (compare to 9.19 for the logistic Ψ above) guarantees that when Weibull and logistic functions have the same support σ, the relevant range of stimulus levels has identical width in both cases.

Throughout our study, π was set at 0.5 when γ was a false positive rate (i.e., in yes–no tasks) and it was set at $(m + 1)/2m$ when γ was a guessing rate fixed at $1/m$ (i.e., in *m*AFC detection tasks; we included 2AFC, 3AFC, and 4AFC). For the 2AFC discrimination task in Section 6.1, where γ is replaced with λ, we set $\pi = 0.5$. Besides this context-dependent value for π and an invariant $\delta = (1 - \lambda - \gamma)/100$, the parameters of Ψ varied systematically across conditions. Lapsing rates λ varied from 0 to 0.06, and γ also took on values in that range when it was a free parameter. Threshold θ ranged from –2.5 to –0.5 and σ ranged from 0.5 to 1.5; the corresponding values for T, b, and β were then obtained from θ and σ by reversing Equations 2, 3, 5, and 6.

3.2. Mathematical Form of the Fitted Function $\breve{\Psi}$

In practice, experimenters choose to fit a function $\breve{\Psi}$ without information as to whether its mathematical form matches that of the actual Ψ. Two popular forms for $\breve{\Psi}$ are Weibull and logistic. The consequences of a mismatch between the forms of Ψ and $\breve{\Psi}$ were evaluated by fitting both logistic and Weibull $\breve{\Psi}$ to data generated with either logistic or Weibull Ψs.

3.3. Overall Number of Trials

Each sampling plan was evaluated for numbers N of trials between 100 and 1000 in steps of 100. An exception had to be made with conventional MOCS, where the number L of levels we used (from 5 to 10; see Section 3.4) and the requirement that all levels are tried the same number of times cannot accommodate our values for N. We gave MOCS a slight advantage and each of its variants ran for the number of trials that fulfills MOCS requirements in the least possible excess of the number of trials ran by its competitors. Thus, conventional MOCS with L levels and N nominal trials ran instead for $L(1 + (N - 1) \div L)$ trials, where ÷ indicates integer division.

3.4. Spacing Between Stimulus Levels and Positioning of the Lattice

In conventional MOCS, the spacing between levels is determined by the number L of levels that cover the range of exploration. We used L from 5 to 10 and spacing amounts to $\sigma/(L - 1)$ because we gave conventional MOCS the apparent advantage that the range of exploration spanned the support of Ψ. Thus, at least one level was included at which the probability of success exceeds 0.95 (as recommended by Wichmann & Hill, 2001b). The lattice was not always centered as in the illustration in Figure 1a. Instead, in each replicate it was jittered from this central position by an amount that was determined by drawing a random number with a uniform distribution on $[-\sigma/2(L - 1), \sigma/2(L - 1)]$, that is, plus and minus half the spacing between levels.

In single-presentation MOCS, spacing is also determined by L (which equals N) and amounts to $(x_u - x_l)/(L - 1)$, where x_u and x_l are the upper and lower limits of the available stimulus range (see Section 3.1). Note that the available stimulus range covers all three log units shown on the ordinate of the panels in Figure 1 and not only the inner two log units that Figure 1b suggested. The lattice was centered within the available stimulus range regardless of the location of Ψ.

QUEST uses its own rules to determine the placement of each trial and, hence, there is no room here for choosing spacing or positioning of the lattice.

Under UDTR staircases, steps up and down have the same size Δ, which will be referred to as the *base spacing*. Under UDTWR staircases and also under k× UDWR staircases, $\Delta^- < \Delta^+$, whereas under $\frac{1}{k}$ × UDWR staircases, $\Delta^- > \Delta^+$. In cases where $\Delta^- \neq \Delta^+$, the smaller was regarded as the base spacing Δ. In Figures 1d–1f, $\Delta = \sigma/4 = 0.25$, but Δ in our study varied between $\sigma/10$ and $\sigma/2$ for all integer values in the denominator, thus yielding nine different variants along the dimension of step size. In all adaptive staircases, positioning of the lattice was set by choosing a random starting point for each replicate, determined either by subtracting from the upper limit x_u or by adding to the lower limit x_l a random number drawn from a uniform distribution on $[0, \Delta]$. The upper starting point was used with 1–d UDTR staircases, k× UDWR staircases, and all UDTWR staircases; the lower starting point was used with u–1 UDTR staircases and $\frac{1}{k}$ × UDWR staircases. Use of this initial jitter with an upper starting point can be noted in the left panels of Figures 1d–1f; lower starting points will be illustrated in Figures 16b and 16c.

3.5. Parameter Estimation

Responses across the N trials in a given replicate were binned by stimulus level (rounded to the nearest ten thousandth) for the J distinct levels tried in that replicate and the numbers c_j and w_j of correct and wrong (or yes and no) responses as well as the proportion p_j of correct (or yes) responses at level x_j were computed (the latter are plotted as circles in the right column of Figure 1). These data were used to obtain ordinary and weighted least squares (OLS and WLS), maximum likelihood (ML), and Bayes quadratic (BQ) estimates of the free parameters of $\breve{\Psi}$. Our interest in studying these alternative estimators lies in that they have all been used in empirical practice, but also in that neither ML nor LS methods are guaranteed to always yield an absolute optimal solution (see Figure 3).

Weighted least squares estimates are obtained by minimizing

$$\sum_{j=1}^{J} a_j \left(\breve{\Psi}(x_j)-p_j\right)^2, \qquad (7)$$

with respect to $\hat{\gamma}$ (when not a fixed guessing rate), $\hat{\lambda}$, $\hat{T}$, and either $\hat{b}$ (for logistic $\breve{\Psi}$) or $\hat{\beta}$ (for Weibull $\breve{\Psi}$). Because the errors are not distributed normally and the variances are not the same at all stimulus levels j, the weights a_j in Equation 7 must be set as $a_j = (c_j + w_j)/c_j w_j$ (Myers, 1990, pp. 317–320). Application of WLS must thus discard data for all stimulus levels at which either c_j or w_j are zero, something that occurs for all j under single-presentation MOCS (see the right panel in Figure 1b), for most stimulus levels under QUEST (see Figure 1c) and UDTWR staircases (see Figure 1f), and for a non-trivial subset of stimulus levels under conventional MOCS (see Figure 1a) and our remaining adaptive staircases (see Figures 1d and 1e). As a result, WLS either is inapplicable or discards a significant amount of relevant data when used with our sampling plans. An alternative is to apply OLS, where $a_j = 1$ for all j in Equation 7. Although OLS is theoretically inappropriate for the reasons stated above, there is still the issue of whether it can reasonably recover the parameters of Ψ when WLS is either inapplicable or its application must discard informative data. We will thus use both WLS and OLS with all sampling plans except single-presentation MOCS, where WLS is not applicable.

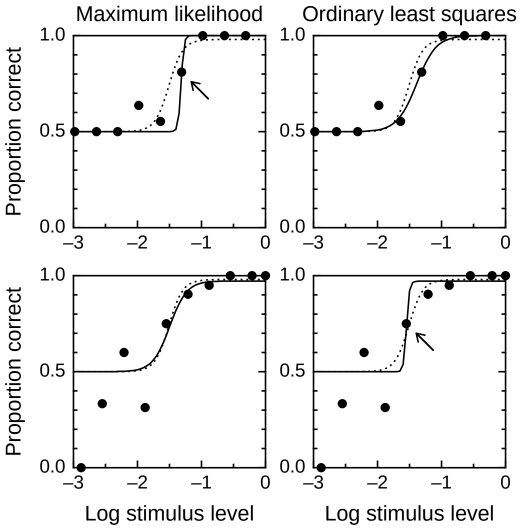

Figure 3. Maximum likelihood can provide a sensible solution when ordinary least squares cannot, and vice versa. The top row shows a sample data set (solid circles) generated by the same Ψ (dotted curve) used in Figure 1: a logistic function with $\gamma = 0.5$, $\lambda = 0.02$, $\theta = -1.5$, and $\sigma = 1$, equivalent to $T = -1.51$ and $b = 9.19$ by Equations 2 and 3. The bottom row shows a second data set generated with the same Ψ. Both data sets arose with 3× UDWR staircases with $\Delta^- = \sigma/3$. Solid curves in each panel represent best-fitting logistic functions obtained via maximum likelihood (left column) or ordinary least squares (right column). Arrows mark points that will be referred to in Section 8.2 below. Each method fails with one of the data sets and does well with the other. Note that this appears to be a fundamental inappropriateness of each parameter estimation method for certain data sets, and does not reveal a failure of the algorithms themselves: The log likelihood is –40.72 for the ML solution in the top left panel ($\hat{\lambda} = 0.000$, $\hat{T} = -1.318$, $\hat{b} = 45.95$), which goes down to –41.16 for the OLS solution rendering the more appealing curve in the top right panel ($\hat{\lambda} = 0.000$, $\hat{T} = -1.388$, $\hat{b} = 6.561$); similarly, the mean square error is 0.0333 for the OLS solution in the bottom right panel ($\hat{\lambda} = 0.029$, $\hat{T} = -1.547$, $\hat{b} = 45.95$), which goes up to 0.0339 with the ML solution that rendered the better-looking curve in the bottom left panel ($\hat{\lambda} = 0.028$, $\hat{T} = -1.496$, $\hat{b} = 7.516$). Results for weighted least squares (not shown) were similar.

Maximum likelihood estimates are obtained by minimizing

$$-\sum_{j=1}^{J}\left[c_j \log\left[\breve{\Psi}(x_j)\right] + w_j \log\left[1-\breve{\Psi}(x_j)\right]\right], \tag{8}$$

and Bayes estimates under a quadratic loss function[3] are obtained by separately evaluating

$$\frac{\int_{\Omega} \hat{\xi} \prod_{j=1}^{J}\left(\breve{\Psi}(x_j)\right)^{c_j}\left(1-\breve{\Psi}(x_j)\right)^{w_j} \mathrm{d}\hat{\gamma}\,\mathrm{d}\hat{\lambda}\,\mathrm{d}\hat{T}\,\mathrm{d}\hat{b}}{\int_{\Omega} \prod_{j=1}^{J}\left(\breve{\Psi}(x_j)\right)^{c_j}\left(1-\breve{\Psi}(x_j)\right)^{w_j} \mathrm{d}\hat{\gamma}\,\mathrm{d}\hat{\lambda}\,\mathrm{d}\hat{T}\,\mathrm{d}\hat{b}} \tag{9}$$

for each $\hat{\xi} \in \{\hat{\gamma}, \hat{\lambda}, \hat{T}, \hat{b}\}$, and where Ω is the parameter space of dimensions $\hat{\gamma}$ (if not a constant), $\hat{\lambda}$, $\hat{T}$, and either $\hat{b}$ or $\hat{\beta}$. In our study, $\Omega = [0, 0.06] \times [0, 0.06] \times [-3, 0] \times [s_{inf}, s_{sup}]$, where $[s_{inf}, s_{sup}] = [1.84, 45.95]$ and $[0.53, 13.3]$ respectively for logistic and Weibull $\breve{\Psi}$ so that in both cases $\hat{\sigma} \in [0.2, 5]$.

Equations 7 and 8 were minimized with NAG subroutine E04JYF (Numerical Algorithms Group, 1999), and each parameter was constrained to remain within the bounds of the corresponding dimension of Ω. This is analogous to using uniform priors over the parameter space, as Wichmann and Hill (2001a) did. Because of the dependence of the final parameter estimates on their initial values (Serrano-Pedraza & Sierra-Vázquez, 2003), we ran the algorithm using a set of vectors of initial values and accepted the solution that yielded the lowest mean square error (under the WLS and OLS approaches) or the highest likelihood (under the ML approach). When γ had to be estimated, 24 initial vectors were used that resulted from the Cartesian product $\{0.015, 0.045\} \times \{0.015, 0.045\} \times \{-2, -1\} \times \{6.13, 9.19, 18.38\}$ (or $\{1.77, 2.66, 5.32\}$) where each set respectively describes alternative initial values for $\hat{\gamma}$, $\hat{\lambda}$, $\hat{T}$, and $\hat{b}$ (or $\hat{\beta}$). When γ was a fixed constant, the first set was dropped and only 12 vectors of initial values resulted.

Finally, for BQ estimates, the integrals in Equation 9 were solved numerically with NAG subroutine D01FCF (Numerical Algorithms Group, 1999). The integrals were solved in only three dimensions (i.e., excluding $\hat{\gamma}$) when γ was a fixed guessing rate.

Once these parameters had been estimated, $\hat{\sigma}$ and $\hat{\theta}$ were obtained from them through Equations 2, 3, 5, and 6, using estimates in place of true parameter values.

3.6. Implementation Details

Conventional and single-presentation MOCS did not require implementation decisions. QUEST was set up using the optimal configuration established by Alcalá-Quintana and García-Pérez (2002, 2004a), which uses a uniform prior on [–4, 1] (because $[x_l, x_u] = [-3, 0]$ here) and the prior mean as placement rule. The extended range was used only in computations, and requests to place trials at levels outside the available range were replaced by the corresponding boundary (see trials 6 and 7 in Figure 1c). QUEST was implemented by tabulating the model function and the prior distribution with a resolution of 1000 samples per unit, and it was set up to target the point at which the probability of success is 0.5 (in yes–no or 2AFC discrimination tasks) or $(m + 1)/2m$ (in mAFC detection tasks). A step by step description of the application of QUEST can be found in Alcalá-Quintana and García-Pérez (2004a, their Figure 1).

All staircases started off with the 1–1 rule either until the first wrong (or no) response (for upper starting points; see Figures 1d–1f) or until the first correct (or yes) response (for lower starting points; see Figures 16b and 16c). These preliminary trials under the 1–1 rule helped to approach the relevant region without wasting many trials. Contrary to Leek et al. (1992), we admit that these trials are part of the procedure. The stimulus range was unbounded low and trials could be placed anywhere on the negative line (as was required on trial 37 by the staircase in Figure 1e). Requests to place trials above the hard upper bound resulted in the stimulus being presented with the boundary level instead. Yet, trials after off-limits requests are placed relative to the required level and not to the boundary level, so that the staircase stays on the same lattice when it bounces off the upper bound (see trials 72 to 73 and 76 to 77 in Figure 1e).

3.7. Simulation Approach and Random Number Generation

To simulate a trial, the stimulus level set by the psychophysical method under consideration was inserted into Ψ to obtain the probability of success. The outcome was then simulated by drawing a Bernoulli variate via NAG subroutine G05DZF (Numerical Algorithms Group, 1999).

Our simulations make the typical assumption that the probability of success on a trial in which the stimulus level is x_i is determined only by $\Psi(x_i)$. In other words, we assume that all trials are statistically independent and that there are no sequential effects along the session.

[3] Use of a quadratic loss function in Bayesian estimation results in estimates valued at the mean of the final posterior distribution (i.e., the posterior distribution at the end of the procedure; see Alcalá-Quintana & García-Pérez, 2004a). In the one-parameter case, this yields what King-Smith et al. (1994, p. 890) called "mean-QUEST." King-Smith et al. (1994) also discussed the two-parameter case in their Appendix, where the estimates of each of the two parameters are given by their Equations 38 and 39. Our Equation 9 is a straightforward generalization to the case of four parameters. To save space, our equation is written in compact form (i.e., not broken up into four separate expressions).

3.8. Data Analysis

The optimal sampling plan is that which provides the best parameter estimates, that is, those with the narrowest symmetrical distribution around the true value. Thus, the empirical distributions of the 10,000 estimates of each parameter under each condition were plotted and mean, standard deviation, and skewness were computed. For clarifications and comments, Figure 4 shows histograms of ML estimates from N = 100 and 1000 for each of the six sampling plans in Figure 1. The actual, logistic Ψ was as shown in Figure 1; the fitted $\breve{\Psi}$ was also logistic.

Note in Figure 4a (for N = 100) that the distributions of estimates of λ have large spikes at either end of the horizontal axis and that the distributions of estimates of b also have a spike at the upper end which, by the inverse relation of b to σ, makes the distributions for σ display the spike at the lower end. (When N = 1000, these spikes are small with all methods except QUEST; see Figure 4b.) In the case of λ, whose true value was 0.02, the spike at $\hat{\lambda}$ = 0 merely reflects data failing to show evidence of lapses. Conversely, the spike at $\hat{\lambda}$ = 0.06 (the arbitrary upper bound for $\hat{\lambda}$ in our parameter space) reflects that the optimization algorithm sought to explore beyond the upper bound. In these cases, an estimate $\hat{\lambda}$ = 0.06 only reflects our own arbitrary decision as to where to place an upper bound and, hence, these values should not enter a statistical description of $\hat{\lambda}$. In line with Leek et al. (1992), we present statistics only for the subset of replicates for which the parameter of concern could actually be estimated. This is why the sample size n given in each panel varies along each row of Figures 4a or 4b, although histograms were drawn using all 10,000 estimates. A new criterion for the comparison of sampling plans should thus be considered, namely, the percentage of times that they produce usable data for the estimation of all parameters (the raw number n_{all} is given in the leftmost panel in each row), which we will refer to as the *usability index*. The relevance of this usability index will be best appreciated when we discuss the case of b next.

The spike at the upper end of the distributions of $\hat{b}$ contains a mixture of authentic estimates that lie beyond the range of the plot (but are represented there for convenience) and of improper estimates at the upper bound ($\hat{b}$ = 45.95; see Section 3.5). Improper estimates would have taken whichever value we could have set as an upper bound. The solid curves in the top left and bottom right panels of Figure 3 have this boundary value for $\hat{b}$, and it is easy to understand this event as a failure to estimate b. These improper estimates were excluded from computations. The reason that the spike at the lower end of the distributions of $\hat{\sigma}$ is generally smaller lies in that all authentic estimates included in the spike in the distribution of $\hat{b}$ are represented elsewhere and only improper estimates are treated for display purposes as if $\hat{\sigma}$ = 0.

4. Results: I. 2AFC Detection Tasks

This section reports results when γ is a guessing rate fixed at 0.5; results for different values (or status) of the lower asymptote of Ψ will be reported in Sections 5 and 6. For simplicity, here we will only present results for the case of logistic Ψ with λ = 0.02, θ = –1.5, and σ = 1 (b = 9.19), logistic $\breve{\Psi}$, and ML estimates. This case was chosen because the results for other cases (to be summarized in Section 6) are easily described with reference to it.

Before we document the merits of the various sampling plans, several differential characteristics of the distributions of estimates across target parameters and sample sizes are worth commenting on, besides the unsurprising result that quality improves as N increases. Parameter λ is rarely well estimated with 100 trials (first column in Figure 4a); with 1000 trials (first column in Figure 4b), it is still slightly underestimated and its estimates are too broadly distributed. Nevertheless, except for the ubiquitous spike at $\hat{\lambda}$ = 0 and the case of QUEST (third row in Figure 4b), the distribution of $\hat{\lambda}$ is approximately symmetric with N = 1000 (it was symmetric for N > 600).

Parameter θ (second column in Figures 4a and 4b) is estimated comparatively much better, never being improper and yielding narrow distributions centered on the true θ: Most methods place the estimate within 0.2 units of its true location when N = 100 and within 0.05 units when N = 1000. (A look at Figure 2 may be useful again, for it shows the actual Ψ used here along with small horizontal segments near the base of the plot that span the ranges just mentioned.) At the same time, the distributions of $\hat{\theta}$ only show mild traces of negative skewness when N = 100.

Parameter b (third column in Figures 4a and 4b) often fails to be estimated when N = 100, and the distribution of proper estimates is severely positively skewed. The indeterminacy of b is almost absent when N = 1000, but the distribution of estimates remains positively skewed and its mean lies above the true value of b. Interestingly, converting the estimate of b into an estimate of σ (simply as $\hat{\sigma}$ = 9.19/$\hat{b}$; see Equation 3) renders the distributions shown in the fourth column of Figures 4a and 4b, which are remarkably symmetric and centered on the true value of σ when N is sufficiently large ($N \geq 400$ suffices except, again, with QUEST). In other words, b is overestimated even when N = 1000 and the corrupting effect of improper estimates has been removed. Conversely, the distribution of $\hat{\sigma}$ is symmetric for all methods except the rowdy QUEST, and shows no evidence of bias. Although with small N the distributions of $\hat{\sigma}$ are skewed (see Figure 4a), this is more a consequence of N being too small than an underlying property of σ itself. Indeed, with small N the distributions of $\hat{\theta}$ are also skewed (see Figure 4a). Because σ is so much easier to interpret than b and its estimates also have much better properties, we will discontinue reporting results for $\hat{b}$. This is equivalent to redefining Equation 1 as

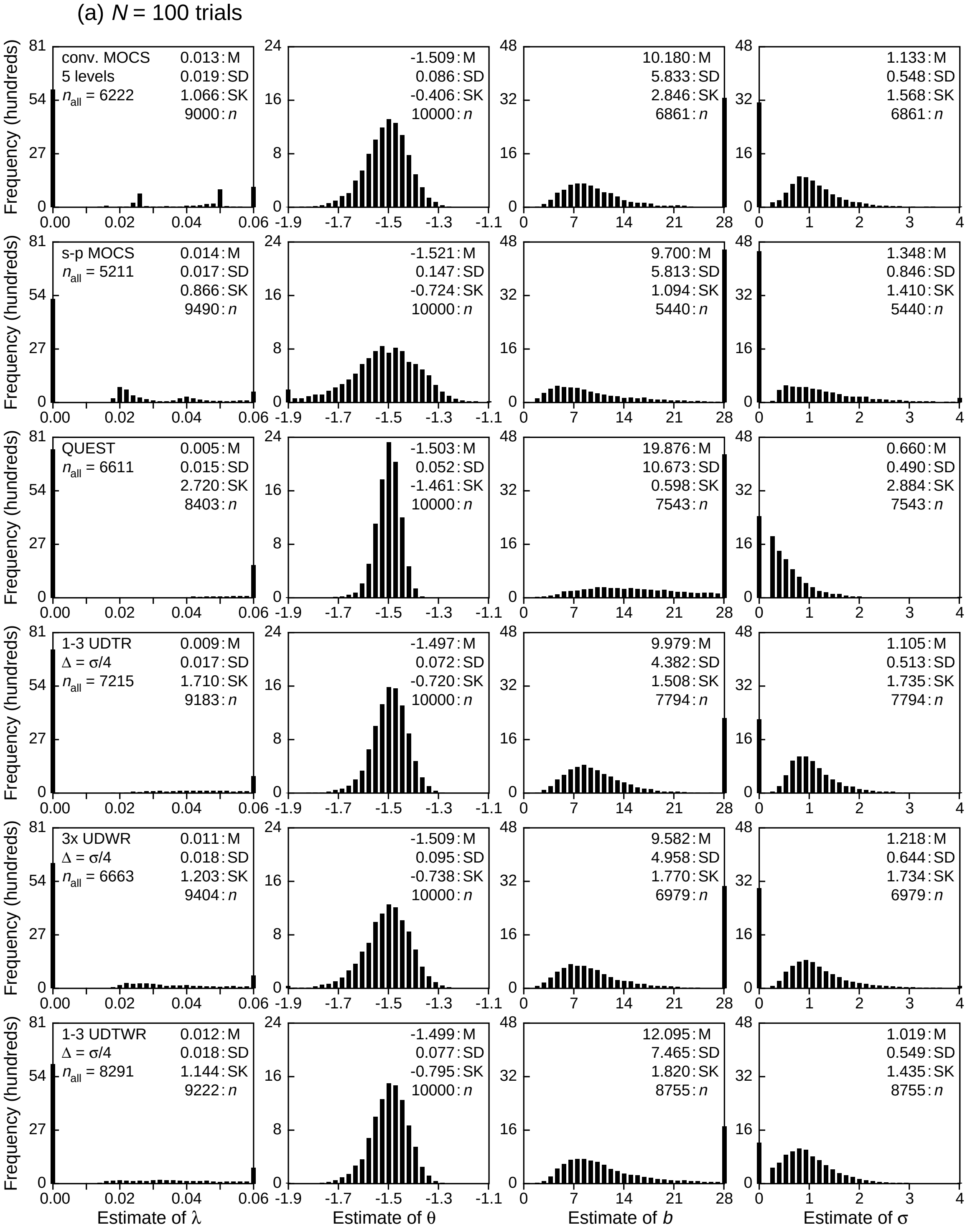

Figure 4. Sample histograms of ML estimates of λ (first column), θ (second column), *b* (third column), and σ (fourth column) from runs of 100 trials (a) and 1000 trials (b) in a 2AFC detection task. Each row pertains to one of the six sampling plans for which results from an individual run were shown in Figure 1 (see the labels inside the top left corner in the leftmost panel in each row). The generating Ψ is also the same as was used there, and the fitted $\breve{\Psi}$ was logistic too. To facilitate visual comparisons within each part, the horizontal and vertical ranges and scales are identical for all panels in each column, although the ranges vary from (a) to (b). The inset list down the top right corner of each panel gives the mean M, standard deviation SD, skewness SK, and sample size *n* (excluding improper estimates) of each distribution. The value of n_{all} in the leftmost panel in each row indicates the overall number of runs (out of 10,000) that yielded data allowing the estimation of all parameters.

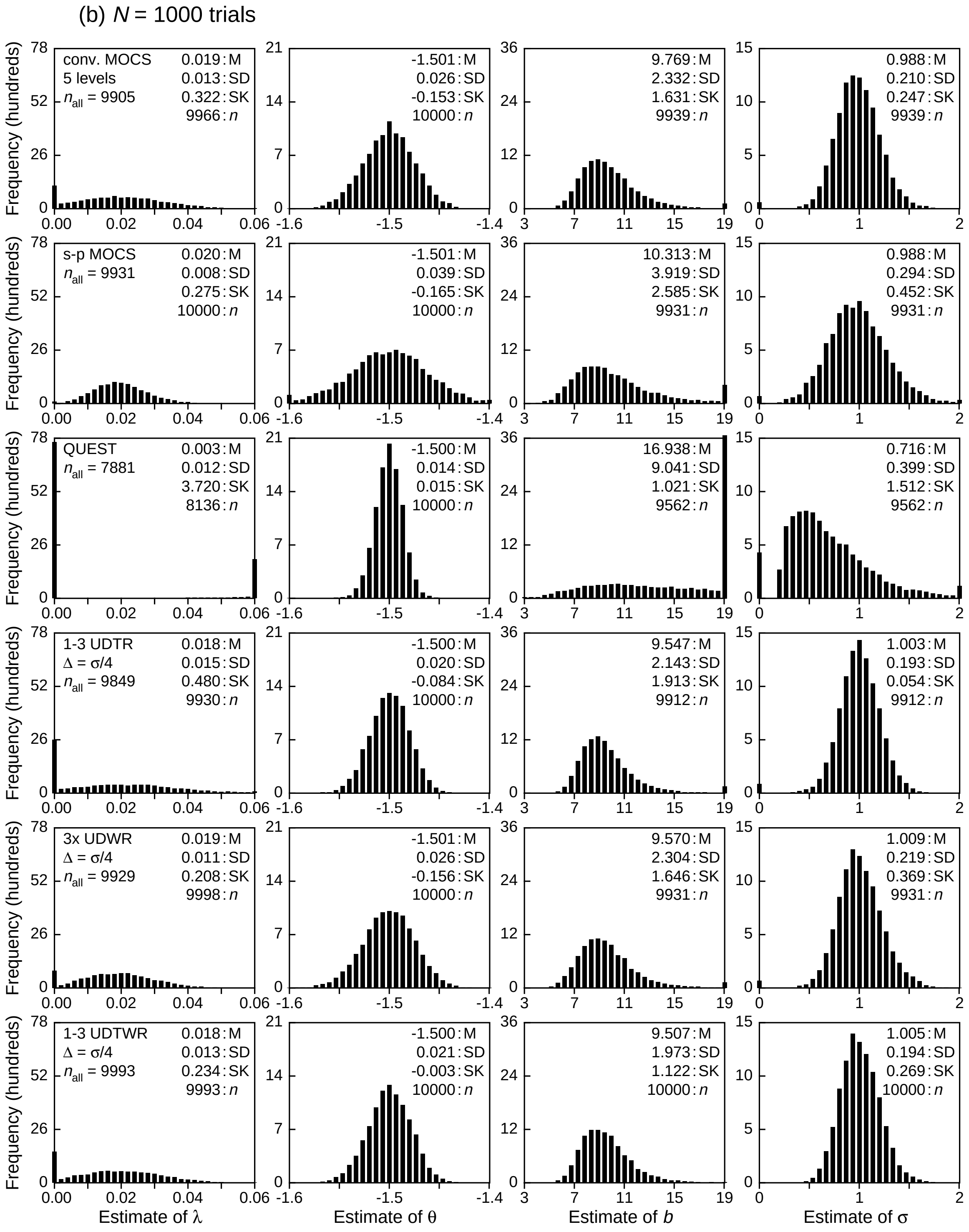

Figure 4 (continued). Sample histograms of ML estimates of λ (first column), θ (second column), b (third column), and σ (fourth column) from runs of 100 trials (a) and 1000 trials (b) in a 2AFC detection task. Each row pertains to one of the six sampling plans for which results from an individual run were shown in Figure 1 (see the labels inside the top left corner in the leftmost panel in each row). The generating Ψ is also the same as was used there, and the fitted $\check{\Psi}$ was logistic too. To facilitate visual comparisons within each part, the horizontal and vertical ranges and scales are identical for all panels in each column, although the ranges vary from (a) to (b). The inset list down the top right corner of each panel gives the mean M, standard deviation SD, skewness SK, and sample size n (excluding improper estimates) of each distribution. The value of n_{all} in the leftmost panel in each row indicates the overall number of runs (out of 10,000) that yielded data allowing the estimation of all parameters.

$$\Psi(x) = \gamma + \frac{1 - \lambda - \gamma}{1 + \exp[-9.19(x - T)/\sigma]}, \qquad (10)$$

An analogous replacement of β with 2.66/σ applies to the Weibull Ψ in Equation 4. We checked that direct estimates of σ from Equation 10 are identical to those obtained indirectly by estimating *b* in Equation 1 and then transforming it into an estimate of σ with Equation 3.

4.1. Usability Indices

A sampling plan must offer some guarantee that the data will yield proper estimates. Figure 5 shows how n_{all} (expressed as a percent) varies with *N* for each of our 98 plans. Some plans move towards 100% usability (albeit at different speeds) as *N* increases whereas other plans have a ceiling usability of 90–95%. QUEST has the lowest usability index (see its trace in the top left panel of Figure 5), implying that even if 1000 trials are given the probability is only 0.7881 that the data will yield proper estimates of all the parameters of Ψ.

Conventional MOCS in any of its variants and single-presentation MOCS (top left panel in Figure 5) require at least 500 trials to yield 95% usability, and their indices go hand in hand beyond 700 trials although none of these plans hits 100% with 1000 trials. At the other extreme, UDTWR staircases (right column in Figure 5) yield 95% usability with only 300 trials and reach almost 100% usability with 600 trials provided that σ/8 ≤ Δ ≤ σ/3. Between UDTR staircases (left column of Figure 5, excluding the top panel) and UDWR staircases (center column of Figure 5), the latter appear preferable in the 4× version, yielding 100% usability with 700 trials when Δ ≥ σ/5.

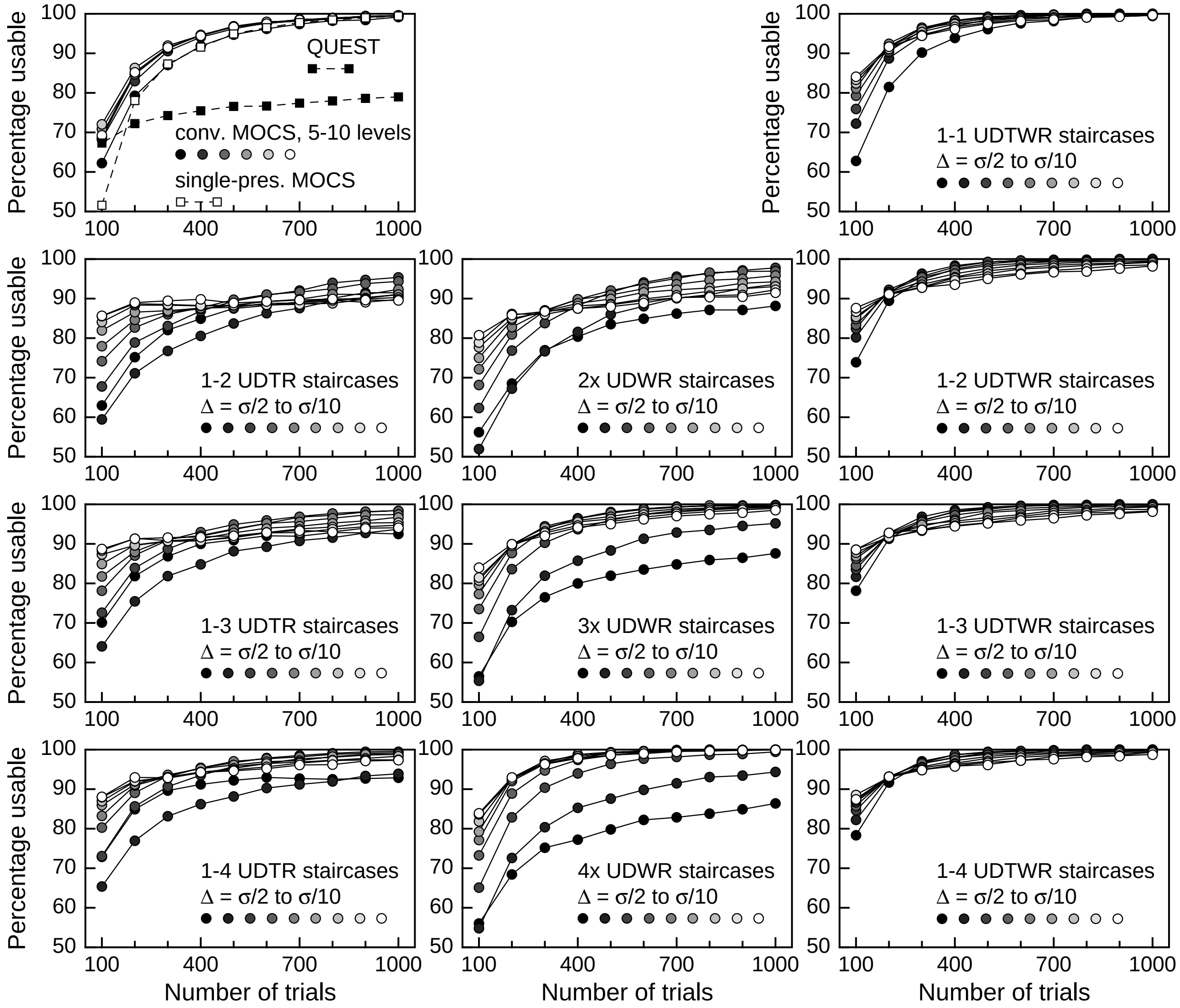

Figure 5. Percentage of runs yielding data that rendered proper ML estimates of all the parameters of Ψ, as a function of the number *N* of trials in 2AFC detection tasks. Generating Ψ and fitted $\breve{\Psi}$ were as in Figure 4. Each panel shows results for a subset of the sampling plans (see insets).

Since different sampling plans may be better suited to estimating different parameters, a look at usability indices for each parameter may be useful for identifying features that allow a plan to yield (or prevent it from yielding) proper estimates of a parameter. Figure 6 shows separate usability indices for the estimation of λ and σ (θ could be estimated 100% of the times in all conditions) as a function of N for QUEST, single-presentation MOCS, and all variants of conventional MOCS (first panel in Figures 6a and 6b) as well as for the best variants of UDTR, UDWR, and UDTWR staircases in terms of overall usability (second to fourth panels in Figures 6a and 6b).

It is clear that the low overall usability of QUEST is a result of its remarkable inability to produce proper estimates of λ being independent from its mild inability to produce proper estimates of σ. Because QUEST has low usability and only produces good estimates of θ, we will discontinue reporting results for it. In single-presentation MOCS, low overall usability is mostly determined by an inability to provide proper estimates of σ. In conventional MOCS, usability for λ and σ also seem independent from one another, thus explaining the lower overall index of usability; and there are little differences among variants of conventional MOCS in these respects. The case of staircases is more interesting because base spacings of different sizes serve competing goals. Large steps (darker circles) prevent obtaining proper estimates of σ but they allow obtaining proper estimates of λ, and the opposite holds for small steps (lighter circles). The ceiling effect observed in most of the panels of Figure 5 for UDTR staircases with small step sizes (light circles) is mostly a consequence of the inappropriateness of small steps for gathering data that will allow proper estimation of λ. Interestingly, UDTWR staircases (fourth panels in Figures 6a and 6b) are affected least by this trade-off, and σ and λ can both be estimated more than 95% of the times with 600 trials regardless of step size. The relative independence of usability on step size under UDTWR staircases was apparent also in the rightmost column of Figure 5.

4.2. Statistical Properties of Estimates From Conventional and Single-Presentation MOCS

The distributions of estimates from conventional MOCS did not vary with the number of levels (results not shown). Thus, when N is fixed (i.e., when more stimulus levels implies fewer trials per level), the number of levels in MOCS does not affect the quality of the estimates. Yet, increasing the number of levels appears to improve usability slightly (see the traces for MOCS in the top left panel of Figure 5). It thus looks as if under conventional MOCS it is only the overall number of trials (regardless of the number of levels) within the region of support of Ψ that matters. This fact is best noted in Figure 7, which shows means and standard deviations of $\hat{\lambda}$, $\hat{\theta}$, and $\hat{\sigma}$ as a function of N for the six variants of conventional MOCS and for single-presentation MOCS.

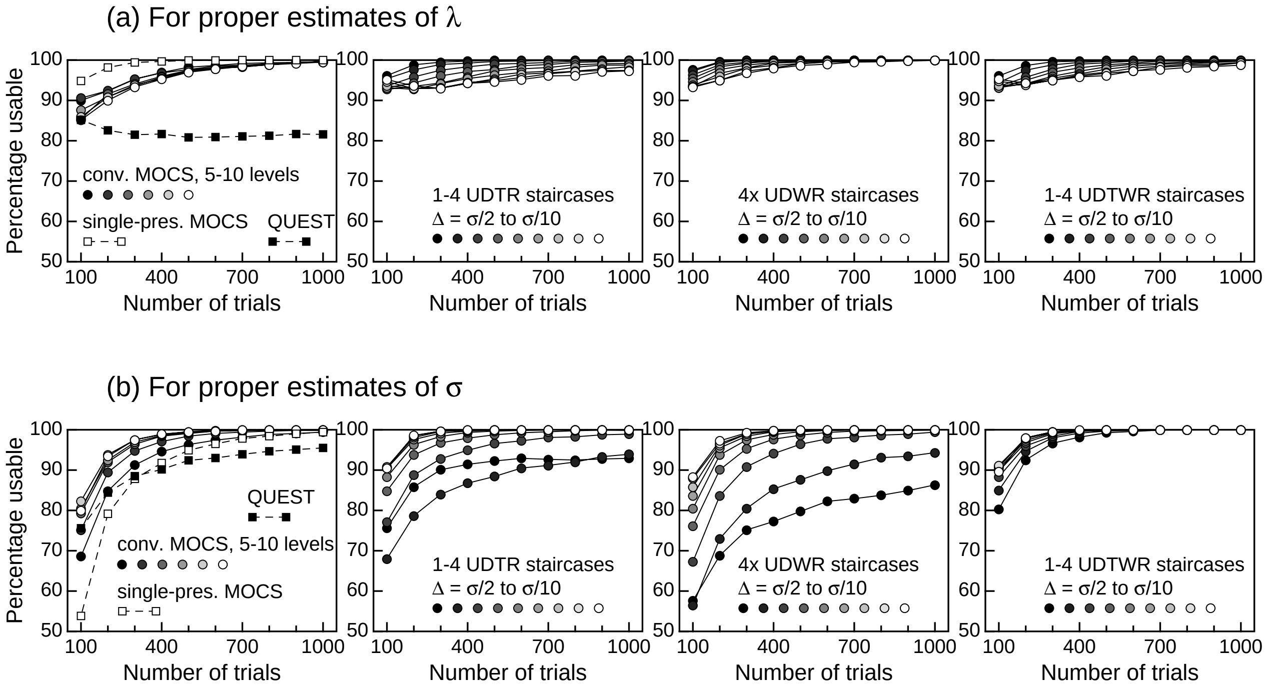

Figure 6. Percentage of runs yielding data that rendered proper ML estimates of λ (a) and σ (b) as a function of the number N of trials in 2AFC detection tasks. Generating Ψ and fitted $\breve{\Psi}$ were as in Figure 4. Each panel shows results for a subset of the sampling plans (see inset labels). Results for adaptive staircases implementing other rules were analogous.

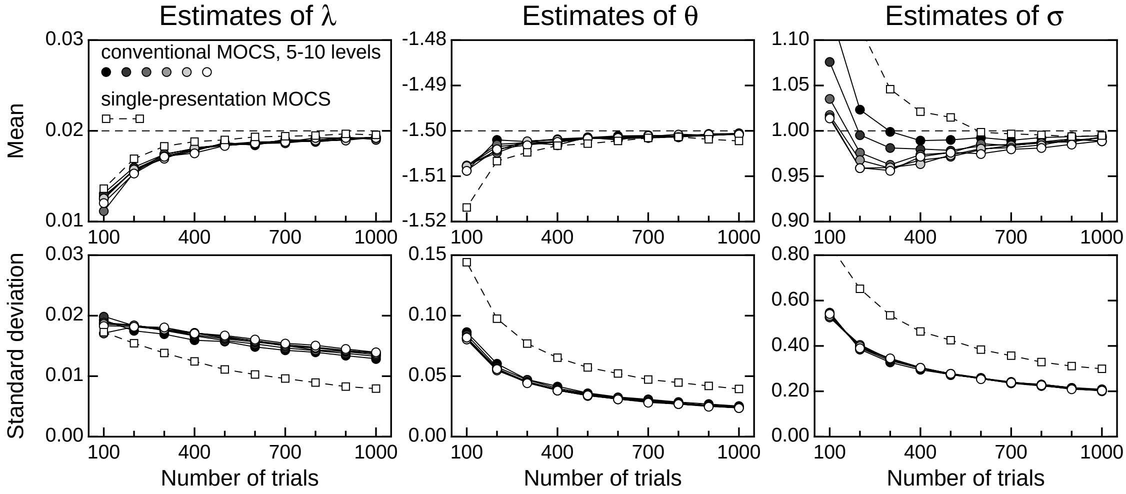

Figure 7. Mean (top) and standard deviation (bottom) of ML estimates of λ (left column), θ (center column), and σ (right column) arising from single-presentation MOCS (open squares in each panel) and six variants of conventional MOCS (circles in each panel) as a function of the overall number N of trials in 2AFC detection tasks. Generating Ψ and fitted $\check{\Psi}$ were as in Figure 4.

Considering means and standard deviations simultaneously, conventional MOCS is poor at estimating λ perhaps because stimulus levels are confined within the region of support of Ψ. Failing to test above this region precludes obtaining evidence of lapses and, thus, many replicates produced $\hat{\lambda} = 0$. Conversely, under single-presentation MOCS, the slightly lower underestimation of λ and the lower variability of its estimates arise from the absence of large numbers of zero estimates. But estimates of θ and σ are poorer under single-presentation MOCS: Their standard deviation is at least 50% larger than standard deviations from conventional MOCS involving the same N. And there are virtually no differences among variants of conventional MOCS as regards the properties of $\hat{\lambda}$, $\hat{\theta}$, and $\hat{\sigma}$ under the ideal placement of levels that we have used (for the effect of variations in the placement of levels, see Wichmann and Hill, 2001a).

4.3. Statistical Properties of Estimates From Adaptive Staircases

Distributions of estimates for our 27 variants of UDTR, 27 of UDWR, and 36 of UDTWR staircases were similar to those displayed in the bottom three rows of Figures 4a and 4b. Figure 8 summarizes the results in the same format as Figure 7 but only for the best rules in terms of properties of the estimates (1–3 UDTR, 3× UDWR, and 1–3 UDTWR staircases; and note from Figure 5 that these variants have usability indices that are respectively similar to those of 1–4 UDTR, 4× UDWR, and 1–4 UDTWR staircases). In general, the properties of all estimates deteriorated slightly as d (in 1–d UDTR and 1–d UDTWR staircases) or k (in k× UDWR staircases) moved away from 3.

Paralleling the implications of step size on usability indices, with all types of adaptive staircase small step sizes (lighter circles) yield poor estimates of λ with large variability (left column in Figure 8) and good estimates of θ and σ with small variability (center and right columns in Figure 8). Large step sizes (darker circles in Figure 8) yield the opposite outcomes.

Piecing it all together, small steps yield 100% usability for θ and σ, their estimates are accurate and have small variability, but the resultant data do not always allow estimating λ and, when they do, those estimates are poor and have large variability. Usability and the quality of estimates of θ and σ deteriorate with large step sizes, with which usability and the quality of estimates of λ improve. UDTWR staircases outperform the other types at balancing this trade-off.

4.4. Discussion

The optimal sampling plan would yield for all parameters the best possible picture of lack of bias and small variability for a given N. From the results presented thus far, none of the plans we analyzed performs optimally with all parameters. QUEST and single-presentation MOCS appear to trade off precision and accuracy too heavily in favor of one of the parameters. Conventional MOCS fails to estimate λ, and users will always have to decide where to place the stimuli thus being exposed to the consequences of an unfortunate decision. Finally, adaptive staircases always provide good estimates of θ and it looks like they can provide good estimates of either λ (when using large steps) or σ (when using small steps) but not both. An improved design that achieves both goals will be presented in Section 7.1.

In our study, spacing in conventional MOCS with L levels was $\sigma/(L - 1)$ whereas spacing in UDTR and UDWR staircases varied between $\sigma/10$ and $\sigma/2$. Then, for each $5 \le L \le 10$ in MOCS there is one set of UDTR and one of UDWR staircases

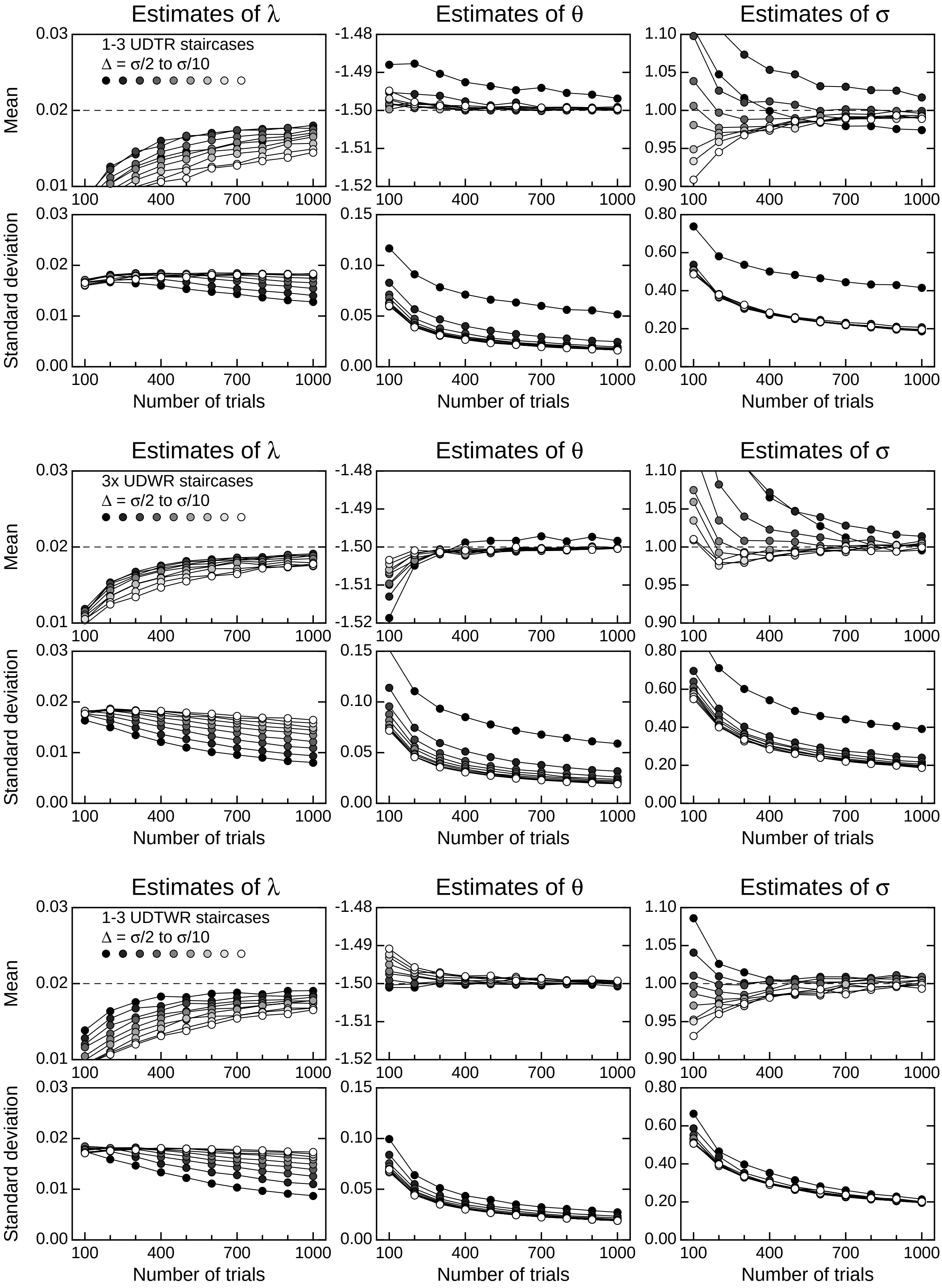

Figure 8. Similar to Figure 7, but for 1–3 UDTR staircases (upper part), 3× UDWR staircases (center part), and 1–3 UDTWR staircases (lower part).

that used the same spacing. It is worth looking at how adaptive placement compares with a fixed placement of levels within the region of support of Ψ with the same spacing. On this issue, McKee, Klein, and Teller (1985, p. 296, original italics) claimed that, with probit analysis, “the variability of [threshold] estimates derived from staircase data *can never be less* than the variability of estimates derived from the method of constant stimuli *selected for the optimal deployment of trials*.”

Figure 9 plots all the relevant data in a format that facilitates this comparison, and it can easily be noted that the statement of McKee et al. (1985) does not hold for ML estimation: The variability of ML estimates of θ from UDTR staircases is always smaller than that arising from comparable MOCS (center panel in the top row of Figure 9), as is the variability from UDWR staircases using fine spacing (lighter symbols in the center panel in the center row). UDTR staircases yield estimates of λ and σ that have about the same variability as those arising from MOCS (left and right panels in the top row of Figure 9), whereas UDWR staircases yield estimates of λ with generally smaller variability and estimates of σ with larger variability than those arising from MOCS (left and right panels in the center row of Figure 9). And there is also the issue that setting up MOCS for “the optimal deployment of trials” is impossible in empirical practice without incurring extra costs. Note also in the bottom row of Figure 9 that UDTWR staircases, which used the same base spacings but rendered a different sampling lattice (see Figure 1f), yield estimates whose variability is related to that of MOCS estimates according to a pattern that is very similar to that shown in the top row of Figure 9 for UDTR staircases.

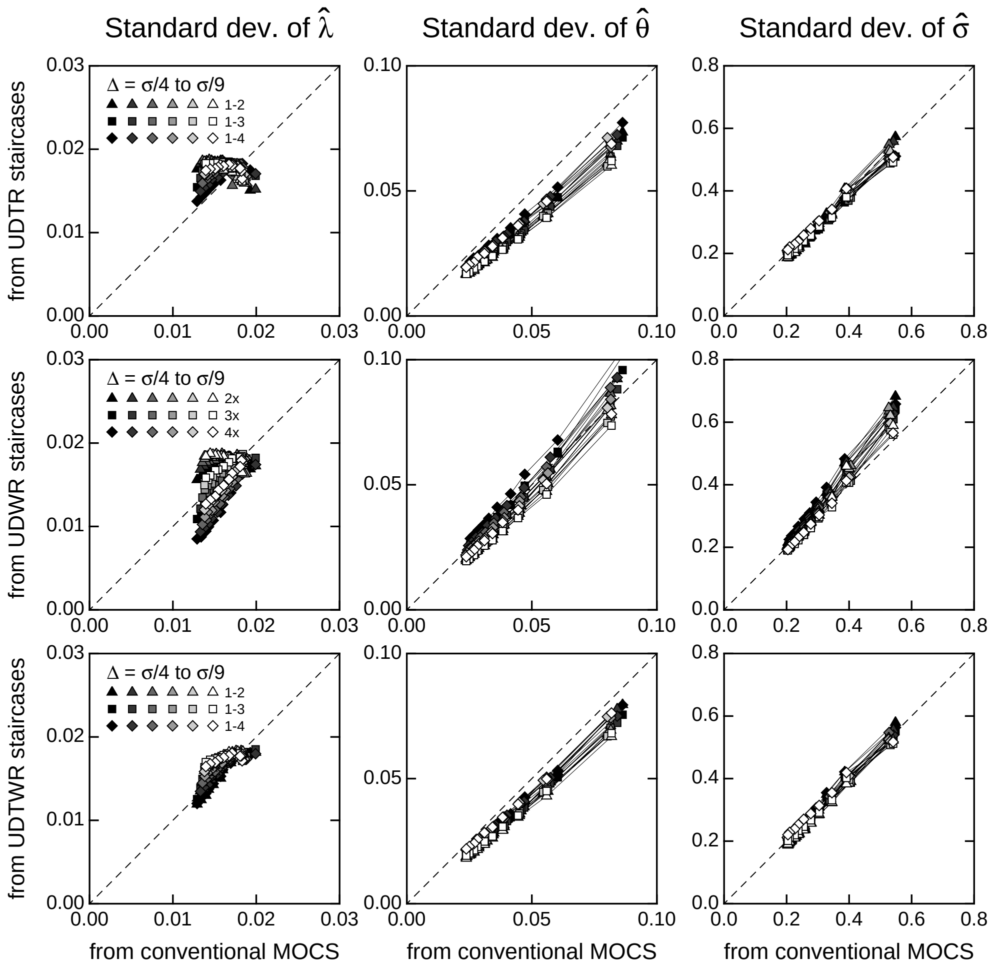

Figure 9. Relationship between the variability of the estimates of λ (left colunn), θ (center column), and σ (right column) obtained with conventional MOCS (abscissa in each panel) and UDTR staircases (ordinate in the top row), UDWR staircases (ordinate in the center row) or UDTWR staircases (ordinate in the bottom row) in 2AFC detection tasks. For reference, the dashed diagonal is the identity line. Gray shading signals spacing between levels and symbol type signals up–down rule (see the inset legend in the leftmost panel in each row); strings of identical symbols with the same shading are connected by thin lines and represent different numbers *N* of trials (unmarked). Results for 1–1 UDTWR staircases are omitted to avoid excessive clutter but the data points were slightly below, above, and above the diagonal respectively in the left, center, and right panels in the bottom row.

5. Results: II. Yes–No Tasks

Sections 5.1–5.3 next describe results for all 125 plans included in our study with yes–no tasks, where γ becomes a free parameter. The first eight sampling plans (six variants of conventional MOCS, one of single-presentation MOCS, and one of QUEST) are analogous to those used with 2AFC detection tasks. The remaining plans comprise 63 variants of UDTR staircases (1–1, 1–2, 1–3, 1–4, 2–1, 3–1, and 4–1 rules each with nine base spacings) and 54 variants of UDWR staircases (2×, 3×, 4×, $\frac{1}{2}$×, $\frac{1}{3}$×, and $\frac{1}{4}$× rules each with nine base spacings). Again for simplicity, we will only report results for logistic Ψ with $\gamma = 0.02$, $\lambda = 0.02$, $\theta = -1.5$, and $\sigma = 1$, logistic $\breve{\Psi}$, and ML estimates. Other results will be summarized in Section 6.

5.1. Usability Indices

Usability (Figure 10a) was lower than for 2AFC detection tasks, undoubtedly because the need to estimate γ provides another independent chance to fail to estimate one parameter. This outcome reveals a failure at placing trials where it is needed to estimate all parameters. The only exception is single-presentation MOCS, which indeed places trials below, within, and above the region of support of Ψ,[4] yielding 100% usability in only 600 trials. We will not report any further results for QUEST because it hardly ever hit 60% usability and always yielded poor estimates of γ, λ, and σ. Finally, all staircases fare worse than conventional MOCS (compare the strings of circles in the top panel of Figure 10a with the strings in the remaining panels down the column). An interesting difference with respect to usability indices in 2AFC detection tasks is that increasing the number of levels in conventional MOCS in yes–no tasks reduces usability.

The problem with adaptive staircases seems to lie in that proper estimates of λ and γ cannot be obtained simultaneously (see Figures 10b and 10c) since each requires trials placed either above or below the region of support of Ψ. In contrast, conventional and single-presentation MOCS do not seem to face any more problems estimating γ than λ (see the top panels in Figures 10b and 10c). Yet, different broad types of either UDTR or UDWR staircases seem well suited to estimating each of these two parameters. In particular, 1–*d* UDTR and *k*× UDWR staircases (see examples in the third and fifth rows in Figure 10) provide proper estimates of λ because they use an upper starting point and they tend to place trials above the midpoint of the region of support of Ψ. Their drawback is that they do not sample sufficiently below this region, which would be required to obtain proper estimates of γ. Conversely, *u*–1 UDTR and $\frac{1}{k}$× UDWR staircases (see examples in the fourth and bottom rows in Figure 10), which use a lower starting point and tend to place trials below the midpoint of the region of support of Ψ, provide proper estimates of γ but fail to sample the region that would allow obtaining proper estimates of λ.

Thus, 1–*d* and *u*–1 UDTR staircases merely differ in that their patterns of usability for γ and λ are interchanged, but they both yield the same patterns for σ and overall (compare the third and fourth rows in Figure 10); *k*× and $\frac{1}{k}$× UDWR staircases differ in the same respects (compare the bottom two rows in Figure 10). And note that UDWR staircases slightly outperform UDTR staircases, yielding higher usability with any given N and Δ. Finally, 1–1 UDTR staircases with large steps are only slightly less efficacious than conventional MOCS in any variant (compare the darker circles in the second row in Figure 10 with the traces for conventional MOCS in the top row). The slightly higher usability of 1–1 UDTR staircases for λ than for γ (compare the second panels in Figures 10b and 10c) arises because these staircases used an upper starting point and, then, placed a few more trials above than below the region of support of Ψ.

5.2. Statistical Properties of Estimates From Conventional and Single-Presentation MOCS

Figure 11 shows histograms of estimates of γ, λ, θ, and σ (left to right) with 1000 trials from 5-level conventional MOCS (first row), 10-level conventional MOCS (second row), and single-presentation MOCS (third row). Single-presentation MOCS outperforms conventional MOCS at estimating γ and λ, but estimates of θ and σ are slightly poorer in that their standard deviation is larger. There appears to be no difference between the distributions arising from 5-level and 10-level conventional MOCS, nor did any of these differ from those obtained with intermediate numbers of levels. When N = 100 (results not shown), γ and λ were not estimated well with any method, but θ and σ were well estimated except by single-presentation MOCS.

Figure 12 plots means and standard deviations of the estimates of γ, λ, θ, and σ (left to right) for all variants of conventional MOCS and for single-presentation MOCS, as a function of N. Similarly to results shown in Figure 7 for 2AFC detection tasks, estimates of λ (and here also γ) from single-presentation MOCS are slightly more accurate and have smaller variability than those obtained from conventional MOCS; however, estimates of θ and σ from single-presentation MOCS have larger variability than those arising from conventional MOCS. Note also by comparison with Figure 7 that the standard deviations of $\hat{\theta}$ and $\hat{\sigma}$ are generally much smaller here, something that has already been reported for $\hat{\theta}$ (McKee et al., 1985). And note again that the number of levels in conventional MOCS does not seem to make any difference when N is fixed.

[4] Of course, this is true only when the region of support of Ψ is narrow and centered with the range of exploration. See Section 6.3 for a broader picture.

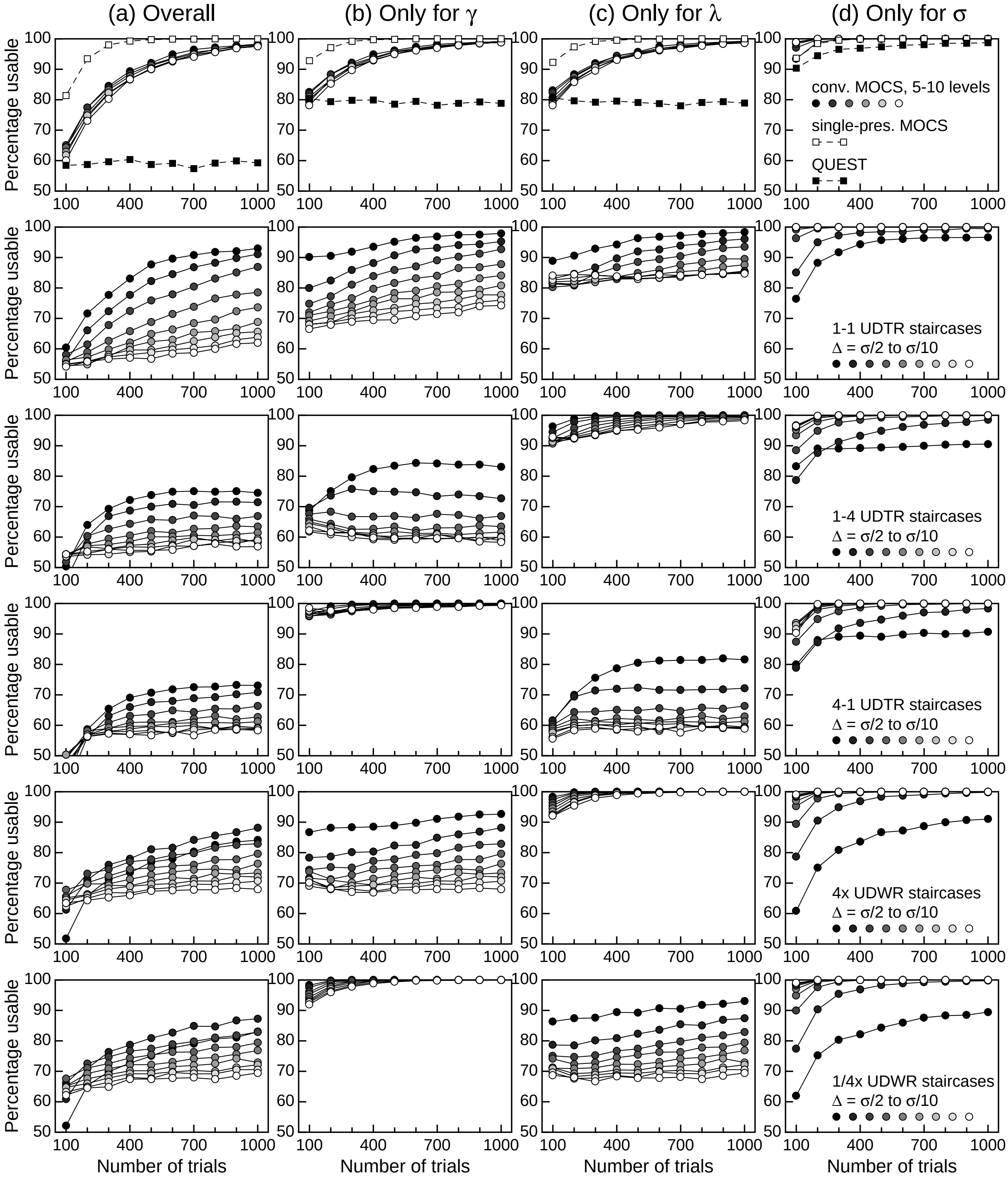

Figure 10. Overall usability indices (a) and usability for γ (b), λ (c), and σ (d) as a function of number *N* of trials in yes–no tasks for conventional MOCS, single-presentation MOCS, and QUEST (first row), 1–1 UDTR staircases (second row), 1–4 UDTR staircases (third row), 4–1 UDTR staircases (fourth row), 4× UDWR staircases (fifth row), and $\frac{1}{4}$× UDWR staircases (sixth row). The generating Ψ was logistic with γ = 0.02, λ = 0.02, θ = –1.5, and σ = 1 (*b* = 9.19), the fitted $\breve{\Psi}$ was logistic too, and usability refers to ML estimation.

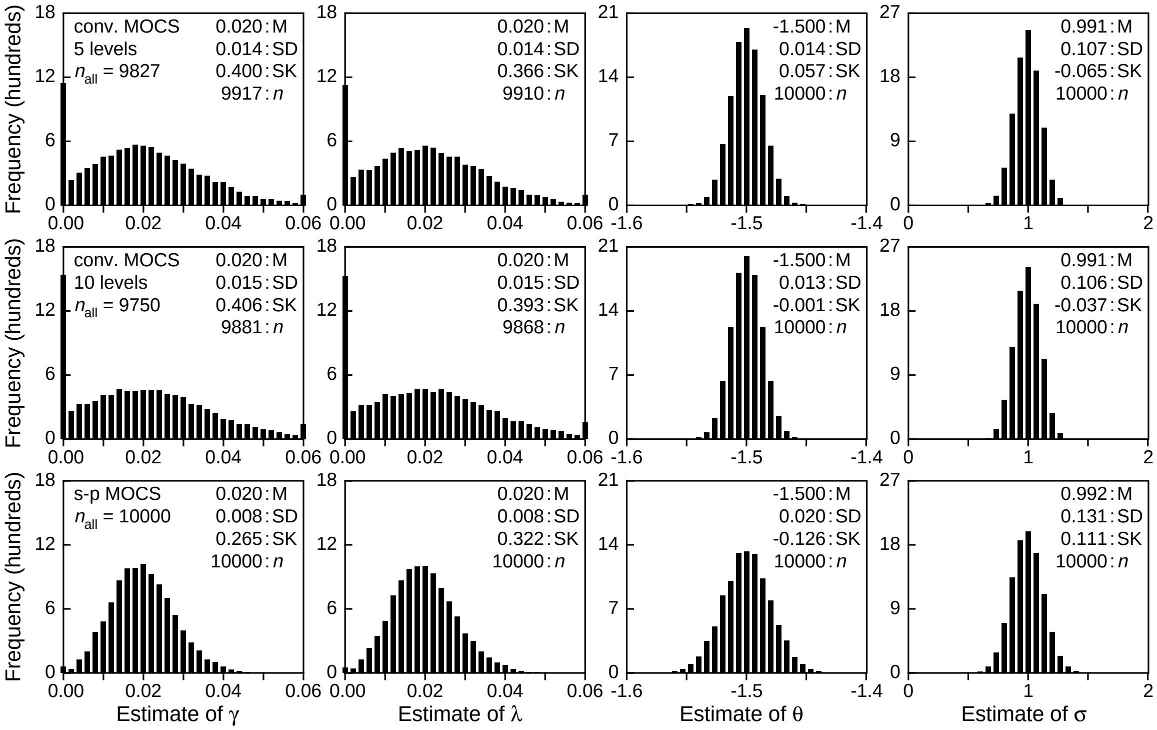

Figure 11. Distributions of ML estimates of γ (first column), λ (second column), θ (third column), and σ (fourth column) from runs of N = 1000 trials arising from 5-level conventional MOCS (top row), 10-level conventional MOCS (center row), and single-presentation MOCS (bottom row) in yes–no tasks. Generating Ψ and fitted $\breve{\Psi}$ were as in Figure 10. The inset list down the top right corner in each panel gives the mean M, standard deviation SD, skewness SK, and sample size n (excluding improper estimates) of each distribution. The value of n_{all} in the leftmost panel in each row indicates the overall number of runs (out of 10,000) that yielded data allowing the estimation of all parameters.

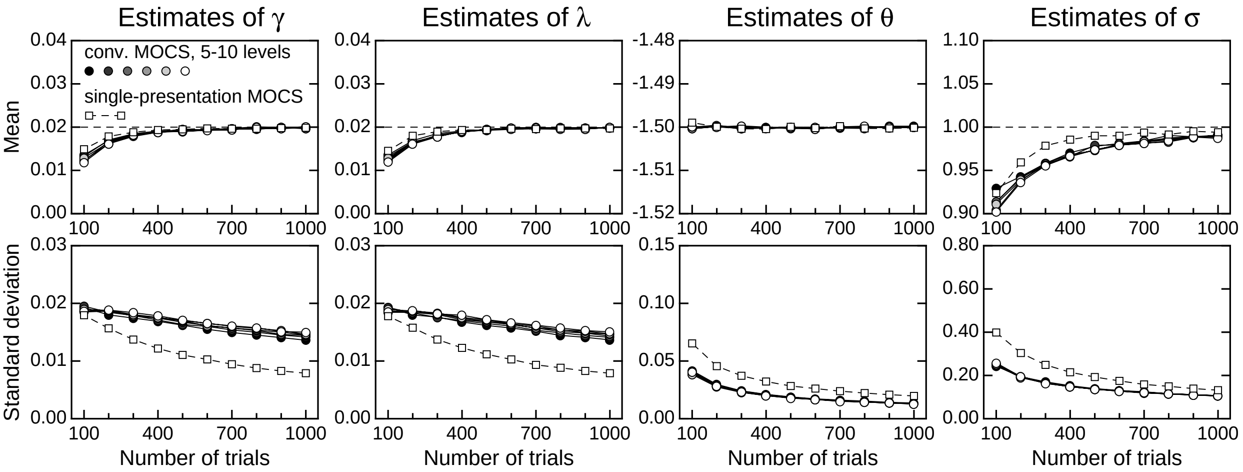

Figure 12. Mean (top row) and standard deviation (bottom row) of ML estimates of γ (first column), λ (second column), θ (third column), and σ (fourth column) arising from single-presentation MOCS (open squares in each panel) and six variants of conventional MOCS (circles in each panel) as a function of the overall number N of trials in yes–no tasks. Generating Ψ and fitted $\breve{\Psi}$ were as in Figure 10.

5.3. Statistical Properties of Estimates from Adaptive Staircases

Histograms of estimates from adaptive staircases had the same aspect as those in Figure 11, and their presentation is omitted. Figure 13 shows means and standard deviations of estimates of γ, λ, θ, and σ (left to right) as a function of N for 1–1 (upper part), 1–4 (center part), and 4–1 (lower part) UDTR staircases; Figure 14 does the same for 4× (upper part) and $\frac{1}{4}$× (lower part) UDWR staircases. Differences between 1–4 and 4–1 UDTR staircases (center and lower parts in Figure 13) and between 4× and $\frac{1}{4}$× UDWR staircases (Figure 14) again lie only in that their patterns of results for γ and λ are interchanged and that misestimation of θ has opposite sign. In general, UDWR staircases do better overall than comparable UDTR staircases. Finally, 1–1 UDTR staircases (upper part in Figure 13) seem superior in that θ and σ are estimated still better and the properties of $\hat{\gamma}$ and $\hat{\lambda}$ are not as imbalanced as they are with other UDTR staircases or with UDWR staircases. A comparison with Figure 8 (for analogous results in 2AFC detection tasks) also reveals that the standard deviations of $\hat{\theta}$ and $\hat{\sigma}$ are much smaller here (see also Kershaw, 1985).

5.4. Discussion

Again, no single sampling plan outperforms the rest in providing better estimates of all four parameters of interest in yes–no tasks. The ideal plan would have the usability indices of single-presentation MOCS (top panel in Figure 10a), but the estimates of γ and λ should have the properties attained with UDWR staircases using large steps (darker circles in the first column in the lower part and in the second column in the upper part of Figure 14), the estimates of θ should have the properties attained with conventional MOCS or 1–1 UDTR staircases (third column in Figure 12 and in the upper part of Figure 13), and the estimates of σ should have the properties attained with 4× UDWR staircases (fourth column in the upper part of Figure 14). A sampling plan that brings together most of these properties will be presented in Section 7.2.

With respect to the variability of estimates from adaptive methods as compared to that from conventional MOCS, in the case of yes–no tasks the comparison yields different results (not shown) for different up–down rules. Only 1–1, 1–2, and 2–1 UDTR staircases render smaller variability for $\hat{\theta}$ than comparable conventional MOCS, but 1–3, 1–4, 3–1, and 4–1 UDTR staircases render larger variability. And only 2× and $\frac{1}{2}$× UDWR staircases render smaller variability for $\hat{\theta}$ than MOCS, whereas 3× and $\frac{1}{3}$× UDWR staircases yield about the same variability as MOCS and 4× and $\frac{1}{4}$× UDWR staircases render more variability. The variability of $\hat{\sigma}$ is larger (and that of $\hat{\gamma}$ and $\hat{\lambda}$ smaller) from adaptive staircases.

6. Results: III. Miscellanea

The results presented in Sections 4 and 5 covered only a small subset of all the conditions described in Sections 3.1–3.5. Results in the rest of the conditions are described next.

6.1. Other Constraints on γ and λ: 2AFC Discrimination Tasks

In 2AFC discrimination tasks, each interval shows a detectable stimulus and the subject must respond which one has, say, higher contrast, guessing at random when uncertain. One of the stimuli is a standard of fixed contrast (say, S), whereas the other is a comparison whose contrast x varies across trials and may be below or above S. Then, Ψ describes the probability that a comparison at x is picked as the stimulus with the higher contrast. Despite the 2AFC format, Ψ ranges from near zero (because the comparison will rarely be perceived as having higher contrast than the standard when x is sufficiently below S) to near one (for an analogous reason). The point θ at which the probability of picking the comparison is 0.5 is called *point of subjective equality*. Moreover, γ is absent and its place in the functional expression of Ψ is taken by λ because lapses may occur with the same probability when x is either well above or well below S. Hence, in 2AFC discrimination tasks Ψ has a range similar to that in yes–no tasks, with the important difference that the lower and upper asymptotes are equidistant from 0 and 1, respectively. Two further consequences are that there are only three parameters to estimate and that Ψ must be odd symmetric about its inflection point (i.e., Ψ cannot be a Weibull function).

Simulations involving all the sampling plans in Section 5 (for yes–no tasks) and differing only in that the logistic Ψ was appropriately modified (so that γ is replaced with λ) rendered results that only differed noticeably in that λ could be estimated more often (which boosted usability indices) and better, undoubtedly because trials placed above or below the region of support of Ψ contribute jointly to the estimation of λ. As for the remaining parameters, the characteristics of the estimates of θ and σ were generally analogous to those reported in Figures 12–14, except that the estimates of σ were slightly better here with all variants of conventional MOCS and the estimates of θ and σ were also slightly better here with k× and $\frac{1}{k}$× UDWR staircases.

6.2. 3AFC and 4AFC Detection Tasks

In 3AFC and 4AFC detection tasks, the chance level γ gets closer to its values in yes–no tasks (although γ is still a constant here). Simulations involving all the sampling plans in Section 4 (except UDTWR staircases), and differing only in that $\gamma = 1/m$ and $\pi = (m + 1)/2m$ with $m = 3$ or 4 (see Section 3.1), rendered analogous results but with the differences described next.

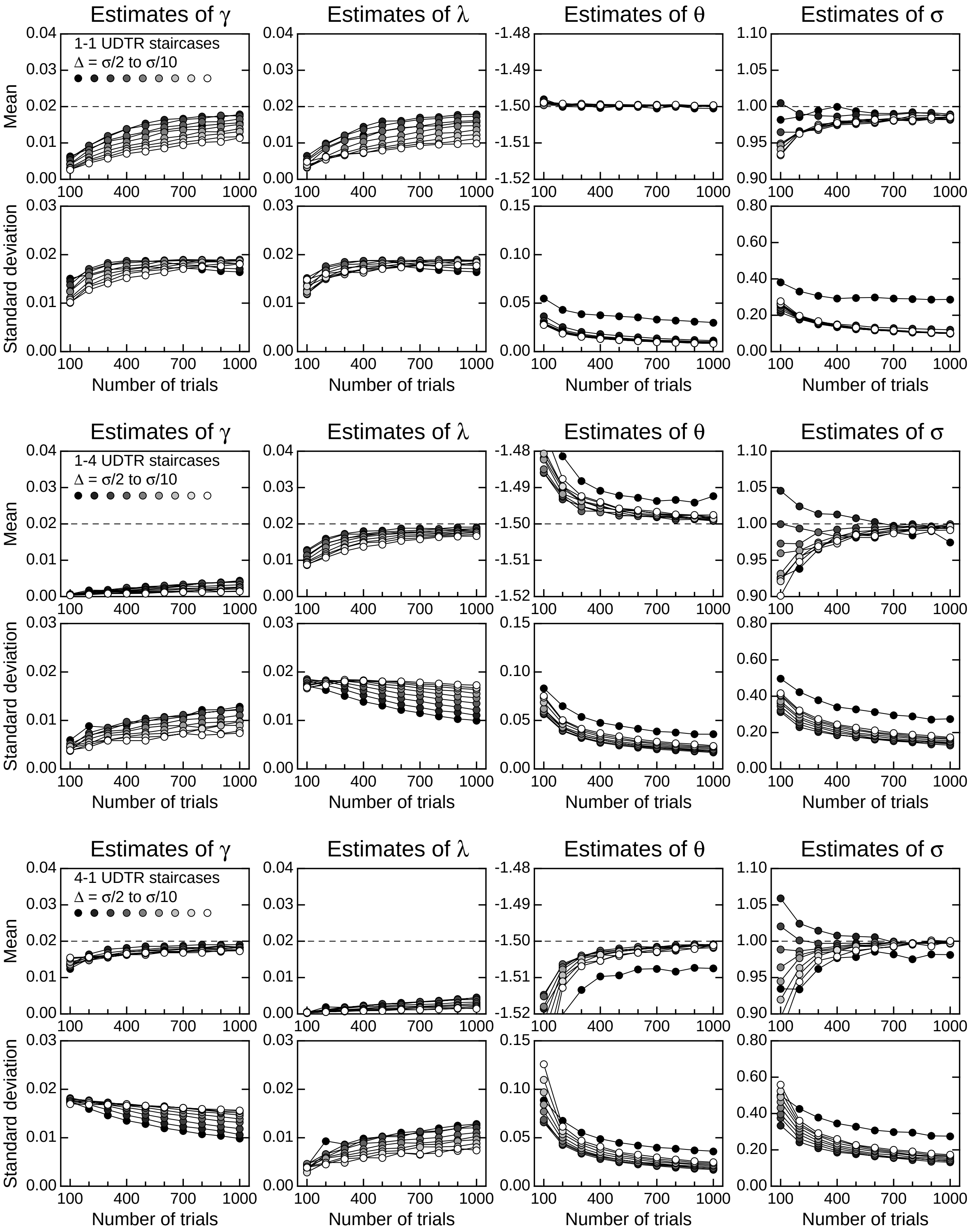

Figure 13. Similar to Figure 12, but for 1–1 UDTR staircases (upper part), 1–4 UDTR staircases (center part), and 4–1 UDTR staircases (lower part).

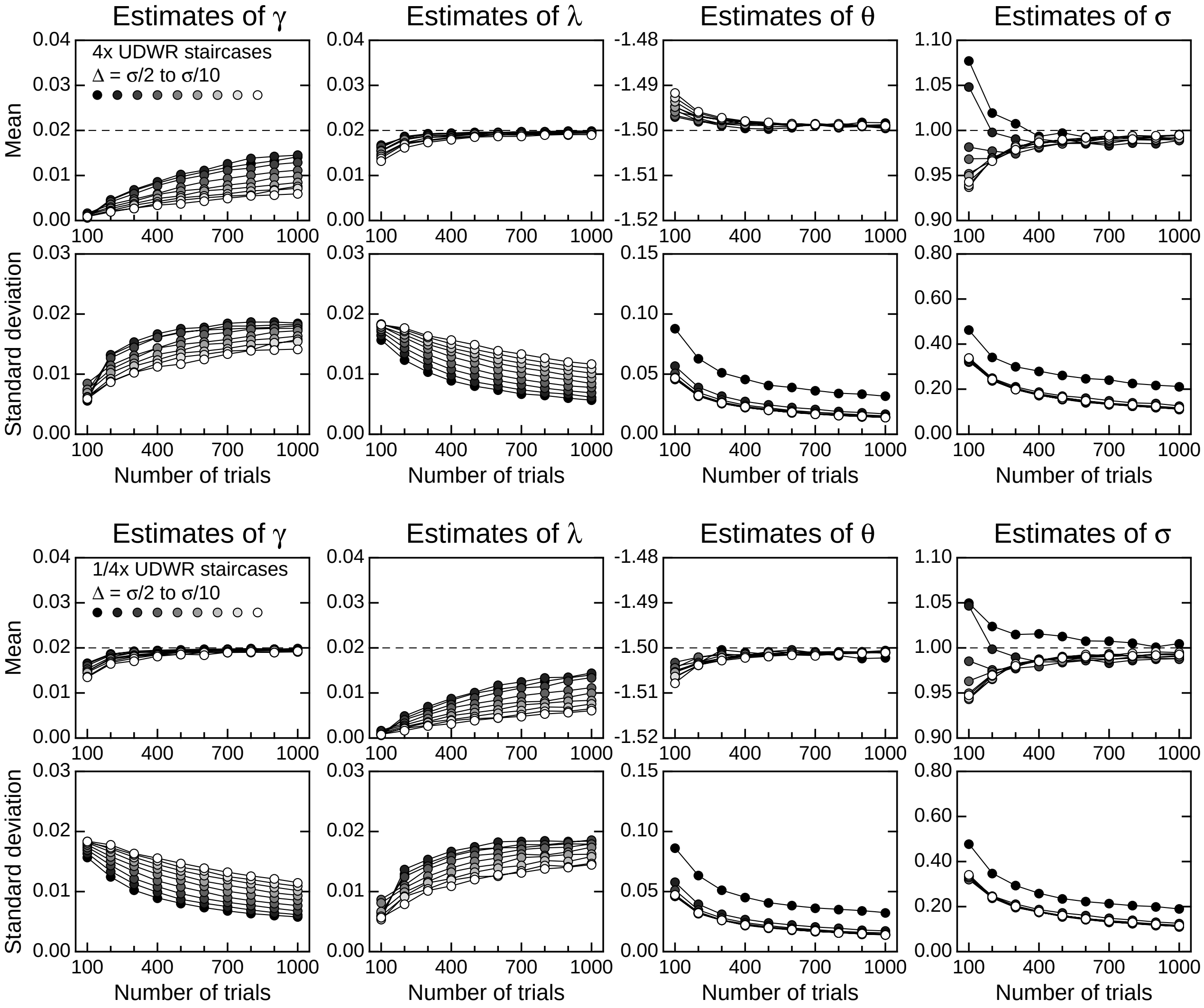

Figure 14. Similar to Figure 12, but for 4× UDWR staircases (upper part) and $\frac{1}{4}$× UDWR staircases (lower part).

Usability increased slightly with all plans as *m* increased. The overall usability of single-presentation MOCS, which lay below that of conventional MOCS in 2AFC detection tasks (top left panel in Figure 5), turned out to be about the same as that of conventional MOCS in 3AFC detection tasks, and slightly surpassed that of conventional MOCS in 4AFC detection tasks. This result reflects a tendency towards higher usability of single-presentation MOCS compared to conventional MOCS as γ approaches zero, the most extreme case of which can be observed in the top panel of Figure 10a for yes–no tasks with γ = 0.02. In addition, the variability of the estimates of all parameters decreased for all sampling plans as *m* increased, again reflecting a progression from the picture of variability shown in Figures 7 and 8 (where γ = 0.5) towards that shown in Figures 12–14 (where γ = 0.02) as γ approaches zero. These progressions are simply caused by the fact that γ gets closer to zero, whether or not γ itself is a free parameter.

6.3. *Other Values for the Parameters of Ψ*

Very few of the plans showed any traces of rendering different results when the true values of γ (only in yes–no tasks), λ, θ, or σ changed within the ranges expressed in Section 3.1. Exceptions for the better occurred when λ = 0 (and also when γ = 0, where applicable), because this precluded lapses (or false positives) and, thus, improper estimates of λ (and γ where applicable). Exceptions for the worse occurred when the values of θ and σ were such that the region of support of Ψ was not entirely contained within the available range of stimulus levels. The deterioration in these cases is as naturally expected as it is uninteresting, for it only reveals that the estimation problem is ill posed. The differences under other circumstances, if any, with respect to the results presented in Sections 4 and 5 are summarized next, one sampling plan at a time.

Because the range and spacing of levels in our implementation of conventional MOCS depends on the true values of θ and σ (see Section 3.4), varying the latter did not have any effect as long as θ and σ were such that the region of support of Ψ was within $[x_l, x_u]$. This raises the question of how a practitioner can figure out an appropriate spacing and positioning under uncertainty as to the values of θ and σ. A discussion of this problem is deferred to Section 8.3.

Single-presentation MOCS spreads stimulus levels evenly over $[x_l, x_u]$. Its performance deteriorated when the region of support of Ψ was too close to the boundary of this range but also when σ was small compared to $x_u - x_l$. A virtue of single-presentation MOCS is that it provides good estimates of γ and λ by placing trials above and below the region of support of Ψ, a characteristic that disappears if this region is too close to the boundaries of the range of stimulus levels. That is, single-presentation MOCS requires that Ψ be more interior within the range of stimulus levels than it is allowed to be under conventional MOCS. On the other hand, when σ is too small, very few trials are placed within the region of support of Ψ, which impairs the estimation of σ.

Because step sizes for our adaptive staircases depend on the value of σ (see Section 3.4), variations in σ cannot affect our results in any way. On the other hand, the location of Ψ, as determined by the value of θ, does not affect the implementation of adaptive staircases, which always started near one of the boundaries of the stimulus range and moved towards the appropriate region adaptively. Thus, when θ was far from starting point and σ was not too large, a few more trials were used to reach the region of support of Ψ. As a result, estimates of λ were a little better and estimates of θ and σ did not suffer much. Conversely, when θ was close to starting point and σ was again not too large, estimates of λ were a little worse and estimates of θ and σ did not improve significantly. Again, from the practitioner's point of view, staircases pose a problem similar to that posed by MOCS, although here the only decision that has to be made right concerns the size of the steps. We also defer a discussion of this issue to Section 8.3.

6.4. Other Forms for Ψ, and Mismatch Between Actual Ψ and Model $\breve{\Psi}$

When actual Ψ and fitted $\breve{\Psi}$ were both Weibull instead of logistic, the results did not show any noticeable difference except that conventional MOCS performed a little worse and adaptive staircases performed a little better in terms of bias and variability. Yet, the mismatch between Ψ and $\breve{\Psi}$ greatly affected the quality of the estimates and the deterioration did not change whether Ψ was logistic and $\breve{\Psi}$ was Weibull or vice versa. In all cases, γ (when applicable), λ, θ, and σ were substantially misestimated. Figure 15 illustrates with representative results from 1–4 UDTWR staircases in 2AFC detection tasks. In comparison with results for logistic Ψ and $\breve{\Psi}$ (lower part of Figure 8), all parameters were overestimated when Ψ was logistic and $\breve{\Psi}$ was Weibull (Figure 15a) and they were underestimated when Ψ was Weibull and $\breve{\Psi}$ was logistic (Figure 15b).

Interestingly, despite the deterioration in the quality of the estimates, the relative performance of the different sampling plans considered in this study remained identical. In addition, the relative performance of variants of these sampling plans (where applicable) also remained identical. For instance, estimates of θ can be seen in the second columns of Figures 15a and 15b to be unbiased and have smaller standard errors for $N > 500$ when $\Delta = \sigma/10$ (open circles), and this was also the optimal spacing for estimation of θ when Ψ and $\breve{\Psi}$ were both logistic (see the lower part of Figure 8). A similar analysis for the remaining parameters reveals comparable relative outcomes across values of Δ in Figure 15 and in the lower part of Figure 8.

Because the actual mathematical form of Ψ is never known, practitioners should realize that there is always a potential for bias in estimates of λ, γ (when applicable), and σ, a bias whose presence can never be assessed because it depends on the similarity of the actual Ψ to the model function $\breve{\Psi}$. The results in Figure 15 indicate that, using the optimal spacing Δ and a sufficiently large number N of trials, estimates of λ and γ will be biased by about ±25% of their actual value, whereas estimates of σ will be biased by about ±10% of its actual value. The reason for this bias lies in the failure of the mathematical form of $\breve{\Psi}$ to accommodate data describing a shape generated by Ψ and further contaminated by random noise, but a thorough analysis of this characteristic is beyond the scope of this paper.

6.5. Other Parameter Estimation Methods

Despite an occasional better behavior of OLS methods (see Figure 3), usability indices for OLS estimation were generally lower than for ML estimation, and the global properties of OLS estimates were often noticeably worse than (occasionally similar to) those of ML estimates. In particular, in 2AFC detection tasks, conventional and single-presentation MOCS render very similar OLS and ML estimates of all three parameters. UDTR and UDWR staircases yield OLS underestimates of λ, and OLS estimates of θ and σ that are similar to their ML counterparts. UDTWR staircases, finally, yield OLS overestimates of λ and θ, and OLS underestimates of σ. In yes–no tasks, on the other hand, OLS and ML estimates had similar distributions with the only exceptions that (1) λ and γ are underestimated with UDTR and UDWR staircases and (2) estimates of σ from all UDWR staircases and from UDTR staircases not using the 1–1 rule have more variability.

The generally inferior performance of OLS estimation may simply reveal its theoretical inappropriateness when the errors are not normally distributed and the variances are not homogeneous across stimulus levels, as was pointed out in Section 3.5. We also discussed there that the appropriate

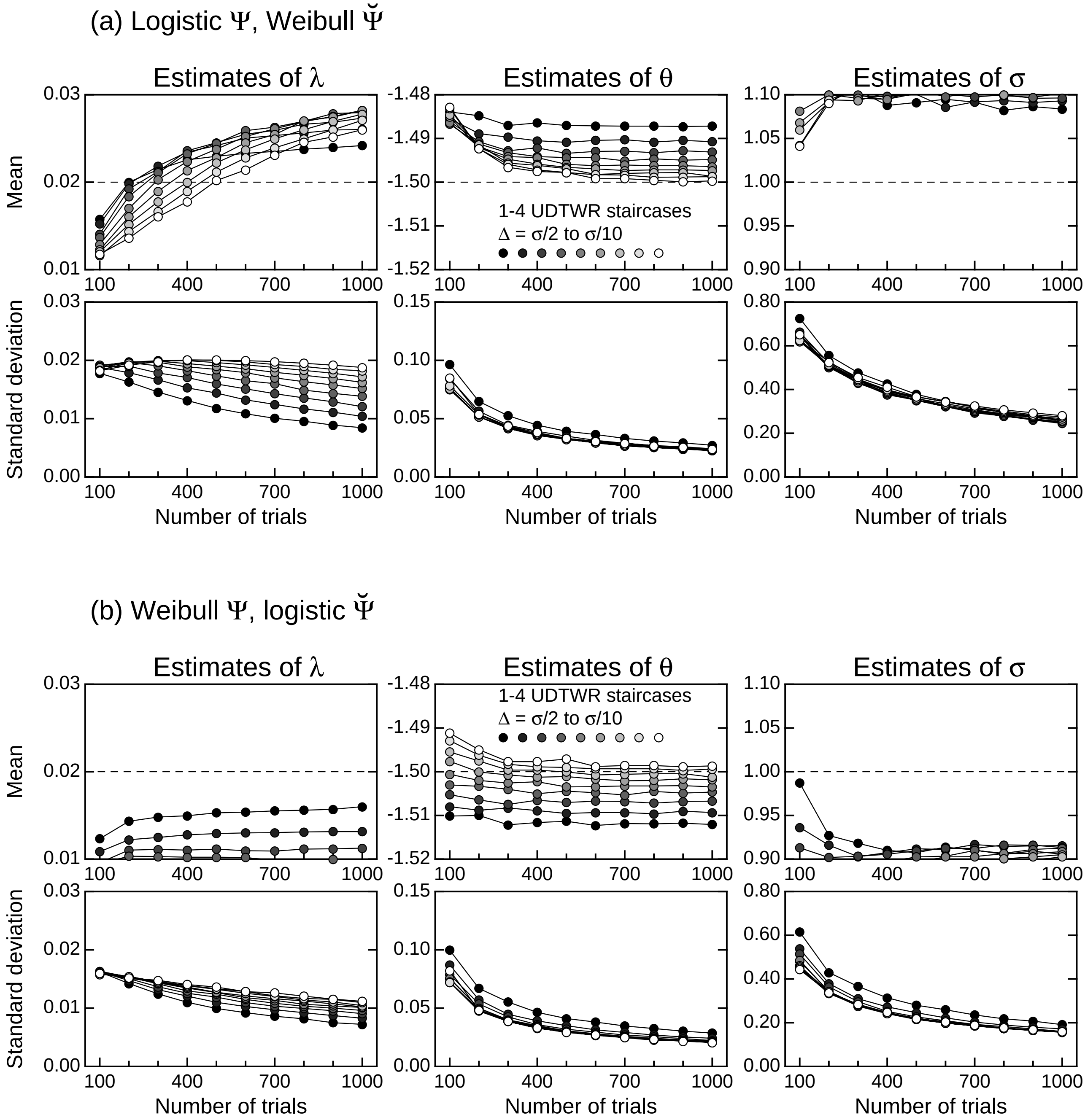

Figure 15. Mean (top row) and standard deviation (bottom row) of ML estimates of λ (left column), θ (center column), and σ (right column) arising from 1–4 UDTWR staircases as a function of the overall number N of trials in 2AFC detection tasks. In (a), the generating Ψ was logistic with γ = 0.5, λ = 0.02, θ = –1.5, and σ = 1 (b = 9.19) and the fitted $\breve{\Psi}$ was Weibull. In (b), the generating Ψ was Weibull with γ = 0.5, λ = 0.02, θ = –1.5, and σ = 1 (β = 2.66) and the fitted $\breve{\Psi}$ was logistic.

alternative, WLS estimation, cannot be used with single-presentation MOCS and that its use must discard non-trivial amounts of data with all of the remaining sampling plans. As a result, in 2AFC detection tasks with N = 1000 trials, usability indices were only 14.51% and 0.33% respectively for QUEST and UDTWR staircases because with these sampling plans most stimulus levels are tried just once. With the remaining plans (conventional MOCS, UDTR staircases and UDWR staircases), usability was slightly inferior than under OLS or ML estimation. The distributions of estimates under the latter sampling plans were nevertheless similar to those described above for OLS estimation. Results of application of WLS estimation in yes–no tasks and 2AFC discrimination tasks yielded a similar picture: only 1.02% usability for

QUEST, slightly lower usability for conventional MOCS and for UDTR and UDWR staircases than under OLS or ML approaches, and distributions of estimates that are thoroughly analogous to those obtained with OLS estimation. The lack of a significant improvement in the performance of WLS compared to that of OLS is understandable because weights are reciprocals of estimated variances at the data points, something that requires large numbers of trials per point in order to obtain reliable estimates of these variances (see Myers, 1990, p. 319).

As for BQ estimation, usability stayed at 100%, but estimates were generally poor: The distributions were either too broad or they were centered too far from true values.

It should be stressed that the input data were exactly the same in all cases (ML, OLS, WLS when reasonably applicable, and BQ) and, thus, the outcomes just discussed only reflect the differential performance of the four estimation methods. These results strongly advice against OLS, WLS, and BQ estimation.

7. The Optimal Sampling Plans

Figure 4 showed representative outcomes of all the types of sampling plan included in this study. Among the entire set of plans (but excluding the inefficacious QUEST), single-presentation MOCS is the worst in that it produces estimates of θ and σ that have larger variability than those obtained with other plans, but it produces better estimates of λ (and γ, where applicable; see Figure 11). Given that one is usually more interested in estimating θ and σ than in estimating λ or γ,[5] the use of single-presentation MOCS is inadvisable. On the other hand, the properties of all parameter estimates are similar with conventional MOCS and all forms of staircase. In principle, this could lead to considering the use of conventional MOCS, but recall that this plan ran here with the advantage that the range of stimulus levels only spanned the region of support of Ψ. Empirical use of conventional MOCS can only be expected to provide estimates with the characteristics reported here if it is set up as we did which, in turn, requires preliminary sessions to locate the region of support of Ψ. When the extra trials thus spent are counted up, the efficiency of conventional MOCS is seriously reduced. And there is also the issue that some variants of adaptive staircase have higher usability than comparable conventional MOCS (see Figure 5).

In contrast, adaptive staircases find the right location in only a few trials that were counted up in our analyses. But results in Sections 4–6 indicate some dependence of usability and the quality of estimates on relative step size, an unknown in practice because step size is set without knowledge of the value of σ. As discussed in Sections 4.3 and 5.3, large relative step sizes allow a precise estimation of γ and λ because a sufficient number of trials is thus placed beyond the region of support of Ψ. But large step sizes impair the estimation of θ and σ. This latter characteristic perhaps reveals only the effect of sampling resolution, which can easily be untangled from step size by interweaving *s* staircases using a common base step size Δ (chosen to be reasonably large) but whose sampling lattices are progressively offset by Δ/*s*. This strategy provides each staircase with large step sizes (valued at Δ) that allow placing trials where they are needed for a proper estimation of λ (and γ when necessary) without compromising the fine resolution (valued at Δ/*s* across the interlaced lattices) that is required for accurate estimation of θ and σ.

Sections 7.1 and 7.2 document the performance of configurations of two interwoven staircases that can be used with each class of Ψ as regards the status of γ. The need for tailored sampling plans arises from the peculiarities of each class of Ψ. All plans should render fine resolution within the region of support of Ψ so that θ and σ can be accurately estimated but,

- — in *m*AFC detection tasks, where the lower asymptote of Ψ is a constant, the uninformative area below the region of support of Ψ should be avoided, whereas the area above the region of support of Ψ need only be coarsely sampled to allow the estimation of λ;
- — in 2AFC discrimination tasks, where both asymptotes are determined by λ, a sufficient number of trials should be placed above or below (or both above and below) the region of support of Ψ because both areas are informative as to the value of λ; and
- — in yes–no tasks, where the lower and upper asymptotes of Ψ are respectively determined by γ and λ, a sufficient number of trials should be placed both below the region of support of Ψ (to allow γ to be estimated) and above it (to allow λ to be estimated).

7.1. 2AFC Detection Tasks

In 2AFC detection tasks, the strategy of interweaving two staircases with offset lattices and base step sizes Δ = σ/2, σ/3, σ/4, and σ/5 was evaluated for all UDTR, UDWR, and UDTWR staircases considered in Section 4. Each interwoven ran for *N*/2 trials so that *N* still represents the cost of the entire procedure, and the two interwovens differed only in that their starting points were offset by Δ/2 (as illustrated in Figure 16a for dual 1–3 UDTWR staircases).

[5] An exception to this general rule occurs when trials in *m*AFC detection tasks have to be segregated by presentation location or interval because Ψ must be fitted separately to each subset of data. In these cases, guessing strategies may result in γ being different within each subset, although it will still be valued at 1/*m* for the overall data (see García-Pérez, Giorgi, Woods, & Peli, 2005, their Appendix B).

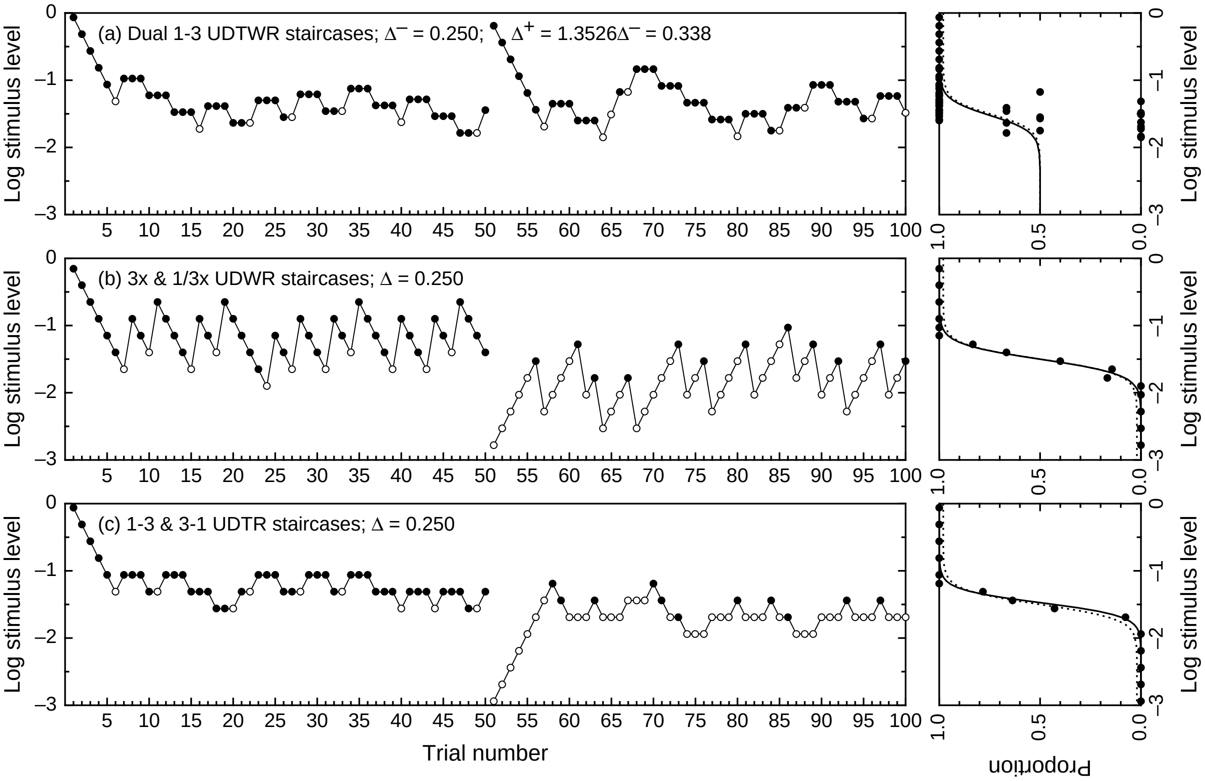

Figure 16. Sample tracks (left column) and results (right column) of the application of three sampling plans with dual staircases. These incidental results do not necessarily reflect the average performance of each plan, which is shown in Figures 17 and 18 below. All methods ran 100 trials of a putative 2AFC detection task (a), 2AFC discrimination task (b), or yes–no task (c). Pictorial conventions as in Figure 1. In all cases Ψ was logistic with λ = 0.02, θ = –1.5 and σ = 1, and the lower asymptote varied with the type of task: γ = 0.5 in (a), γ absent and replaced with λ = 0.02 in (b), and γ = 0.02 in (c). The two staircases in each track actually ran with their trials randomly interwoven but here they are shown unscrambled for clarity, the first staircase running for trials 1–50 and the second one running for trials 51–100. (a) Dual 1–3 UDTWR staircases (thus using $\Delta^+ = 1.3526\Delta^-$) each with $\Delta^- = \sigma/4 = 0.25$ that differ only in that their starting points are offset by $\Delta^-/2 = \sigma/8$. (b) Dual 3× and $\frac{1}{3}$× UDWR staircases with Δ = σ/4 too, one with an upper starting point and one with a lower starting point at locations such that their lattices yield a joint resolution of Δ/2 = σ/8 in the region of overlap, which coincides with the region of support of Ψ. (c) Dual 1–3 and 3–1 UDTR staircases with Δ = σ/4, one with an upper starting point and one with a lower starting point at locations such that their lattices also yield a joint resolution of Δ/2 = σ/8 within the region of support of Ψ.

Compared to single-staircase designs (see Figure 5), dual staircases improved usability and reduced or removed its dependence on relative step size for all types of staircase, although there were still differences among them in these respects. In 1–*d* UDTR staircases, 100% usability was attained regardless of relative step size only when *d* = 4 and after 900 trials; in *k*× UDWR staircases, 100% usability could not be achieved when *k* = 2 or Δ = σ/2 but otherwise it was attained in 700 trials regardless of relative step size; finally, in 1–*d* UDTWR staircases, 100% usability was attained irrespective of relative step size with 700 trials provided *d* ≥ 2. Independence of usability from relative step size is desirable because values for relative step size can only be chosen in practice when σ is known. Overall usability improved because λ and σ could both be estimated more often: λ as a result of the two sets of trials placed above the region of support of Ψ (see Figure 16a) and σ as a result of the double resolution within the region of support of Ψ. Proper estimates of θ could be obtained 100% of the times. Usability for both λ and σ showed much less (if any) dependence on relative step size than was reported in Figure 6.

Figure 17 shows the properties of the estimates obtained with dual-staircase designs involving the same rules for which single-staircase results were presented in Figure 8. The most conspicuous differences in favor of dual-staircase designs are a reduced underestimation of λ (compare the left columns in Figures 17 and 8) and a reduction of the dependence of the properties of the estimates of θ and σ on relative step size (center and right columns in Figures 17 and 8).

Considering usability and the properties of estimates, dual 1–3 UDTWR staircases with upper starting points offset by Δ/2 appear to be the best sampling plan in 2AFC detection tasks (see the lower part of Figure 17). This design renders data that plotted along with the fitted function may not look appealing (see the right panel in Figure 16a). If graphical appeal is wanted, dual 3× UDWR staircases render

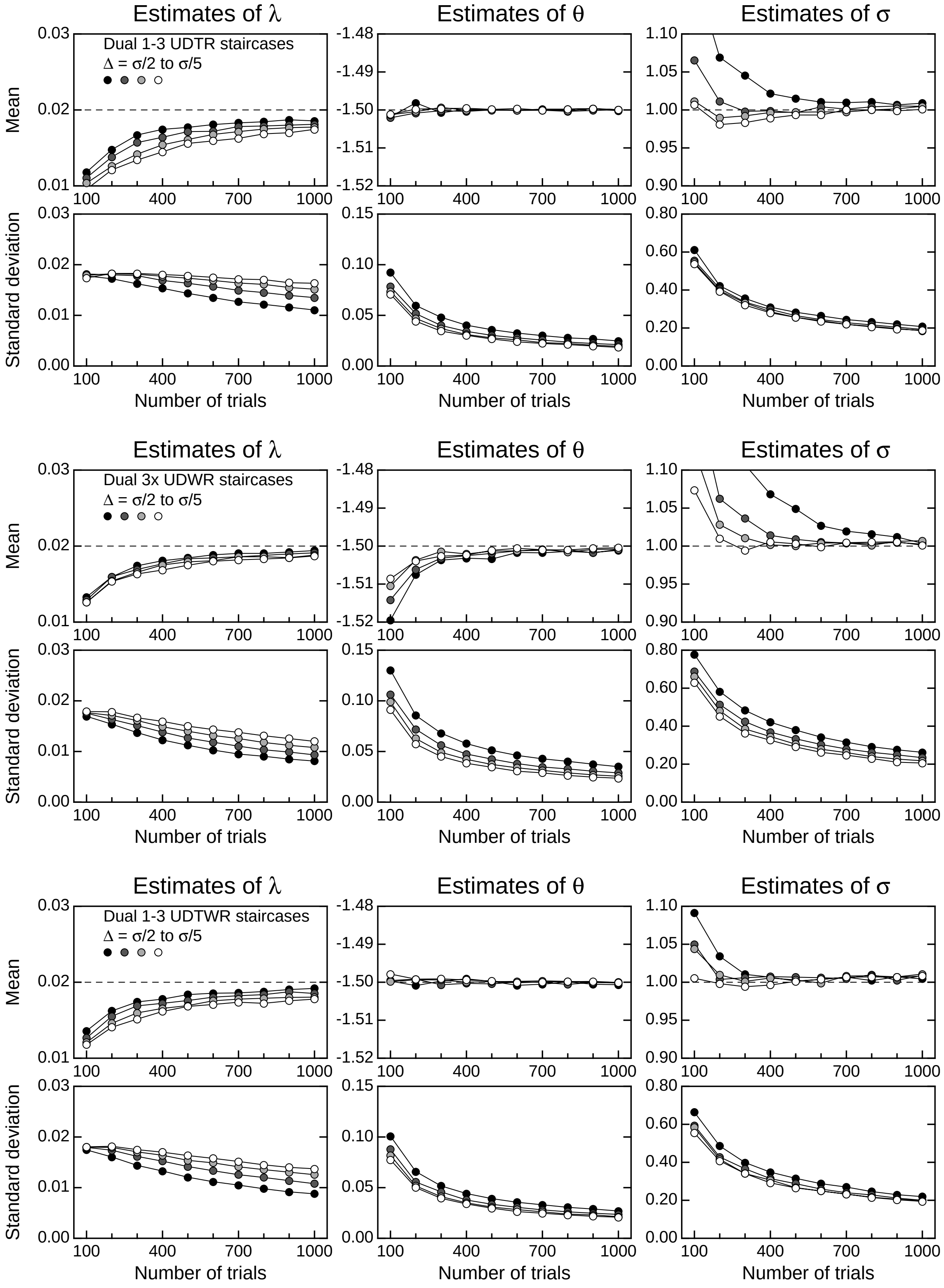

Figure 17. Similar to Figure 8, but 2AFC detection data were gathered with pairs of interwoven staircases each running for *N*/2 trials and whose upper starting points were offset by Δ/2. The upper, center, and lower parts respectively describe results for dual 1–3 UDTR staircases, dual 3× UDWR staircases, and dual 1–3 UDTWR staircases.

estimates with similar characteristics provided $\Delta \geq \sigma/2$ (center part in Figure 17). In either case, 500 trials ensure 98% overall usability (always implying 100% usability for θ and σ), negligible misestimation of θ and σ, and acceptable misestimation of λ.

7.2. Yes–No Tasks and 2AFC Discrimination Tasks

The same interweaving strategy was evaluated with yes–no and 2AFC discrimination tasks using base step sizes $\Delta = \sigma/2$, $\sigma/3$, $\sigma/4$, and $\sigma/5$ for all UDTR and UDWR staircases in Sections 5 and 6.1. Each interwoven ran also for $N/2$ trials. The two interwovens implemented inverse rules and had starting points at opposite ends of the range of stimulus levels (see Figures 16b and 16c).

In yes–no tasks, usability increased considerably in comparison with results reported in Figure 10a for single-staircase designs. Yet, overall usability of 100% irrespective of relative step size was only achieved with dual UDWR staircases involving $k = 4$ and $N \geq 600$. Higher usability arises because starting points on opposite ends of the range of stimulus levels allow gathering data to estimate both γ and λ, instead of only one of them in single-staircase designs (see Figures 10b and 10c). The interlaced lattices provide fine resolution within the region of support of Ψ so that all dual-staircase designs rendered 100% usability for σ regardless of relative step size with $N \geq 400$, besides the usual 100% usability for θ in all conditions. Then, limited usability arises from occasional failures to estimate γ and λ, not because σ or θ cannot be estimated.

Figure 18 shows the properties of the estimates from dual-staircase designs involving the rules for which single-staircase results were shown in Figures 13 and 14. Dual-staircase designs reduce misestimation of γ and λ, equalize the properties of $\hat{\gamma}$ and $\hat{\lambda}$ (compare the first and second columns in Figures 13, 14, and 18), and reduce or remove the dependence of the properties of the estimates of θ and σ on step size (third and fourth columns in Figures 13, 14, and 18). Also, dual UDWR designs (lower part of Figure 18) outperform dual UDTR designs (upper and center parts of Figure 18), providing unbiased estimates of all parameters regardless of step size when $N \geq 300$ trials. Among them, dual UDWR staircases with $k = 4$ represent the best choice. In terms of bias, this design yields estimates of γ and λ with properties similar to the optimum provided by single-presentation MOCS (compare with the first and second columns in Figure 12) along with estimates of θ that are comparable to those provided by conventional MOCS (compare with the third column in Figure 12) and better estimates of σ than those provided by any individual plan (compare with the fourth column in Figures 12–14). In terms of variability, the dual-staircase design also reaches an optimum compared to individual plans.

In 2AFC discrimination tasks, usability improved even further, understandably a consequence of improved usability for λ because data collected on both ends of the range of stimulus levels contribute to its estimation. This, in turn, resulted in still more accurate estimates of λ with smaller variability. Usability for θ and σ was not altered. Thus, the best dual-staircase design for use in 2AFC discrimination tasks involves UDWR staircases with $k = 4$ and provides unbiased estimates of θ and σ as well as acceptable (though minimally negatively biased) estimates of λ with $N \geq 300$ trials. Nevertheless, with dual UDWR staircases using $k = 2$ or 3, λ was only insignificantly underestimated and with a slightly larger variability but, in return, estimates of θ and σ (which were also unbiased) were noticeably less variable.

7.3. Discussion

A comparison of Figure 17 with Figure 8, on the one hand, and of Figure 18 with Figures 13 and 14, on the other, reveals that the best dual setup is much more so in the case of yes–no and 2AFC discrimination tasks than in the case of 2AFC detection tasks. Regardless of its magnitude, the improvement consists of (1) a reduction in bias with small numbers of trials, where it was more needed, and (2) a reduction or elimination of variations related to relative step size, which is useful in case of uncertainty as to the value of σ.

One might interweave four staircases instead of just two. The set of staircases would render lattices that are progressively offset by $\Delta/4$ units, and their starting points should balance the number of upper and lower positions so as to yield estimates of γ and λ with similar properties (if this strategy is used in 2AFC detection tasks, only upper starting points should be used as illustrated in Figure 16a). We have run simulations of 4-staircase designs that are otherwise analogous to those described in Sections 7.1 and 7.2 above. The results (not shown) indicate that usability increases because γ and λ can be estimated yet more often as a consequence of the extra sets of initial trials, whereas the properties of the estimates of θ and σ do not suffer. At the same time, the properties of the estimates became much more independent of step size, since even our largest steps ($\Delta = \sigma/2$) render a resolution of $\sigma/8$ within the region of support of Ψ.

8. General Discussion and Conclusion

8.1. Summary of Our Results

The most salient finding across Sections 4–6 is that no individual sampling plan provides optimal estimates of all parameters for each class of Ψ. Each parameter is optimally estimated by a different plan and the conjunction of those optimal properties across parameters define the target performance of the ideal sampling plan. The reason for the suboptimal behavior of all plans with some of the parameters

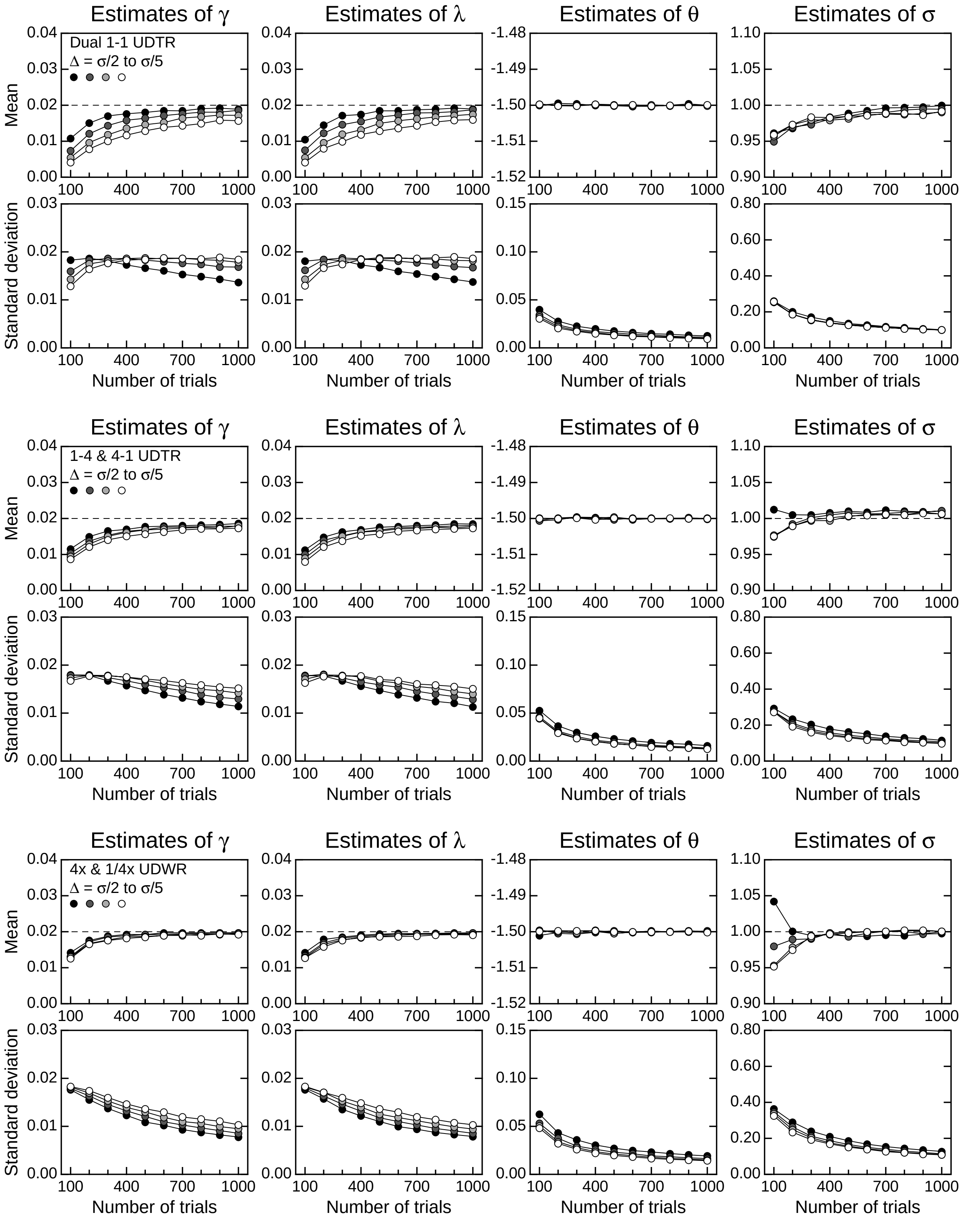

Figure 18. Similar to Figures 13 and 14, but yes–no data were gathered with pairs of interwoven staircases each running for *N*/2 trials, implementing inverse rules, and with starting points at either end of the range of stimulus levels. The lattices of the two staircases in each pair are also offset by Δ/2. The upper, center, and lower parts respectively describe results for dual 1–1 UDTR staircases, dual 1–4 and 4–1 UDTR staircases, and dual 4× and $\frac{1}{4}$× UDWR staircases.

is that each plan samples with a fixed pattern, when accurate estimation of different parameters appears to require that trials be placed with different patterns. Knowledge of these characteristics helped design optimal sampling plans that were shown in Section 7 to approach ideal performance. The optimal sampling plan differed for each class of Ψ because the classes are defined by the status of the lower asymptote and, therefore, by whether parameter γ needs to be estimated and also by whether or not both asymptotes are determined by the same parameter. In any case, the optimal sampling plan represents a larger improvement over individual plans in yes–no and 2AFC discrimination tasks than in 2AFC detection tasks.

Second, a mismatch between the actual form of Ψ and that of the fitted $\breve{\Psi}$ introduces systematic errors in the estimates. In practice, little can be done to prevent these errors because the form of Ψ is unknown. Thus, one can only hope that the actual form of Ψ across the set of subjects from which results will be aggregated and across the set of stimuli for which results will be compared does not vary, so that systematic errors caused by a potential mismatch between the forms of Ψ and $\breve{\Psi}$ do not contaminate the results differentially. An interesting and useful result in this respect is that this deterioration does not have implications on the choice of an optimal sampling plan.

A third major finding arises from our comparison of the estimates obtained with different estimation methods. From results discussed in Section 6.5, only the use of ML methods is advisable; OLS, WLS (when applicable) and, especially, BQ estimates are remarkably poorer.

Finally, psychophysical methods that place most trials in a narrow range of stimulus levels and fail to sample across the region of support of Ψ are remarkably inefficacious and greatly underestimate σ. This may seem natural in retrospect, but QUEST was included in our study because it has been (and still is) widely used for the estimation of Ψ, as mentioned in Section 2.3. Other adaptive methods exist that produce tracks similar to that shown for QUEST in Figure 1c in that they seek some target point quickly and then place all subsequent trials within a very narrow vicinity of that point. These include stochastic approximation (Robbins & Monro, 1951; used in simulations by Treutwein & Strasburger, 1999), YAAP (Treutwein, 1997; applied to empirical data by Treutwein & Strasburger, 1999), or ML-PEST (Hall, 1981; Harvey, 1997; used in empirical studies by Strasburger, 2001). All of these methods are bound to be inadequate for the same reason. Klein (2001, p. 1450, original italics) stressed that "*if one wants to measure (...) slope using adaptive methods, one should use a method that places trials at well separated levels.*" Our results with QUEST indicate that failure to do so overestimates slope (i.e., underestimates σ). Next we analyze the causes of slope bias in QUEST and other methods.

8.2. Slope Bias, or Support Instead of Slope

Several causes contribute to the slope bias documented in the literature (Berkson, 1955; Kaernbach, 2001; Leek et al., 1992; Maloney, 1990; O'Regan & Humbert, 1989) and that has also come out in our study. Not all of these causes have been previously acknowledged.

The most important and least known of them is, so to speak, constitutional. Although regularity conditions ensure that ML estimators of either b or σ are asymptotically normal and unbiased, this does not imply that the rate of convergence towards asymptotic behavior is identical for both. We have examined preasymptotic distributions of logistic $\hat{b}$ and $\hat{\sigma}$ obtained either by estimating b in Equation 1 or by estimating σ in Equation 10 for values of N up to 15,000 and using 10-level conventional MOCS. This analysis revealed that the pre-asymptotic distribution of $\hat{b}$ is positively skewed with its mode (not its mean) near the true value of b, whereas that of $\hat{\sigma}$ is positively skewed but peaks below the true value of σ. These characteristics were apparent in Figure 4. Moreover, the rate of convergence towards asymptotic normality and unbiasedness is slower for $\hat{b}$ than for $\hat{\sigma}$ (results not shown). Because bias must be defined as the difference between the expected value of an estimator (not the mode or the median) and the true value of the parameter, slope estimators must be positively biased in pre-asymptotic situations which, according to our results, imply all $N < 5000$ trials in 2AFC detection tasks. Conversely, estimates of σ reach asymptotic normality when $N \geq 400$ trials. In retrospect, there is probably a reason that statistical texts (e.g., Balakrishnan, 1992; Evans, Hastings, & Peacock, 1993) define the logistic distribution using a mathematical form that involves a divisive parameter such as σ in Equation 10 instead of a multiplicative parameter such as b in Equation 1.

A second cause of slope bias is that not every data set can yield a proper slope estimate (this was clear in Figure 4) and lousy data are likely to yield overestimation. Because slope estimation requires data collected at well separated locations, insufficient resolution may leave basically only one point at which the proportion of correct responses is far from the asymptotic regime of Ψ. The remaining points are then absorbed into the estimated asymptotes. This naturally inflates the slope estimate as seen in the top left and bottom right panels of Figure 3. The reason appears to be that, under the metric implied in the criterion function, there is basically only one point that is unmistakably within the region of support of Ψ (see the points marked with arrows in the top left and bottom right panels of Figure 3). This point dictates a location, the estimated curve passes right through it, and there is no extra information in the data that can set an upper limit for the slope of the curve once the asymptotes are established. It looks as if stray data with this characteristic were more prevalent than data with the characteristics shown on the right panel of Figure 1e, which tend to deflate slope estimates. Inadequate sampling plans like QUEST often produce data with a similar characteristic (see the right panel in Figure 1c), namely, a few points within an extremely narrow range of stimulus levels and displaying percentages dominated by binomial variability. Because ML-PEST is similar to QUEST in this respect, it seems clear that the high slopes reported by Strasburger (2001) are inflated and flawed, a result of the inadequate distribution of slope estimates arising from QUEST or QUEST-like methods (see Figure 4).

Slope bias has often been understated through inadequate statistical analyses. Either to compensate for the presence of improper estimates or directly to compensate for the positive skewness of the distribution of slope estimates, some authors have indicated bias by reporting median estimated slope (Kaernbach, 2001; Wichmann & Hill, 2001a), by reporting the geometric mean of estimated slope (Leek et al., 1992), or by reporting the mean of log estimated slope (Swanson & Birch, 1992). All these actions have the effect of making bias look smaller than it really is, but slope bias is the difference between mean estimated slope and true slope.

Kaernbach (2001) claimed that slope bias only occurs with data gathered with adaptive methods and he went to great lengths to make his point. He claimed without proof that slope estimates from MOCS are unbiased. The third and fourth columns in our Figures 4a and 4b show that conventional MOCS overestimates slope just as adaptive staircases do, and that estimates of σ with either method are unbiased when N is sufficiently but not unreasonably large. To look into this issue more closely, we carried out a simulation for one of Kaernbach's conditions, namely, 10,000 replicates of 1–1 UDTR staircases starting on threshold, $\Delta = \sigma/11$ (comparable to the step sizes used by Kaernbach), 48–120 yes–no trials (Kaernbach used 10–100), and a logistic Ψ (Kaernbach used a cumulative Gaussian) with $\lambda = \gamma = 0$, $\theta = -1.5$, and $\sigma = 1$; and we also simulated MOCS set up as always in this paper, running for the same numbers of trials, and using $L = 12$ so as to render the same sampling lattice as the adaptive staircase. The distributions of $\hat{b}$ turned out to be positively skewed both with adaptive staircases and with MOCS. In all cases the mode occurred at $\hat{b} \approx b = 9.19$ and the mean (ranging between 9.8 and 10.2) was still too high when $N = 120$. Then, slope bias is definitely not caused by the nature of adaptive data.

It should further be stressed that not all adaptive methods are equally prone to large slope bias and that a statement to that effect by Strasburger (2001) is in error. From our results, adaptive staircases that use fixed step sizes overestimate slope to the same extent as conventional MOCS, whereas adaptive methods that progressively reduce step size to end up placing most trials virtually at a single stimulus level (e.g., stochastic approximation, QUEST, ML-PEST, YAAP, etc) overestimate slope to a much larger extent. And not for being adaptive but for placing most trials within too narrow a range of stimulus levels.

Given all of the above, the practical question is how to go about this nagging property of slope estimates. We believe that our alternative measure of support σ is the solution for a number of reasons. First, σ and b (or β) are merely inversely related to one another (see Section 3.1) and, then, they do not carry different information: They only differ as to how they carry it. Second, σ has a much more direct interpretation by naturally describing the width of the region over which performance depends on stimulus level. Third, estimates of σ have much better statistical properties than estimates of b or β: For one thing, the distribution of $\hat{\sigma}$ is symmetric provided that N is not ridiculously small or an inappropriate method such as QUEST is used (see Figure 4). Fourth, because σ is a measure of length, interval estimates yield a range of values that are much easier to interpret (as a range of sizes) than the analogous interval estimates for b or β, which are obscure transformations of the slope of Ψ at a usually irrelevant point. Finally, the logistic distribution is more often defined in statistical texts using a divisive parameter as in Equation 10 than using a multiplicative parameter as in Equation 1, so there is no reason to linger on b. And other forms for Ψ (e.g., cumulative Gaussian) do not even have a slope parameter.

Support σ further provides a common metric for the comparison of psychometric functions of different form, but it also allows a better understanding of the differences between functions of the same form. For instance, if two logistic Ψ differ only in that one has $b = 10.21$ and the other has $b = 9.19$, little can be said immediately except that the former is steeper; but the first one has $\sigma = 0.9$ whereas the second has $\sigma = 1$, which leads to picturing the region of support of the first as being 10% narrower than that of the second. Also, suppose a psychophysicist claims that the slope of a Weibull Ψ is $\beta = 3.5$ for a condition where others have reported logistic Ψ with $b = 12.09$; they both would have found that $\sigma = 0.76$ and, hence, the range of stimulus levels that affects performance is identical in both cases.

8.3. Practical Recommendations

When the goal is estimating all the parameters of Ψ, our results advise against the use of QUEST and similar methods (stochastic approximation, ML-PEST, YAAP, etc) that progressively reduce step size to end up placing most trials within a narrow range of stimulus levels. This would only result in underestimates of support and improper estimates of the asymptotes.

Our results also recommend steering away from conventional MOCS. Its performance has been documented here for the ideal case in which its levels span the region of support of Ψ. However, setting up conventional MOCS in this way requires that the location and width of the region of support of Ψ be known in advance for every combination of stimulus and participant. Although rough estimates could be obtained in preliminary sessions, the cost of obtaining them should be taken into account when assessing the efficiency of the method. This cost was not included in our results, but it would greatly diminish the efficiency of conventional MOCS if the number of trials invested in the preliminary sessions is sufficiently large so as to provide dependable data. Trying to save trials in this preliminary phase can only result in initial estimates that are too rough, which will in turn affect the appropriateness of the choice of levels for conventional MOCS. In any case, for those who will not let go of conventional MOCS, our results can tell post-hoc what the quality of their final estimates was: Once Ψ has been estimated from the data, simply determine how many levels fell within the region of support of the estimated Ψ and how many trials were given altogether in that subset of levels; as shown in Sections 4.2 and 5.2, the overall number of trials within the region of support of Ψ indicates the quality of the estimates of θ and σ, provided there were at least five levels within the region of support of Ψ.

It is far more sensible to allow an adaptive staircase to find out the region of support of Ψ. This only leaves uncertainty as to the appropriate size for the steps. Interestingly, multiple-staircase designs discussed in Section 7 minimize or eliminate the effect of relative step size on the properties of parameter estimates. The only psychophysical method that our study found appropriate is the optimal dual-staircase design that was identified in Section 7 for each class of Ψ, namely, 1–3 UDTWR (or, as a second-best option, 3× UDWR) staircases in 2AFC detection tasks and 4× (but also 3× and 2×) UDWR staircases in yes–no or 2AFC discrimination tasks. Users need only guesstimate σ and set $\Delta \geq \sigma/3$ so that there are at least six samples within the region of support of Ψ. Of course, with UDTWR staircases (only recommended for use in 2AFC detection tasks) this issue is less critical because the sampling plan that they render explores the region of support of Ψ densely (see Figures 1f and 16a). Then, understandably, usability indices as well as the statistical properties of parameter estimates from UDTWR staircases depend on relative step size less than in other types of staircase (see Figures 5, 8, and 17). The only trouble may then come when UDTWR staircases cannot be used (i.e., in yes–no tasks, mAFC detection tasks with $m > 2$, and 2AFC discrimination tasks) and the guesstimate of σ turned out to be in gross error, so that Δ happened to be too large and fewer than four samples lay within the region of support of Ψ. Interweaving $s > 2$ staircases with offset lattices will protect against the effects of base steps Δ that were too large in retrospect, because the resolution within the region of support of Ψ is Δ/s.

Finally, the properties of parameter estimates vary with N. Figures 17 and 18 showed, for the optimal dual-staircase designs, that bias is negligible in all cases and that variability decreases as N increases. Those relations can be used to determine how many trials are needed to ensure that the standard errors stay within predetermined bounds. Our results indicate that 300 trials (in yes–no or 2AFC discrimination tasks) or 500 trials (in 2AFC detection tasks) suffice to obtain unbiased estimates with usability above 95% (always implying 100% usability for θ and σ) and acceptable standard errors. But our experience is that human observers are less dependable than simulated subjects in ways that the simulations reported here do not mimic (Alcalá-Quintana & García-Pérez, 2004b). As a rule of thumb, determine a minimum number of trials using the relationships of variability to N given in Figures 17 and 18 and, if at all possible, run 50% more trials.

References

Alcalá-Quintana, R., & García-Pérez, M.A. (2002, August). *Bias and standard errors in forced-choice Bayesian staircases*. Paper presented at the 33rd European Mathematical Psychology Group Meeting, Bremen, Germany.

Alcalá-Quintana, R., & García-Pérez, M.A. (2004a). The role of parametric assumptions in adaptive Bayesian estimation. *Psychological Methods*, *9*, 250-271.

Alcalá-Quintana, R., & García-Pérez, M.A. (2004b). Empirical performance of optimal Bayesian adaptive psychophysical methods. *Perception (Suppl.)*, *33*, 178.

Balakrishnan, N. (1992). *Handbook of the logistic distribution*. New York: Marcel Dekker.

Berkson, J. (1955). Maximum likelihood and minimum χ^2 estimates of the logistic function. *Journal of the American Statistical Association*, *50*, 130-162.

Brand, T., & Kollmeier, B. (2002). Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. *Journal of the Acoustical Society of America*, *111*, 2801-2810.

Brown, L.G. (1996). Additional rules for the transformed up-down method in psychophysics. *Perception & Psychophysics*, *58*, 959-962.

Dixon, W.J., & Mood, A.M. (1948). A method for obtaining and analyzing sensitivity data. *Journal of the American Statistical Association*, *43*, 109-126.

Evans, M., Hastings, N., & Peacock, B. (1993). *Statistical distributions* (2nd ed.). New York: Wiley.

Foster, D.H., & Bischof, W.F. (1991). Thresholds from psychometric functions: Superiority of bootstrap to incremental and probit variance estimators. *Psychological Bulletin*, *109*, 152-159.

Freeman, P.R. (1970). Optimal Bayesian sequential estimation of the median effective dose. *Biometrika*, *57*, 79-89.

García-Pérez, M.A. (1998). Forced-choice staircases with fixed step sizes: Asymptotic and small-sample properties. *Vision Research*, *38*, 1861-1881.

García-Pérez, M.A. (2001). Yes-no staircases with fixed step sizes: Psychometric properties and optimal setup. *Optometry & Vision Science*, *78*, 56-64.

García-Pérez, M.A., Giorgi, R., Woods, R.L., & Peli, E. (2005). Thresholds vary between spatial and temporal forced-choice paradigms: The case of lateral interactions in peripheral vision. *Spatial Vision*, *18*, 99-127.

Hall, J.L. (1981). Hybrid adaptive procedure for estimation of psychometric functions. *Journal of the Acoustical Society of America*, *69*, 1763-1769.

Harvey, L.O., Jr. (1997). Efficient estimation of sensory thresholds with ML-PEST. *Spatial Vision*, *11*, 121-128.

Kaernbach, C. (1991). Simple adaptive testing with the weighted up–down method. *Perception & Psychophysics*, *49*, 227-229.

Kaernbach, C. (2001). Slope bias of psychometric functions derived from adaptive data. *Perception & Psychophysics*, *63*, 1389-1398.

Kershaw, C.D. (1985). Statistical properties of staircase estimates from two interval forced choice experiments. *British Journal of Mathematical & Statistical Psychology*, *38*, 35-43.

King-Smith, P.E., & Rose, D. (1997). Principles of an adaptive method for measuring the slope of a psychometric function. *Vision Research*, *37*, 1595-1604.

King-Smith, P.E., Grigsby, S.S., Vingrys, A.J., Benes, S.C., & Supowit, A. (1994). Efficient and unbiased modifications of the QUEST threshold method: Theory, simulations, experimental evaluation and practical implementation. *Vision Research*, *34*, 885-912.

Klein, S.A. (2001). Measuring, estimating, and understanding the psychometric function: A commentary. *Perception & Psychophysics*, *63*, 1421-1455.

Kontsevich, L.L., & Tyler, C.W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*, *39*, 2729-2737.

Lam, C.F., Mills, J.H., & Dubno, J.R. (1996). Placement of observations for the efficient estimation of a psychometric function. *Journal of the Acoustical Society of America*, *99*, 3689-3693.

Leek, M.R., Hanna, T.E., & Marshall, L. (1992). Estimation of psychometric functions from adaptive tracking procedures. *Perception & Psychophysics*, *51*, 247-256.

Maloney, L.T. (1990). Confidence intervals for the parameters of psychometric functions. *Perception & Psychophysics*, *47*, 127-134.

Marks, B.L. (1962). Some optimal sequential schemes for estimating the mean of a cumulative normal quantal response curve. *Journal of the Royal Statistical Society, Series B*, *24*, 393-400.

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

Miller, J., & Ulrich, R. (2001). On the analysis of psychometric functions: The Spearman–Kärber method. *Perception & Psychophysics*, *63*, 1399-1420.

Müller, H.G., & Schmitt, T. (1990). Choice of number of doses for maximum likelihood estimation of the ED50 for quantal dose-response data. *Biometrics*, *46*, 117-129.

Myers, R.H. (1990). *Classical and modern regression with applications*. Boston, MA: PWS-KENT.

Numerical Algorithms Group (1999). *NAG Fortran library manual, Mark 19*. Oxford: Author.

O'Regan, J.K., & Humbert, R. (1989). Estimating psychometric functions in forced-choice situations: Significant biases found in threshold and slope estimations when small samples are used. *Perception & Psychophysics*, *46*, 434-442.

Owen, R.J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, *70*, 351-356.

Ramsey, F.L. (1972). A Bayesian approach to bioassay. *Biometrics*, *28*, 841-858.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, *22*, 400-407.

Santoro, L., Burr, D., & Morrone, M. C. (2002). Saccadic compression can improve detection of Glass patterns. *Vision Research*, *42*, 1361-1366.

Serrano-Pedraza, I., & Sierra-Vázquez, V. (2003, August). *Comparison between two methods for estimation of the parameters of a psychometric function: Effect of initial guess*. Poster presented at the 34th European Mathematical Psychology Group Meeting, Madrid, Spain.

Simmers, A.J., Bex, P.J., Smith, F.K.H., & Wilkins, A.J. (2001). Spatiotemporal visual function in tinted lens wearers. *Investigative Ophthalmology & Visual Science*, *42*, 879-884.

Snoeren, P.R., & Puts, M.J.H. (1997). Multiple parameter estimation in an adaptive psychometric method: MUEST, an extension of the QUEST method. *Journal of Mathematical Psychology*, *41*, 431-439.

Snowden, R.J., & Hammett, S.T. (1998). The effects of surround contrast on contrast thresholds, perceived contrast and contrast discrimination. *Vision Research*, *38*, 1935-1945.

Solomon, J.A., & Morgan, M.J. (2000). Facilitation of collinear flanks is cancelled by non-collinear flanks. *Vision Research*, *40*, 279-286.

Strasburger, H. (2001). Invariance of the psychometric function for character recognition across the visual field. *Perception & Psychophysics*, *63*, 1356-1376.

Swanson, W.H., & Birch, E.E. (1992). Extracting thresholds from noisy psychophysical data. *Perception & Psychophysics*, *51*, 409-422.

Treutwein, B. (1997). YAAP: Yet another adaptive procedure. *Spatial Vision*, *11*, 129-134.

Treutwein, B., & Strasburger, H. (1999). Fitting the psychometric function. *Perception & Psychophysics*, *61*, 87-106.

Watson, A.B., & Pelli, D.G. (1983). QUEST: A Bayesian adaptive psychometric method. *Perception & Psychophysics*, *33*, 113-120.

Watson, A.B., & Turano, K. (1995). The optimal motion stimulus. *Vision Research*, *35*, 325-336.

Werkhoven, P., & Snippe, H.P. (1996). An efficient adaptive procedure for psychophysical discrimination experiments. *Behavior Research Methods, Instruments, & Computers*, *28*, 556-562.

Wetherill, G.B., & Levitt, H. (1965). Sequential estimation of points on a psychometric function. *British Journal of Mathematical & Statistical Psychology*, *18*, 1-10.

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

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

Received April 28, 2005
Revision received June 13, 2005
Accepted June 24, 2005