Psychophysics with children: Evaluating the use of maximum likelihood estimators in children aged 4–15 years (QUEST+)

Mahtab Farahbakhsh
Child Vision Lab, Institute of Ophthalmology, University College London (UCL), London, UK

Tessa M. Dekker
Child Vision Lab, Institute of Ophthalmology, University College London (UCL), London, UK
UCL Division of Psychology and Language Sciences, London, UK

Pete R. Jones
Child Vision Lab, Institute of Ophthalmology, University College London (UCL), London, UK
NIHR Moorfields Biomedical Research Centre, London, UK

Maximum Likelihood (ML) estimators such as QUEST+ allow complex psychophysical measurements to be made more quickly and precisely than traditional staircase techniques. They could therefore be useful for quantifying sensory function in populations with limited attention spans, such as children. To test this, the present study empirically evaluated the performance of an ML estimator (QUEST+) versus a traditional Up-Down Weighted Staircase in children and adults. Seventy-one children (4.7–14.7 years) and 43 adults (18.1–29.6 years) completed a typical psychophysical procedure: Contrast Sensitivity Function (CSF) determination. Some participants were tested twice with the same method (QUEST+ or Staircase), allowing test-retest repeatability to be quantified. Others were tested once each with either method (QUEST+ and Staircase), allowing accuracy to be quantified. The results showed that QUEST+ was more efficient: In both children and adults, approximately half the number of ML trials were required to attain comparable levels of accuracy and reliability as a traditional Staircase paradigm, and plausible CSF estimates could be made in even the youngest children. The ML procedure was also as robust as the Staircase to lapses in concentration, and its performance did not depend on prespecifying correct model priors. The results show that ML estimators could greatly improve our ability to study sensory processes and detect impairments in children, although important practical considerations for-and-against their use are discussed.

Introduction

Rapid measures of sensory function are vital, both for basic research and clinical practice. To quantify sensory abilities or detect impairments, we may, for example, wish to know the smallest increment in luminance a child can see (Contrast Sensitivity), or the faintest intensity of sound they can hear (Audiometry). The most straightforward way to answer such questions is to present the full range of stimulus intensities, and to then determine, posthoc, the smallest intensity below which the child responded correctly (Method of Constant Stimuli). This approach is often impractical, however, as the required number of stimulus presentations quickly exceeds both the child's patience and the experimenter/clinician's time.

The traditional solution to limited testing time is to use adaptive procedures such as up-down Staircases to adjust the magnitude of the stimulus dynamically, trial-by-trial, in order to quickly locate the observer's threshold (or "just noticeable difference"). Over the years, a number of such Staircase methods have been developed (e.g., Method of Limits, Transformed Up-Down, Weighted Up-Down, Bekesy Tracking, PEST), and Staircases are used extensively throughout psychophysics (for reviews, see Ginsburg & Cannon, 1983; Leek, 2001; Treutwein, 1995). Many standard clinical tests (e.g., letter charts, acuity cards, and some forms of perimetry/microperimetry in vision, and audiometry in hearing) also consist, fundamentally of traditional Staircase algorithms.

Citation: Farahbakhsh, M., Dekker, T. M., & Jones, P. R. (2019). Psychophysics with children: Evaluating the use of maximum likelihood estimators in children aged 4–15 years (QUEST+). *Journal of Vision*, *19*(6):22, 1–19, https://doi.org/10.1167/19.6.22.

https://doi.org/10.1167/19.6.22 Received November 16, 2018; published June 27, 2019 ISSN 1534-7362 Copyright 2019 The Authors

This work is licensed under a Creative Commons Attribution 4.0 International License.

Downloaded from jov.arvojournals.org on 08/25/2019

In the last 40 years, however, various statistical optimization procedures have been developed which, in theory, have a number of advantages over traditional Staircase methods. This includes algorithms such as Best-PEST (Pentland, 1980), QUEST (Watson & Pelli, 1983), QUEST+ (Watson, 2017), ZEST (King-Smith, Grigsby, Vingrys, Benes, & Supowit, 1994), FAST (Vul, Bergsma, & MacLeod, 2010), Psi (Kontsevich & Tyler, 1999), Psi-marginal (Prins, 2013), qCSF (Lesmes, Lu, Baek, & Albright, 2010), MUEST (Snoeren & Puts, 1997), UML (Shen & Richards, 2012), and various unnamed methods (Green, 1993; King-Smith & Rose, 1997; Kujala & Lukka, 2006); for reviews, see Emerson (1986); Kingdom and Prins (2010); and Madigan and Williams (1987). These include both maximum likelihood and maximum a priori methods; however, following convention, we shall hereafter refer to both collectively as Maximum Likelihood (ML) estimators. In all cases, the variable(s) of interest are treated as unknown values in a parametric model, and after every trial the probability of each possible parameter value being true is computed explicitly (for mathematical details, see Kontsevich & Tyler, 1999; Watson, 2017). Framing the problem in this way confers several advantages. First, it becomes possible to compute the expected most informative stimulus to present on the next trial, thereby making the test more efficient—preventing, for example, the "slow downward crawl" that is often observed at the start of Staircases. Second, information can be integrated across multiple sources, including prior information (e.g., from normative data, or the individual's previous test results). Third, multiple parameters can be estimated simultaneously. For instance, the whole psychometric function can be measured instead of only its threshold, or we can quantify how a given threshold covaries with some second parameter—such as how detection thresholds vary with frequency, in the case of contrast sensitivity and audiometry. Finally, ML estimators also have a number of other attractive features, including the ability to specify dynamic stopping criteria based on statistical confidence (Alcala-Quintana & García-Pérez, 2005; Anderson, 2003; McKendrick & Turpin, 2005), and the ability to explicitly model and account for lapse rates (Prins, 2012, 2013; Wichmann & Hill, 2001).

Despite these theoretical advantages, many psychophysicists—the present authors included—have continued to favor traditional Staircases when working with children. This methodological inertia likely has many causes (see Discussion), but most fundamental is the concern that complex ML estimators may simply fail to function effectively when applied to children. This is for two main reasons. Firstly, ML estimators require us to make a number of simplifying assumptions (see Discussion). These assumptions may be acceptable in well-trained, relatively homogeneous groups of adults, but may be inappropriate for children, who, for example, often exhibit high levels of inattentiveness (Godwin et al., 2016; Jones, 2018b; Jones, Kalwarowsky, Braddick, Atkinson, & Nardini, 2015; Kaunhoven & Dorjee, 2017; Manning, Jones, Dekker, & Pellicano, 2018; Moore, Ferguson, Halliday, & Riley, 2008; Smallwood, Fishman, & Schooler, 2007; Wightman & Allen, 1992; Witton, Talcott, & Henning, 2017), response bias (Trehub, Schneider, Thorpe, & Judge, 1991; Werner, Marean, Halpin, Spetner, & Gillenwater, 1992), and other nonstationary behaviors. The concern is that such deviations from an "ideal" observer might at best degrade the efficiency of the test compared to standard Staircases, and at worse may cause the results to become excessively noisy or biased. Secondly, while ML estimators allow stimuli to be selected in a statistically optimal manner, there is a worry that more efficient stimulus sequences may "throw the baby out with the bathwater" by sacrificing some of the inadvertent beneficial properties that traditional Staircases possess: for example, the slow lead-in phase which gives the participant time to learn the task (Consolidation Trials), or the guarantee of an easier trial following an error (Motivational Trials).

The easiest and most common way to assess the efficacy of psychophysical methods is through Monte Carlo simulations. However, this approach is appropriate when considering "nonideal" observers such as children. A large number of variables can influence the responses of a human observer, and these variables are liable to interact in complex ways. For example, increased lapse rates can lead to longer test durations which can lead to fatigue, which can lead to further increases in lapse rates. For tractability, simulations invariably require the experimenter to make a large number of simplifying assumptions. Often, however, these are the same contentious assumptions that are made by the psychophysical algorithm themselves, and therefore beg the question of how well the algorithm can cope in the real world. Furthermore, and likely as a result, such simulations are often inconsistent with empirical data (see García-Pérez & Alcalá-Quintana, 2009). In short, the only way to be certain whether a given psychophysical method is effective is to assess their performance empirically.

The purpose of the present study was therefore to assess empirically the performance of a modern ML procedure (QUEST+; Watson, 2017) versus a traditional adaptive Staircase (Weighted Up-Down; Kaernbach, 1991). Specifically, we quantified their relative accuracy, speed, reliability, and robustness in children aged 4.7–14.7 years, and also in adults. For the test algorithm we selected QUEST+, as this is the most flexible/powerful procedure currently available, and essentially represents the superset of most ML algorithms to date (see Methods). For the comparison we

Downloaded from jov.arvojournals.org on 08/25/2019

selected a Weighted Up-Down Staircase method, as we have previously found it to be fast and reliable in children—though we do not believe our conclusions would have differed if another Staircase method had been used. For the psychophysical task, we used a four-alternative choice Gabor detection procedure to measure the contrast sensitivity function [CSF]. This procedure was intended primarily as representative of a "typical" psychophysical task, and again we believe that the present findings will generalize to other tasks. CSF determination was of particular interest, however, owing to it being a well understood task with previous normative data for comparison, and because of its particular importance both in basic research and clinically (see Hou et al., 2010; Lesmes et al., 2010; Rosén, Lundström, Venkataraman, Winter, & Unsbo, 2014).

Methods

Overview

Spatial Contrast Sensitivity Functions (CSFs) were measured in healthy children and adults using a four-alternative forced choice (4AFC) Gabor detection task. The task and stimulus-type remained the same throughout, but two different psychophysical procedures were employed: (a) a traditional Staircase procedure, and (b) a novel QUEST+ procedure similar to the "quick CSF" (qCSF; Hou et al., 2010; Lesmes et al., 2010; Rosén et al., 2014). Some participants were tested with both methods once, allowing their results and performance to be compared within-subjects. Other participants performed a single method twice, allowing test-retest repeatability to be quantified.

Participants

Participants were 114 normally-sighted individuals: 71 children aged 4.7–14.7 years (*M*: 9.0; *SD*: 2.1) and 43 adults aged 18.1–29.6 (*M*: 21.8; *SD*: 2.7). An additional three children were recruited, but their data are not reported as they did not pass the screening criteria for normal-vision (see below).

As detailed in Table 1, participants were randomly assigned to complete either both tests once, or one test twice. To evaluate test-retest reliability, 19 children and 30 adults repeated the same test twice (Table 1, Rows 1–2). To evaluate the relative accuracy and performance of the two methods, 34 children and 13 adults performed both tests once (Table 1, Row 3). An additional 18 children were assigned to complete both tests once, but ultimately contributed data for only one

	Participants			
	Children		Adults	
Test condition	*N*	Mean age (range)	*N*	Mean age (range)
2 × QUEST+ only	10	8.5 (6.1–9.9)	15	22.3 (18.1–29.6)
2 × Staircase only	9	8.4 (6.1–9.8)	15	20.8 (18.4–23.7)
1 × both	34	9.9 (7.2–14.7)	13	22.4 (18.1–25.3)
1 × QUEST+ only	6	5.3 (4.7–6.3)	0	
1 × Staircase only	12	9.2 (5.6–12.3)	0	
1 × QUEST+ with Bad Priors (extra)	20	9.4 (5.3–12.3)	0	
Total	71	9.0 (4.7–14.7)	43	21.8 (18.1–29.6)

Table 1. Breakdown of participants and test conditions. *Notes*: All participants completed one of the test conditions shown in rows 1–5 (see body text for details). Some children additionally completed the Bad Priors test condition (row 6); however, these individuals are not included in the total as they are already counted in rows 1–5.

(Table 1, Rows 4–5). Their failure to contribute data for the second test was for one of two reasons: Twelve children (the first 12 to be seen) completed both tests successfully, but the QUEST+ data were not analyzed due to a critical error in how the algorithm was implemented, leading to thresholds being grossly misestimated (see Parameter Domain, below). Six children were deemed too young to complete the Staircase condition within the allotted time, so only performed the shorter QUEST+ condition (see Results: Speed).

Finally, in addition to the main test conditions described above, 20 of the 71 children also completed an additional second variant of the QUEST+ algorithm in which a prior was used, but was intentionally misspecified (Table 1, Row 6; for further details see Results: Robustness to incorrect priors). These children were selected quasirandomly, primarily when scheduling permitted.

The distribution of children's ages can be seen below in Figure 9. Note that although a wide range of ages were examined, the majority of children (75%) fell between 6–11.5 years.

All participants were required to have normal or corrected-to-normal vision, as defined by no reported history of eye disease, and a binocular letter acuity score of 0.16 logMAR (6/9) or better, assessed using an ETDRS chart at 4 m (Precision Vision Ltd., La Salle, IL).

Adults were recruited through the UCL Psychology Subject Pool ("SONA"), and received £7/hour compensation. Children were recruited through the UCL Child Vision Lab volunteer database, and received certificates, small toys, and transportation costs. Informed written consent was obtained from all adults

Downloaded from jov.arvojournals.org on 08/25/2019

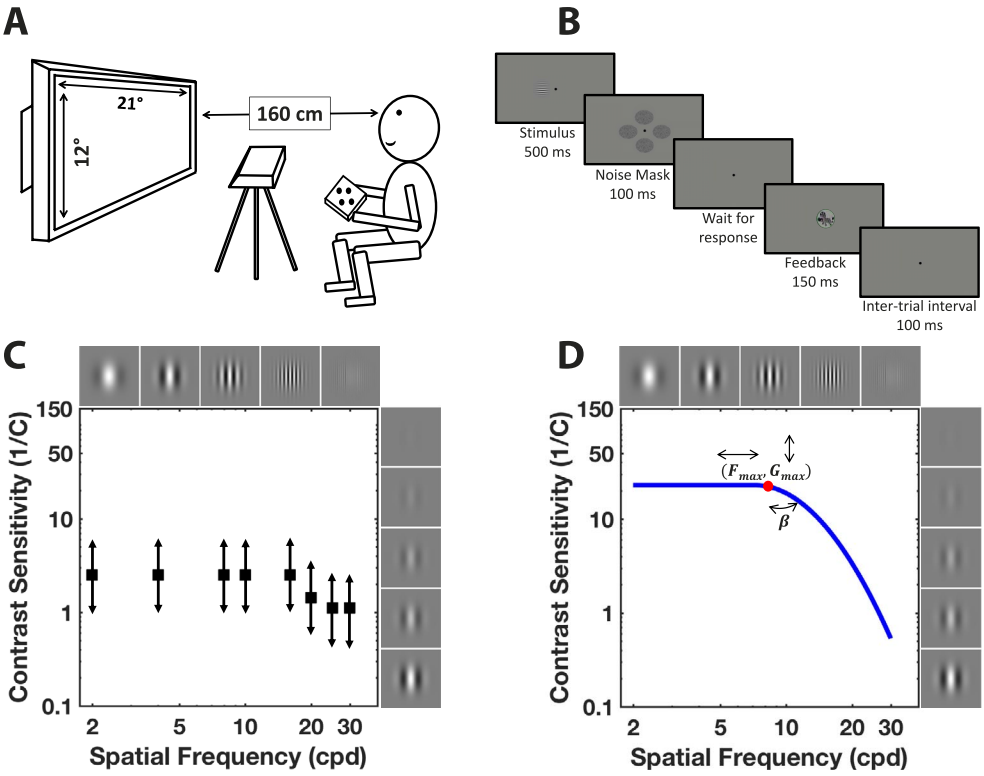


Figure 1. Methods. (A) Apparatus. Participants viewed stimuli binocularly at a distance of 160 cm. A remote eye-tracker was used to ensure central fixation. (B) General procedure. The task was to locate a single Gabor, presented at one of four locations (4AFC). (C) Staircase procedure. Eight independent, Weighted, Up-Down adaptive tracks were run at predetermined frequencies. Black squares denote the starting value of each adaptive track. (D) QUEST+ Procedure. The algorithm attempted to fit the three free parameters in Equation 2 (G_{max}, F_{max}, β) to the trial-by-trial response data.

and parents, and children provided verbal assent. The research was carried out in accordance with the tenets of the Declaration of Helsinki, and was approved by the UCL Ethics Committee (#1153/001).

Stimuli and apparatus

The stimulus was a horizontal Gabor patch of variable contrast and spatial frequency. The standard deviation of the Gaussian hull was 2.53° visual angle, and the total spatial support was 17.71°. The presentation duration was 500 ms, including 83 ms raised-cosine on/off ramps. The mean luminance of the Gabor was 136 cd/m^2, and it was presented against an equiluminant gray background.

On each trial, a single Gabor was presented at one of four cardinal locations, selected at random. The center of each location was 3.8° from a central fixation point, which consisted of a black circle 0.19° in diameter (see Figure 1B). The spatial extent of a Gabor is technically infinite; however, if we consider ±1 *SD* of the hull to be the "edge" of the stimulus, then the distance from the fixation point to the nearest edge was ~1.3°.

Stimuli were presented on a 27″, 10-bit IPS monitor (EIZO ColorEdge CG2730; 2560 × 1440 pixels; EIZO Co., Ltd., Birmingham, UK), connected via DisplayPort to a 10-bit graphics card (Nvidia GeForce GTX 650Ti; Nvidia Corp., Santa Clara, CA). The screen was viewed binocularly at a distance of approximately 160 cm, although viewing distance was not strictly controlled (Figure 1A).

Throughout the experiment, participants were regularly reminded to fixate the central spot. To ensure compliance, gaze location was monitored continuously using a remote eye tracker (Tobii X120; Tobii Technology AB, Danderyd, Sweden). If at any point the participant's gaze deviated by more than 2° from the central fixation spot, the experiment would automatically pause, and the fixation point would turn gray.

Hardware were controlled with custom MATLAB code (R2016b, MathWorks, Natick, MA), using the Psychophysics Toolbox v3 (Brainard, 1997; Kleiner et al., 2007) and Tobii SDK 3.0 (Tobii Technology AB, Danderyd, Sweden). The monitor was calibrated using a ColorCal colorimeter (Cambridge Research Systems, Cambridge, UK), and the calibration was validated using both a Minolta CS-100 (Minolta Camera Co.,

Downloaded from jov.arvojournals.org on 08/25/2019

Osaka, Japan), and also by the monitor's own integrated photometer (EIZO Co., Ltd., Birmingham, UK).

Testing took place in a quiet room under mesopic illumination (12.6 lx; Amprobe LM-120 Light Meter; Danaher Corporation, Washington, DC). Participants were seated throughout the test. Family members were discouraged from being present in the room during testing. When they were present, they sat outside the child's eyeline, and were asked to remain silent during testing. In order to avoid any potential distractions (or glare), the area around the participant and screen was also separated from the main room by a thin black cotton curtain.

Procedure: General

Participants performed a 4AFC Gabor detection task. This was presented as a game in which participants were asked to "find where the zebra is hiding."

During each trial, a single Gabor was presented for 500 ms, followed immediately by 100 ms white noise masks at all four potential target locations (see Figure 1B). Participants were then given unlimited time to indicate the location of the Gabor, which they did by pressing one of four arrows on a custom keypad. Participants generally pressed the response buttons themselves, although occasionally the experimenter would press the button for a period under instruction from the participant (i.e., if they believed that the participant was becoming inattentive). After a response was entered, veridical auditory and visual feedback were presented in the form of a happy/sad cartoon zebra and a corresponding sound. The next trial then commenced automatically after an intertrial interval of 100 ms. The stimulus parameters on each trial (contrast, spatial frequency), and the overall number of trials, were determined by the psychophysical algorithm (Staircase or QUEST+), the details of which are described below.

Furthermore, regardless of the procedure, ~30 additional catch trials were quasirandomly interleaved throughout the test trials (uniformly-randomly distribution across every six test trials in QUEST+; across every 13 test trials in Staircase). The stimulus on these catch trials consisted of a highly suprathreshold Gabor (spatial frequency: 3–16 cycles/°; contrast = 0.8–1.0). These stimuli were expected to be visible to all participants, and this was confirmed posthoc from their empirical data. The intended function of these trials was to quantify lapse rates (i.e., false negative responses), though as a secondary function they may also have served to motivate participants. These trials were not used when fitting the CSF.

Trials were divided into blocks ("levels" of the game). In the Staircase condition, there were eight blocks: Each block consisted of a single adaptive-track/spatial-frequency, and consisted of 47 trials on average (including catch trials). In the QUEST+ condition, there were six blocks: Each block consisted of exactly 35 trials (including catch trials), and the same instance of the algorithm ran continuously across all blocks. Participants were encouraged to take short breaks between blocks as required.

Prior to testing, participants also completed two practice blocks of first nine (block 1) and then fifteen trials (block 2). During the first practice block, the target Gabor continued to remain visible until the participant responded. Furthermore, the experimenter pressed the response key for the first three trials. This was to teach participants to the concept of the game. In the second block, the trials were identical to the main experiment; however, a fixed sequence of stimulus levels was used, designed to demonstrate a representative range of possible frequency/contrast levels: including both sub- and suprathreshold magnitudes. The criterion for completing the two practice blocks successfully was $\geq$ 90% correct responses on those trials expected to be suprathreshold. Most participants (97%) achieved this on their first attempt. Three individuals failed to reach this criterion on their first attempt, and so repeated both practice blocks, at which point the criterion was met.

Procedure: Staircase

As shown in Figure 1C, the Staircase procedure consisted of eight independent adaptive tracks, each of which independently estimated contrast sensitivity for a particular spatial frequency: 2, 4, 8, 10, 16, 20, 25, and 30 cycles/°. The order of the adaptive tracks was randomized between participants. Within each adaptive track, Michelson contrast was varied using a down-1 up-2 Weighted Staircase, which targets the 66.7% correct point on the psychometric function (NB: this is almost the same as the 62% threshold parameter of the Weibull function fitted by QUEST+; Madigan & Williams, 1987). Step sizes were multiplicative, and decreased every four trials from: 3, 2, 1.5, 1.25, remaining at 1.25 thereafter (e.g., a step size of 2 meant that the contrast halved/doubled after a correct/incorrect response). As illustrated in Figure 1D, each adaptive track started two steps away from the expected threshold at that frequency, as determined by piloting.

Note that a number of actions were taken to optimize the overall speed/efficiency of the Staircase procedure, including the use of a Weighted (rather than Transformed; Levitt, 1971) Staircase, multiplicative steps, progressively decreasing step sizes, and a starting point that varied with spatial frequencies. These

Downloaded from jov.arvojournals.org on 08/25/2019

optimizations were important to ensure that any observed differences in performance between the QUEST+ and Staircase procedure were not artefacts of a poor Staircase implementation.

Each adaptive track continued until at least 19 reversals had occurred (mean N trials = 47). Contrast thresholds were then calculated by geometric-mean-averaging the last eight reversals for each spatial frequency. The final output was a vector of eight values (contrast detection thresholds)—one per spatial frequency. During analysis, these values were then converted to a single CSF measurement by numerically fitting Equation 2 (see below) to the data. This fitting was performed using a bounded nonlinear minimization procedure (MATLAB's fminsearchbnd routine), with parameters constrained to fall within the same limits imposed by QUEST+ (see below).

Procedure: QUEST+

Detailed background information regarding QUEST+ is available elsewhere (Watson, 2017). However, in general terms it consists of a single, flexible algorithm that can (a) dynamically vary multiple properties of the stimulus simultaneously (here, both spatial frequency and contrast); (b) fit any arbitrary model to the raw trial-by-trial data (here, a three parameter CSF), and (c) evaluate the estimated likelihoods of all possible parameter values to determine the most informative stimulus to present on the next trial. For an intuitive graphical overview of how Maximum Likelihood approaches such as QUEST+ can be applied to the specific problem of CSF estimation, see figure 2 of Vul and colleagues (2010).

Model

The model that QUEST+ attempted to fit consisted of a traditional Weibull psychometric function (see Watson, 2017), which describes the expected proportion of a correct response, P_{correct}, as a function of stimulus contrast, c:

$$P_{\text{correct}} = \gamma + (1 - \gamma - \lambda) \times \left[1 - \exp_e\left(-10^{\varphi(\log_{10} c - \log_{10}\alpha)}\right)\right]. \quad (1)$$

The function's lower asymptote, γ, upper asymptote, λ, and slope, φ, were fixed parameters, with values 0.25, 0.1, and 3 respectively. The lower asymptote ("guess rate") was known a priori (i.e., in an mAFC paradigm $\gamma = 1/\text{m}$). The upper asymptote and slopes were set based on pilot data, and were only intended as approximations. The values were similar to those used elsewhere in the literature (e.g., $\varphi = 2$, $\lambda = 0.04$ in the qCSF method of Lesmes et al., 2010), but were somewhat greater to reflect the poorer concentration and/or lower sensitivity of some children. Note that, given the nature of QUEST+, λ and φ could have also been made free parameters, but this would have been impractical, given the additional data/trials required to constrain a five-dimensional parameter domain (though see Prins, 2013). The key "threshold" parameter, α, was a free parameter that varied with spatial frequency in accordance with the following three parameter CSF:

$$\alpha = \begin{cases} 1 \Big/ \exp_{10}\left(\log_{10}(G_{max}) - \log_{10}(2)\left(\frac{\log_{10}(f) - \log_{10}(F_{max})}{\log_{10}(2\beta)/2}\right)^2\right) & \text{if } f > F_{max} \\ \log_{10}(G_{max}) & \text{otherwise} \end{cases}, \quad (2)$$

where G_{max} represents peak gain (contrast sensitivity), F_{max} peak spatial frequency, and β the rate of fall-off at high frequencies (full width half maximum, in octaves). The action of these three parameters is illustrated graphically in Figure 1D.

Note that this formulation of the CSF represents a modified version of the log-parabola model recommended previously by Lesmes et al. (2010) and others (Watson & Ahumada, 2005). The only difference is that there is no fall-off/truncation of sensitivity at low frequencies (Lesmes et al.'s δ parameter). The reason for this difference was practical, not theoretical, and simply reflects the fact that no low frequency stimuli were presented. In the longer term, the decision to omit low frequencies was motivated by the fact that we are interested in developing a clinically relevant measure, and lower frequencies are difficult to spatially localize, exhibit greater individual variability (Watson, 2000), and are potentially redundant, with only acuity and peak sensitivity sufficient to describe most CSF curves to a first-degree of approximation (Pelli & Robson, 1991; though for dissenting opinions, see Dorr et al., 2017; Watson & Ahumada, 2005).

Stimulus domain

The stimulus domain (i.e., the set of potential Gabor parameter values) was bivariate, and consisted of 20 possible spatial frequencies spaced log-linearly from 2 to 30 cycles/°, and 20 possible grating contrasts spaced log-linearly from 0.01% to 100%.

Downloaded from jov.arvojournals.org on 08/25/2019

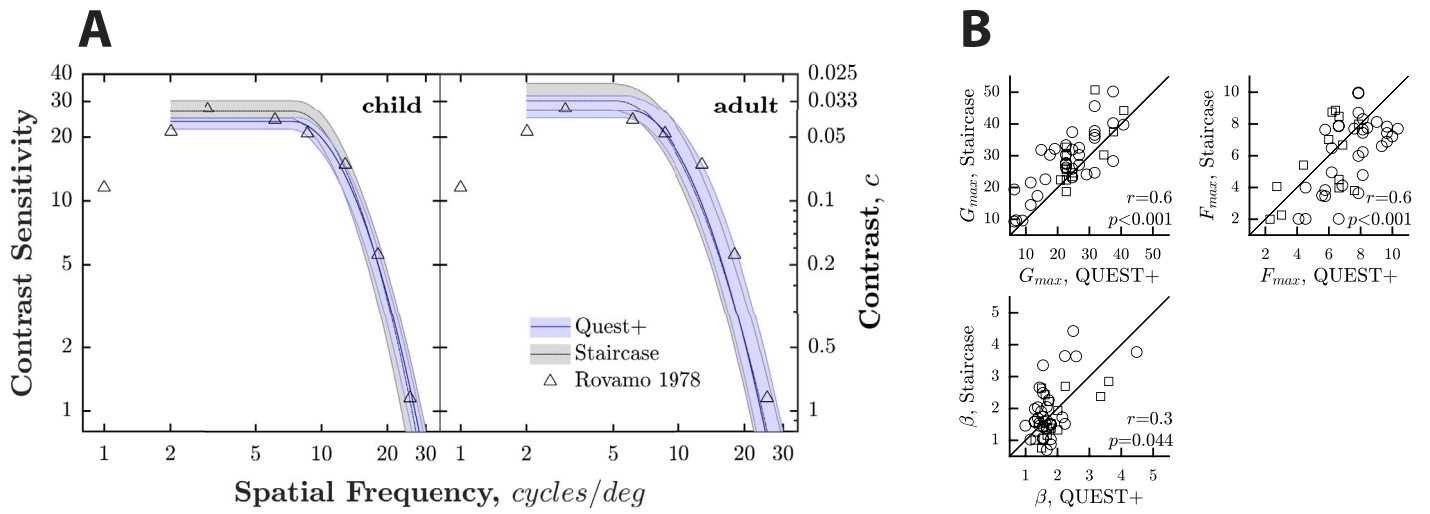


Figure 2. Accuracy. (A) Group-median CSFs ± 95% CI (derived using bootstrapping; $N = 20{,}000$, bias-corrected accelerated percentile method) for both children (left panel) and adults (right panel). Separate lines indicate measurements made using QUEST+ (blue, solid) and a Weighted Staircase procedure (Black, dashed). Triangles indicate previous data from the 1.5° eccentric condition of Rovamo and colleagues (Rovamo, Virsu, & Näsänen, 1978), angular distance from fixation point to nearest edge of grating. (B) Scatter plots for each of the three constituent parameters in the CSF model equation (Equation 2), as measured using QUEST+ (abscissa) and Staircase (ordinate). Each marker indicates an individual child (circle) or adult (square). Black lines indicate identity (perfect correlation).

Parameter domain

The parameter domain (i.e., the set of hypotheses that the model evaluated) was trivariate (i.e., the three free variables in Equation 2), and consisted of: 20 values of G_{max}, spaced log-linearly from 2 to 1500; 20 values of F_{max}, spaced log-linearly from 2 to 30; and 20 values of β, spaced log-linearly from 0.5 to 9. These values were determined based on piloting, and also on previous data from Lesmes and colleagues (2010; see also Hou et al., 2010). Due to user error, 12 children were tested using values of G_{max} spaced 30–1500. This range failed to include the likely true value for some participants (e.g., see Figure 3), and so could not possibly provide meaningful data. The data from these individuals are therefore not reported (see Methods: Participants).

Response domain

The response domain consisted of a single binary variable: correct or incorrect.

Priors

For the majority of testing, the prior probabilities for all three parameters were flat (all values equally likely). This is equivalent to not including an explicit Bayesian prior (though note that some prior assumptions are nonetheless always implicit in the choice of model, and in the specification of the stimulus and parameter domains). The decision not to use explicit priors was taken in order to avoid the suspicion that any observed benefits were due purely to the test being biased towards giving the correct result (see Discussion). The only exception to this was in the "Bad Priors" test condition (row 6 in the Table 1), in which incorrect priors were purposefully selected in order to probe the robustness of the method. The values of these incorrect priors can be seen below in the Results (Figure 10).

Output

The QUEST+ algorithm continued for 180 trials (fixed number of trials). In contrast with the Staircase procedure, no additional analysis was required to fit the CSF function posthoc, as the algorithm fits the three parameters in Equation 2 directly, after every trial. In practice, however, the QUEST+ routine was rerun during analysis using 80 steps per parameter, in order to minimize quantization error.

Code

QUEST+ was run using a custom MATLAB (MathWorks, Natick, MA) implementation (Jones, 2018a), which is freely available online under an Open Source license at www.github.com/petejonze/QuestPlus. At the time of writing, an independent MATLAB implementation of QUEST+ is also available as part of PsychToolBox (Brainard, 1997; Kleiner et al., 2007).

Downloaded from jov.arvojournals.org on 08/25/2019

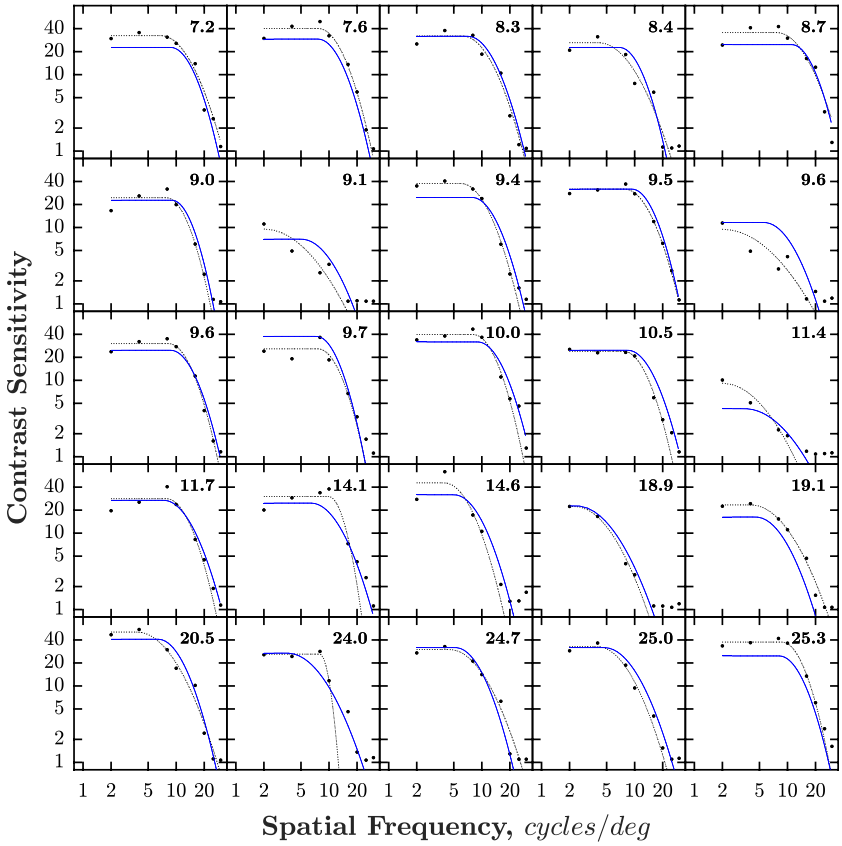


Figure 3. Accuracy: CSFs for 25 individuals (participants sampled uniformly across age). Curves represent the best fitting CSF, as measured using QUEST+ (blue, solid), and Staircase (black, dashed). Black circles represent the raw threshold measurement for each of the constituent Staircase measurements (corresponding constituent data for QUEST+ not shown). Text in each panel denotes the individual's age at time of testing, in years.

Results

Accuracy

Group-median CSF curves for both methods (QUEST+ vs. Staircase) are shown in Figure 2. There was good agreement between the two methods, and also with previous normative data (triangles in Figure 2A).

The only marked difference was a small but consistent tendency for G_{max} (the peak of the CSF) to be greater in the Staircase than the QUEST+ condition, and also greater for adults than children. Both differences were confirmed formally by fitting a linear mixed effects model with fixed terms for age group (Child vs. Adult) and condition (QUEST+ vs. Staircase), and random intercepts for each participant (i.e., to account for individuals who performed both conditions). Both the fixed terms of age, $t(171) = -3.60$, $p < 0.001$, and method, $t(171) = 3.75$, $p < 0.001$, were significant.

The age difference appears to be the result of a general developmental trend, which we report in more detail in a future manuscript. The reason for the difference between methods can be seen by inspection of the individual data shown in Figure 3. There, it can be seen that some observers (e.g., ages 7.6, 8.7, 14.6), though not all, exhibited a distinct peak in sensitivity between 4—8 cycles/°, with a fall off of sensitivity at 2 cycles/°. Because, as shown below in Figure 7, QUEST+ tended to primarily use the sensitivity at 2 cycles/° to define the upper asymptote, these higher sensitivities to midrange frequencies were not captured by the ML routine.

These systematic differences notwithstanding, there was good agreement between the two methods. Thus, the parameters G_{max} and F_{max} were strongly correlated ($r = 0.6$; see Figure 2B). The parameter β was also correlated, though the effect was weaker, likely due to the relatively smaller amounts of individual variability within each condition. Visually, there was good agreement between the CSF curves for each individual (Figure 3). Furthermore, to quantify the overall concordance of CSF estimates, we took the median thresholds shown in Figure 2A, and calculated the root mean squared error (RMSE) between the two sets of values, using the eight spatial frequencies common to

Downloaded from jov.arvojournals.org on 08/25/2019

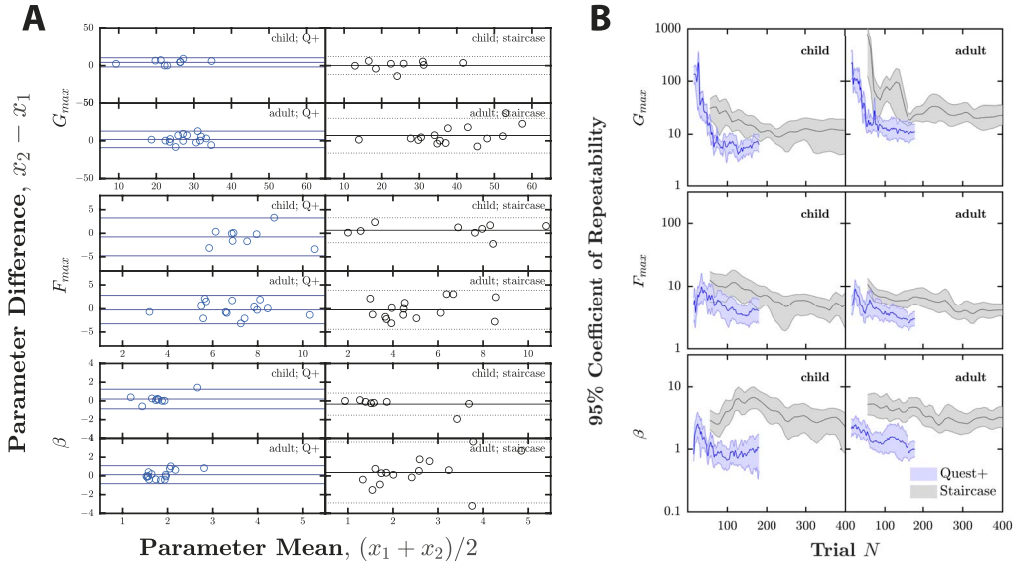


Figure 4. Reliability. (A) Bland-Altman plots for each CSF parameter and age group (NB: Using all trials). Circles represent individual participants. Dashed lines indicate the 95% Limits of Agreement. (B) 95% Coefficients of Repeatability for each parameter/age-group (panels), and for both QUEST+ (blue, solid lines) and Staircase (Blacked, dashed lines), computed using variable numbers of test trials. Note that this analysis involved performing repeated Bland-Altman analyses: The final point on each curve therefore corresponds to the data shown in Panel A.

both methods. The RMSE between methods was 1.04 dB for children, and 0.80 dB for adults; NB: following convention (Watson, 2000), 1 dB = 20 $\mathrm{Log}_{10}c$. For reference, the RMSE *within* methods (i.e., the difference between the median values from the 1st and 2nd run in those participants who performed the same method twice), was similar or greater: 0.66/2.45 dB for children (QUEST+/Staircase), and 1.28/0.99 dB for adults (QUEST+/Staircase). In this context, the concordance between methods appears strong.

Reliability

To assess the reliability of each method, Bland-Altman tests (Bland & Altman, 1999) were used to quantify test-retest repeatability in those participants who performed the same test twice (Rows 1 and 2 of Table 1). As shown in Figure 4, for five of the six parameter/age-group combinations, the QUEST+ data (with 180 trials) were more reliable across repeats than the Staircase method (with 350–400 trials). This indicates that QUEST+ was more reliable than the longer Staircase procedure.

To assess this difference formally, the Bland-Altman analysis was repeated for increasing numbers of trials, and bootstrapping was used to derive 95% CI for the 95% Coefficient of Repeatability. The results are shown in Figure 4B. By inspection of the confidence intervals, it can be seen that for any given number of trials, QUEST+ results were significantly more reliable (all parameters/ages). Furthermore, with QUEST+ the reliability of the test reached an approximate asymptote by 100 trials, whereas the reliability of the Staircase continued to improve gradually until at least 200 trials.

Speed

In terms of overall test duration, QUEST+ was faster than the Staircase condition, with average durations of 7.2 ± 2.9 versus 12.1 ± 2.3 minutes (median ± IQR), respectively. The ~40% difference between methods was consistent across all ages. However absolute test times did vary with age, increasing substantially for children younger than around 9 years (Figure 5A). Anecdotally, this change with age was largely due to the need for additional explanation, encouragement, and breaks. One corollary of this was that none of the very youngest children (< 6 years) were given the Staircase condition, as it was unlikely that they would complete it within a single test session. The difference in test durations between methods was largely due to the difference in number of trials (QUEST+: 180 test trials, plus 30 catch trials; Staircase: 319—429 test trials, plus 30 catch trials). When overall test duration was divided by number of trials, the duration *per trial* were similar for the two methods (in line with the use of identical within-trial procedures), although the same qualitative age effect remained (Figure 5B).

Since the stopping criterion was different for the two algorithms (see Methods), the question is whether the

Downloaded from jov.arvojournals.org on 08/25/2019

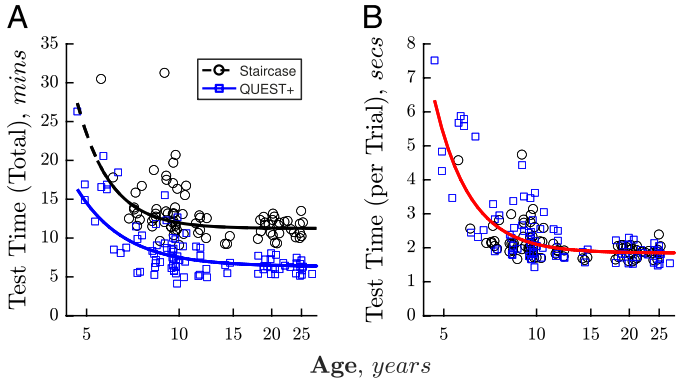


Figure 5. Speed: temporal duration. (A) Whole test duration (including breaks and catch trials) as a function of age, for QUEST+ (blue squares) and Staircase (black circles) conditions. Each marker represents an individual test run. Lines represent the best fitting two-term power series ($y = ax^b + c$). (B) Per-trial duration, as a function of age. These represent the same data as Panel A, but normalized by number of trials. The red line represents the best fit to both data sets.

additional trials in the Staircase were necessary. To answer this, Figure 6A shows group-median CSFs fitted using progressively more trials. From these it can be seen that QUEST+ converged on the reference value (the group-median CSF estimated using all available test data) in fewer trials than the Staircase procedure, suggesting that the former was more efficient. To quantify this difference more formally, we computed the extent to which an estimated CSF changed following each observer response. Specifically, as illustrated in Figure 6B, we computed Δ_{CSF} as the area of the difference between successive CSF estimates, computed every eight trials (i.e., one trial per adaptive track in the Staircase procedure, or every eight QUEST+ trials). This value provides an index of how stable the CSF measurement was, and is conceptually related to the "sweat factor" often used in simulations to evaluate the efficiency of a psychophysical algorithm (Treutwein, 1995). Values of Δ_{CSF} were computed independently for every test and then group-averaged. From the results shown in Figure 6B, it can be seen that the QUEST+ estimates were largely stable after 100 trials, whereas the Staircase estimates continued to vary even after 300 trials. Furthermore, for any given number of trials, the Staircase estimates were less stable than those for QUEST+. For example, the stability of the QUEST+ procedure (with 180 trials) was not reached by the Staircase procedure until approximately 300 trials). These results are in qualitative agreement with the test-retest reliability analyses presented previously (Figure 4B), and together, they indicate that QUEST+ was substantially (~50%) more efficient.

To begin to understand the reason for this difference in test efficiency, Figure 7 shows the distribution of

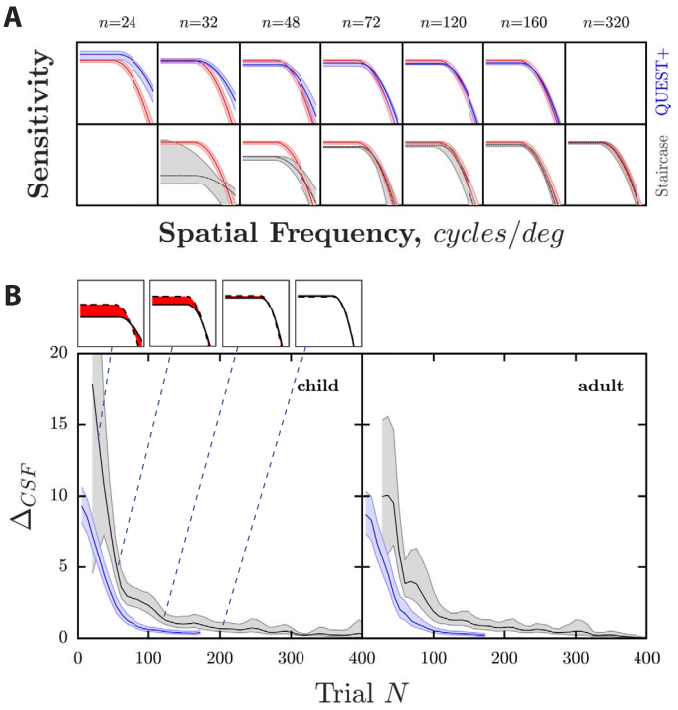


Figure 6. Speed: Test efficiency. (A) Group-median CSFs, computed using different numbers of test trials (same format as Figure 2A). The top row (blue curves) represent data collected using QUEST+. The bottom row (gray curves), represents data collected using a Weighted Staircase. The red curve in every panel is the fit to the Staircase data using all available trials, which we take as the reference value (i.e., the closest available approximation to the ground truth) (B) Mean (±95% CI) change in the CSF, as a function of number of trials. Change magnitude, Δ_{CSF}, was computed by numerically integrating the trial-by-trial difference in CSF, as illustrated by the red regions in the breakout panels (NB: these schematics are illustrative only, and not shown to scale). Values of Δ_{CSF} were computed independently for each individual (and each trial pair), and then group-averaged.

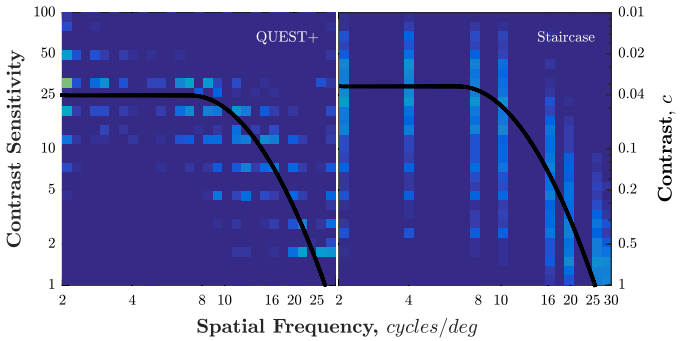


Figure 7. Heatmaps indicating the distribution of stimulus values when using QUEST+ (left) and Staircase (right). Trials from every condition/participant are included. Lighter colors indicate a greater proportion of trials (irrespective of whether responses were correct or incorrect). For reference, black curves give the group-median CSF values from Figure 2A.

Downloaded from jov.arvojournals.org on 08/25/2019

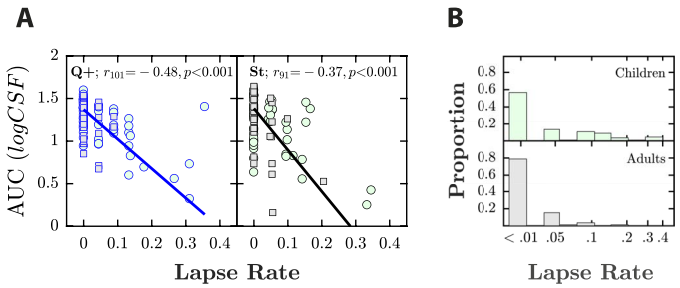


Figure 8. Robustness to lapses in concentration. (A) Scatter plot showing overall contrast sensitivity (Area under the CSF Curve) as a function of lapse rates (proportion of incorrect responses in false-negative catch trials). Markers represent individual tests for children (circles) and adults (squares). Data from all test conditions (Table 1, Rows 1–5) were included, so some individuals contributed multiple data points. Lines represent least-square geometric-mean regression slopes. *P* values are for Spearman correlations. (B) Histograms showing the distribution of lapse rates among children (top) and adults (bottom).

stimulus values under each of the two regimens. By inspection, it can be seen that QUEST+ rapidly homed in on observers' CSF, with the majority of stimuli presented near-threshold. Conversely, the stimuli in the Staircase condition were considerably more variable. The initial "crawl" down to threshold is evident in the greater proportion of high contrast values, and there was also a substantial overshoot, with many more very low contrast (subthreshold) values presented. A relatively large spread of stimulus contrast values is also evident in the QUEST+ at the highest spatial frequencies (see Figure 7, left panel). This is because the algorithm typically commenced testing at these higher frequencies (at which point true sensitivity was entirely unknown), and used the estimates of contrast sensitivity made there to constrain the rest of the curve (see Supplemental Figure S1).

Robustness: Lapses in concentration

To assess the resilience of each method to lapses in concentration, lapse rates were estimated from the ~30 false-negative (highly suprathreshold) catch-trials in each test, as follows:

$$\widehat{\text{Lapse Rate}} = \frac{m(1 - P_{\text{correct}})}{m - 1}, \quad (3)$$

where $P_{correct}$ was the proportion of correct responses, and m was the number of response alternatives ($m = 4$). Estimated lapse rates were then correlated against the area under the CSF curve (AUC), which we used as an overall index of sensitivity. The results are shown in Figure 8A. For both methods, higher lapse rates were correlated with lower estimated sensitivity (Spearman rank-order; $p < 0.001$), indicating that the accuracy of both methods was affected negatively by lapses in concentration. There was no significant difference between the best fitting geometric-mean regression slopes for the two conditions ($t_{194} = 1.30$, $p = 0.194$), indicating that neither method was substantially more or less resilient to lapses in concentration.

There was a significant negative correlation between age and lapse rates (Spearman rank-order; $r_{194} = -0.31$, $p < 0.001$), with children exhibiting more lapses than adults (see Figure 8B). These differences may explain—in part or in full—the developmental differences in sensitivity observed across both methods (see Results: Accuracy). This more general developmental question was not explored systematically in the present work, however. It is also worth noting that most children (57%) and adults (79%) exhibited no lapses (Lapse Rate = 0), as shown in Figure 8B.

Robustness: Young children

A particular concern was that ML estimators may work well in general, but produce highly inaccurate or imprecise estimates in some individuals, particularly the very young. To explore this possibility, Figure 9A shows the absolute difference in CSF parameter estimates for those individuals who completed the same method twice. From this it can be seen that there was no indication of a systematic age difference in terms of test repeatability (except for a very weak tendency for measurements to be more variable in adults in the Staircase condition). One possible exception to this is the very youngest observer, who did produce the most/second-most variable values of F_{max} and β. However, it should be noted that even greater variability was exhibited by some adults in the Staircase condition. Furthermore, values of G_{max} tended to be relatively consistent in younger observers in both conditions. Figure 9B also shows QUEST+ CSF curves for the six children younger than 6 years old (4.7–5.7 years). These curves appeared well formed, and differed only in values of $G_{max,}$ tending to be somewhat lower than those of their peers. In summary, while the small numbers preclude any firm conclusions, it appeared that even very young children were able to complete the QUEST+ test, and to do so in a manner that produced plausible and relatively consistent results.

Robustness: Incorrect priors

In all of the foregoing analyses, the QUEST+ algorithm was implemented using uniform priors. A unique advantage of ML estimators, however, is their

Downloaded from jov.arvojournals.org on 08/25/2019

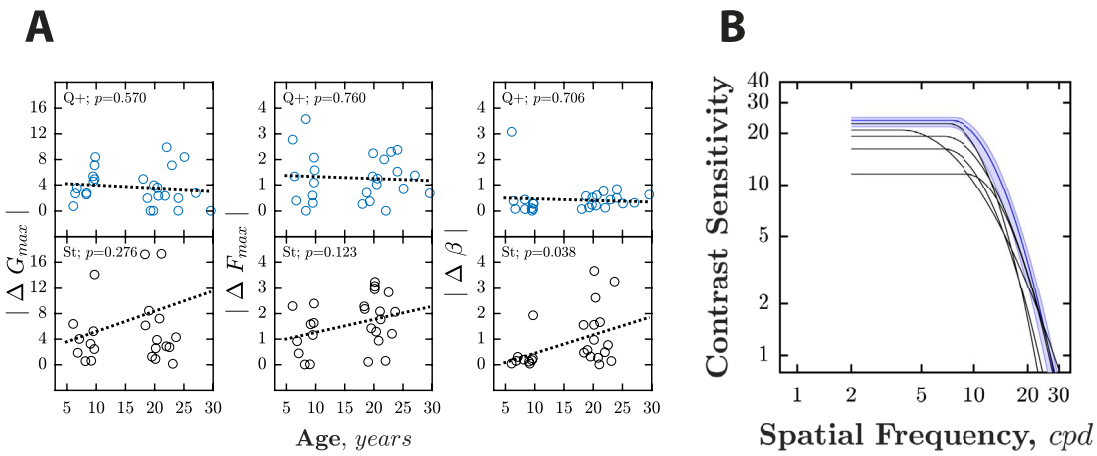


Figure 9. Robustness to young children. (A) Absolute test-retest *difference* in CSF parameters as a function of age. Each marker represents a single individual (only those individuals who performed the same test twice included). Black lines represent least-square geometric-mean regression slopes. *P* values are for Pearson correlations (NB: all nonsignificant, after correcting for multiple comparisons). (B) Individual CSF curves for the six children younger than six years measured using QUEST+, versus the group-median average across all children (blue shaded curve).

ability to integrate prior information to constrain the search space. If used incorrectly though, explicit priors can also be a source of potential measurement error. For example, if priors based on adult data are misapplied to children with lower sensitivities, then the test may become slower, less reliable, or systematically biased. A full characterization of the robustness of ML estimators to mis-specified priors was beyond the scope of the present work. However, to provide an initial assay, 20 children (Row 6 of Table 1) additionally completed a version of QUEST+ in the prior estimates for G_{max} and β were initially underestimated (see Figure 10B heatmaps). As shown in Figure 10, the final parameter estimates in the standard, "flat prior" condition (Figure 10A, black circles) were statistically indistinguishable from those in the corresponding, "bad prior" condition (Figure 10B, black circles), as indicated by their overlapping 95% CI. This implies that, in this instance, QUEST+ was able to recover from incorrect prior information, although whether

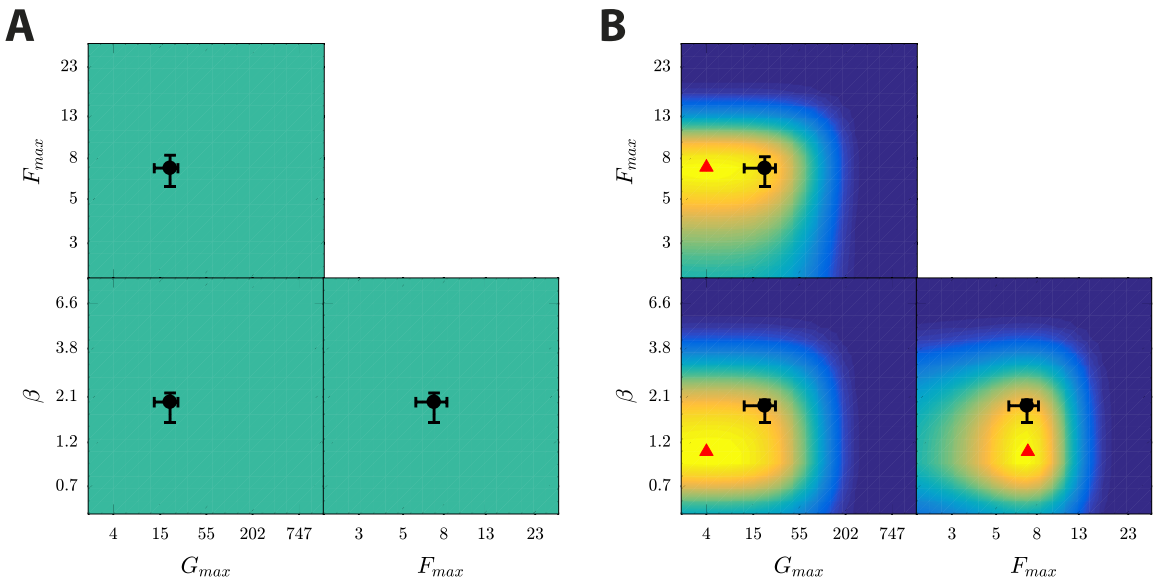


Figure 10. Robustness of QUEST+ to incorrect priors. (A) Group-median (±95% CI) parameter estimates (black circles) in the standard, flat-prior condition. The background heatmap indicates the prior likelihood of each possible parameter value (NB: In this instance, all values are equally likely). (B) Same as A, but when intentionally incorrect priors were initially specified (Bad Priors condition). Red triangles highlight the mode of the initial prior distributions. Note the similar parameter estimates in both subfigures.

Downloaded from jov.arvojournals.org on 08/25/2019

this continues to hold in more extreme cases remains unknown.

Discussion

The present study provides empirical evidence that ML estimators such as QUEST+ constitute an effective and efficient psychophysical tool for testing children aged 5 to 15 years, as well as adults. Even the youngest participants exhibited no difficulties in completing the QUEST+ assessment of CSF, and posthoc analyses indicated that approximately 50% fewer ML trials were required to attain comparable levels of accuracy and reliability versus a traditional Staircase paradigm.

The exact difference in performance between ML estimators and traditional Staircase designs will depend on the particulars of a given experiment. For example, it may have been possible to further optimize the Staircase condition by using fewer runs, or by modifying the steps sizes or starting location. Equally, it is likely that the speed and/or reliability of the QUEST+ paradigm could have been further improved, for example by varying the range or distribution of parameter values, or by using prior probabilities to explicitly constrain the search space. It is impossible, therefore, to put a definitive number on their relative efficiency. However, what is clear is that in the present study—in which both methods were implemented to the best of our abilities—the performance of QUEST+ substantially exceeded that of a Weighted Up-Down Staircase. It was also particularly reassuring to note that even the youngest children were able to successfully complete the QUEST+ procedure, and that our initial concerns regarding the suitability of ML estimators for children, such as the lack of a gradual "lead-in" phase, appeared unfounded (see Introduction).

The decision not to incorporate explicit priors in our implementation of QUEST+ is noteworthy in that the reported benefits are potentially an underestimate: It is likely that if appropriate priors had been specified then the relative speed and reliability of QUEST+ would have further improved. The decision not to use explicit priors was necessary, to avoid the suspicion that any benefits were purely an artefact of our a priori assumptions (a "lucky guess"). However, it is encouraging that even when prior information is not employed, ML methods still confer substantial benefits. This is particularly relevant for researchers working with children, where there is often a paucity of reliable prior information. Even in cases where prior data are available, some researchers may, on the basis of the present results, reasonably choose to not to incorporate them, to avoid the risk of introducing bias or user-error (see below).

Comparison to previous literature

The present empirical results are consistent with simulations, which have shown that ML estimators are highly efficient under idealized conditions (Kontsevich & Tyler, 1999; Prins, 2013; Turpin, McKendrick, Johnson, & Vingrys, 2003; Watson & Pelli, 1983). In terms of empirical data, the present results are most directly comparable to studies of the "quick CSF" (qCSF; Hou et al., 2010; Lesmes et al., 2010; Rosén et al., 2014): a specific implementation of a QUEST+ type method that also attempts to fit a variant of Equation 2. In those studies, it has been reported (in adults), that the majority of the CSF measurement error is reduced between 50–100 trials, and that by ~100 trials ML estimates are approximately as accurate and precise as more exhaustive measurements made using traditional "pointwise" threshold estimates at fixed frequencies (e.g., involving ~300 trials). The present findings are in good qualitative agreement, and suggest further that the same level of performance can be similarly expected in children as young as 5 years. This is particularly encouraging as CSF measurements have significant clinical utility as a more complete and precise measure of spatial vision than current gold standards (Arden, 1978; Ginsburg, 2003; Hess & Howell, 1977; Hou et al., 2016; Marmor, 1986; Onal, Yenice, Cakir, & Temel, 2008). The present results suggest that effective (~100 trial) measurements could be achieved in ~3.5 min in children 8+ years, and in 3.5–7 min in younger children. Previous reports have further indicated that, when Bayesian priors are employed, ML estimators can rapidly classify typical/atypical vision in adults with as few as 25–50 trials (Hou et al., 2010; Lesmes et al., 2010). Because the present study only examined normally sighted children, we are unable to say whether visual impairments could be detected as efficiently in children also; however, the present data give us no cause for doubt.

Wider considerations for and against the use of QUEST+

The present results demonstrate that ML estimators such as QUEST+ are more efficient than traditional Staircase algorithms in children, allowing a given degree of measurement precision to be attained in substantially fewer trials. This could prove vital in allowing detailed, rigorous measurements of vision even when testing times are constrained (e.g., in clinical environments) or when the observer's attention span is

Downloaded from jov.arvojournals.org on 08/25/2019

limited (e.g., in children). However, test efficiency is not always the only consideration when designing/implementing a psychophysical test. Here we consider other key concerns that may remain.

User error

Probably the greatest reason against the use of ML methods is the added potential for experimenter error. Measurements can be potentially rendered biased, noisy, or—in extreme cases—unusable due to a range of factors, including mathematical errors in the model formulation, an inappropriate choice of model, an inappropriate range of parameter values, inappropriate spacing of parameter values, inappropriate priors, an inappropriate stopping rule, or incorrect implementation of the algorithm itself.

Traditional Staircase designs are not immune from user error either. However, those errors tend to be relatively obvious and easy to diagnose. In contrast, errors with QUEST+ are often pernicious. For example, even after careful piloting in the present study, the QUEST+ data from the first 12 children (Table 1 Row 5) had to be discarded due to user experimenter in specifying the parameters domain (see Methods); and similar reports of "catastrophic failures" have been reported anecdotally by colleagues. Put simply, with complex statistical methods there is more to go wrong, and it is harder to notice when it does. In practical terms, this means that a much greater level of detailed testing and development is required, and these added costs must be traded-off against the gain in test efficiency. The increased potential for user error is of particular concern for work involving children, where overheads are often especially high (e.g., recruitment costs, labor costs, the cost of acquiring ethical permissions, etc.), and where repeating a failed experiment is often not an option.

Model assumptions

ML methods require the user to make a large number of assumptions. For example, the methods themselves typically assume that each response is made independently, and that the observer's ability to perform the task remains stationary throughout the test. Furthermore, the particular model implemented will also contain many further assumptions (e.g., pertaining to the shape of psychometric function, the values of its fixed parameters, and the possible values of its free parameters, see Methods). It is worth noting that "nonparametric" designs such as Staircases also contain a number of assumptions. For instance, Staircases likewise assume explicitly that responses are independent and the target parameter is stationary (see Levitt, 1971), and the chosen step size and starting positions implicitly express the experimenters' beliefs about the underlying psychometric slope and possible distribution of threshold values. However, their assumptions are generally fewer in number, and weaker in strength (i.e., that the psychometric function is monotonic, but not its precise form). It was encouraging that, in the present data, the ML estimator tended to be relatively robust even when these assumptions were breached (though see Limitations and Future work), and it could be argued that by making all assumptions explicit, methods such as QUEST+ actually encourages good practice. However, it is clear that need to specify statistical models in detail often necessitates much more extensive piloting, and can be particularly daunting when working with highly heterogenous populations such as children or clinical groups.

Loss of information

The efficiency of ML estimators follows in part from the fact that they minimize redundancy across measurements. For example, the Staircase procedure used in the present study estimated contrast sensitivity at eight discrete spatial frequencies, whereas with QUEST+, measurements were clustered primarily at a few key locations (see Figure 7), and these few locations were sufficient to constrain the three parameters in the CSF model. Other regions of the stimulus domain were tested rarely, if at all. For example, given that the model assumes contrast sensitivity is invariant at midfrequencies, all sensitivities between 3 cycles/° and F_{max} could be inferred exactly from sensitivity at 2 cycles/°. Such extrapolation can greatly reduce test durations, but also means that any localized deviations from the expected model may be missed. For example, as has been noted previously (Lesmes et al., 2010), the present model would be insensitive to "notches" in the CSF due to neurological disorders (Bodis-Wollner, 1972) or the adaptation of specific frequency-filters (Blakemore & Campbell, 1969). Furthermore, even in the present study, the assumed model meant that G_{max} was underestimated in those observers whose contrast sensitivities exhibit a distinct peak between 4–8 cycles/°. Elsewhere, fitting a continuous monotonic curve would be likewise inappropriate if trying to detect isolated (e.g., noise induced) defects in audiometry, or scotomas in the visual field.

To some extent, the fact that the ML method used in the present study would be insensitive to acute "notches" in sensitivity is a criticism of our implementation, rather than the ML approach per se. For example, it would be perfectly possible to fit a more complex spline type model, or to use multiple ML routines to independently estimate thresholds at predefined stimulus location (i.e., as required in the

Downloaded from jov.arvojournals.org on 08/25/2019

Staircase design). Indeed, the former approach has been used successfully to estimate equal-loudness contours in hearing (Shen, Zhang, & Zhang, 2018), while the latter is the approach currently used in visual field testing (Turpin et al., 2003). However, it is important to note that in doing so, the speed of the technique will be reduced concomitantly, and in the limiting case, there may well be relatively little difference in efficiency between the Staircase and ML procedures—though residual benefits may remain, such as the ability of ML procedures to integrate priors and dynamic stopping rules (Shen et al., 2018).

Not always necessary

As discussed previously, the setup costs (e.g., in terms of time, effort, and expertise) associated with ML algorithms can be substantial. In this context, it is important to note that simpler techniques may sometimes be perfectly sufficient. For instance, in the present study, both QUEST+ and a traditional Staircase ultimately produced concordant measurements with similar reliability. The difference was primarily that QUEST+ produced these measurements more quickly. This is consistent with previous reports that have found that, as long as the starting point and percent correct target are well chosen, "a simple Staircase is [often] as good as a fancy likelihood method" (Klein, 2001, p. 1436; see also Green, 1990).

Summary

Advanced ML estimators have tremendous potential for allowing detailed and precise psychophysical measurements to be made in otherwise hard to reach populations. They are particularly well suited to situations in which testing time is critical, such as when the search space is large, the number of participants is great, when the observer's concentration span is limited (as is often the case with young children), or when the test is only one component of an extensive battery. However, ML estimators are not always necessary or appropriate, and, in our experience, are more susceptible to experimenter errors, particularly among inexperienced users. Whether their increased efficiency justifies the additional investment in time, effort, and risk will depend on the circumstances of a given experiment. In many cases, the decision of whether to use ML procedures is ultimately a question of time-costs, and whether the cost is better born by the participant (in the form of longer testing times) or the experimenter (in the form of longer development times). If testing time is no object, then it may well be reasonable to use simpler psychophysical designs, such as Staircases.

Limitations and future work

Age

The majority of the children in the present study were aged 6 years and above. It remains an open question as to whether there is a lower limit on which ages ML estimators are suitable. The present results were encouraging, insofar as the six youngest children (4.7–5.7 years) were all able to complete the QUEST+ procedure (albeit more slowly than their peers), and there was no evidence of test reliability decreasing with age (Figure 9A). From this, it may appear that ML estimators are superior at all ages. However, this may not be correct. Consider, for example, the fact that young children and infants are liable to become highly despondent or distracted when the stimulus approaches threshold. This tendency is evident in the characteristic "saw tooth" pattern sometimes observed with traditional Staircase, whereby the child loses all interest when the task becomes impossible, and does not resume answering correctly for many trials thereafter (Jones et al., 2015; Moore et al., 2008). Such nonstationary behaviors can be accommodated to a degree when using a Staircase. For example, when the guessing (false positive) rate is low, it may be appropriate to simply use the lowest intensity correctly-responded-to stimulus as an index of threshold, rather than averaging reversals across trials (see Jones et al., 2015). However, it is not obvious what corresponding heuristics could be with QUEST+. Furthermore, it is unclear whether some children would ever regain interest without the run of easy trials guaranteed by a traditional Staircase following an incorrect response. Ultimately, these are empirical question, which can only be answered in future by testing cohorts of infants and preschool children.

Robustness

It remains unclear precisely how tolerant ML estimators are to "user error," either in the form of model misspecification or incorrect priors. With regards to model misspecification, it is likely, for example, that the values we selected for the slope and upper asymptote of the psychometric function (both of which were fixed parameters), were not strictly correct for any individual. They nevertheless appeared to serve as acceptable approximations, and this is consistent with previous observations that statistical methods tend to be reasonably tolerant of incorrect specifications of psychometric slope (Madigan & Williams, 1987). However, a breakdown point must exist, so the question remains: How accurate do the fixed parameters in a psychometric model need to be in order for it to operate acceptably?

Downloaded from jov.arvojournals.org on 08/25/2019

With regards to incorrect priors, the present results were encouraging, insomuch as the results remained largely invariant even when we intentionally included incorrect priors that (falsely) suggested that other values were more likely. However, as with model misspecification, it is trivially true that in some circumstances erroneous priors will lead to biased or misleading results. It would be likewise important to know how great a deviation is tolerable before recovery is no longer guaranteed, and what the associated impacts on reliability and test durations are.

Answering these questions is not trivial. There are a large number of interacting variables to consider, and a large quantity of data (both simulated and empirical) would be required to formulate any laws or heuristics. It would be extremely useful, however, to be able to specify what the reasonable operating range of a given model is, both when designing one's own tests and judging others'.

The bigger picture

Rapid, reliable tests are vital for detecting disease, evaluating the effectiveness of novel interventions, and studying how the visual system develops with age. Even with a maximally efficient algorithm, however, most psychophysical measurements will still require several minutes of sustained testing. In this context, it remains crucial that the child is sufficiently motivated. This can generally be achieved through positive feedback and encouragement, by introducing a sense of competition, and—perhaps most importantly of all—by presenting the task in the context of an intuitive and engaging narrative. Some experimenters additionally advocate the use of elaborate graphics or game mechanics during the task itself ("gamification"). While this can be effective, such measures can also prove counterproductive, introducing unnecessary confounds and/or prolonging the duration of the test.

Furthermore, even with a maximally compelling task, it is inevitable that some observers will lose focus at some point during the test. Such lapses in concentration can add measurement error, and/or cause sensory abilities to be systematically underestimated (Jones, 2018b; Manning et al., 2018). There have been suggestions in the past that ML estimators are more resistant to such lapses than traditional Staircase procedures (Manning et al., 2018; though see also García-Pérez & Alcalá-Quintana, 2009). This was not observed in the present study, however (see Figure 8). We therefore remain reliant, in the short term, on the shrewd judgements and timely interventions of an experienced experimenter to ensure good quality data.

In the longer term, however, the recent proliferation of affordable head-, face-, eye- and body-tracking sensors could allow us monitor task compliance autonomously (Jones, 2018b). And a key further advantage of ML estimators is that they provide a statistical framework for integrating these objective measures directly into the threshold estimation procedure. Thus, responses from "high concentration" trials can be given greater weight, while responses from "low concentration" trials are partially ignored (for details, see Jones, 2018b). This means that more attentive observers are required to complete fewer trials (incentivizing good behavior), and could drastically improve test reliability and accuracy in cases where trained experimenters are unavailable or too costly.

Summary and concluding remarks

Maximum Likelihood estimators (QUEST+) are more efficient than traditional psychophysical procedures (Up-Down Staircases) in children aged 4.7–14.7 years, and in adults. Given a large number of trials, both methods converged on the same estimates of an observer's Contrast Sensitivity Function. However, with a limited number of trials, an ML estimator was more accurate and reliable, and was as robust to lapses in concentration (though no better). ML estimators are therefore particularly well suited to situations where testing speed is imperative.

Keywords: maximum likelihood, psychophysics, children, contrast sensitivity function, QUEST+

Acknowledgments

This work was supported by Moorfields Eye Charity (#R160035A); the NIHR Biomedical Research Centre located at (both) Moorfields Eye Hospital and the UCL Institute of Ophthalmology; the European Social Research Council (#ES/N000838/1); an Ardalan Family Scholarship; the Persia Educational Foundation Maryam Mirzakhani Scholarship; and the Sir Richard Stapley Educational Trust (#313812).

Commercial relationships: none.
Corresponding author: Pete R. Jones.
Email: p.r.jones@ucl.ac.uk.
Address: Child Vision Lab, Institute of Ophthalmology, University College London (UCL), London, UK.

Downloaded from jov.arvojournals.org on 08/25/2019

References

Alcala-Quintana, R., & García-Pérez, M. A. (2005). Stopping rules in Bayesian adaptive threshold estimation. *Spatial Vision*, *18*(3), 347–374.

Anderson, A. J. (2003). Utility of a dynamic termination criterion in the ZEST adaptive threshold method. *Vision Research*, *43*(2), 165–170.

Arden, G. B. (1978). The importance of measuring contrast sensitivity in cases of visual disturbance. *British Journal of Ophthalmology*, *62*(4), 198–209.

Blakemore, C., & Campbell, F. W. (1969). On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images. *The Journal of Physiology*, *203*(1), 237–260.

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, *8*(2), 135–160.

Bodis-Wollner, I. (1972, November 17). Visual acuity and contrast sensitivity in patients with cerebral lesions. *Science*, *178*(4062), 769–771.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433–436.

Dorr, M., Lesmes, L. A., Elze, T., Wang, H., Lu, Z.-L., & Bex, P. J. (2017). Evaluation of the precision of contrast sensitivity function assessment on a tablet device. *Scientific Reports*, *7*:46706.

Emerson, P. L. (1986). Observations on maximum-likelihood and Bayesian methods of forced-choice sequential threshold estimation. *Attention, Perception, & Psychophysics*, *39*(2), 151–153.

García-Pérez, M. Á., & Alcalá-Quintana, R. (2009). Empirical performance of optimal Bayesian adaptive estimation. *The Spanish Journal of Psychology*, *12*(1), 3–11.

Ginsburg, A. P. (2003). Contrast sensitivity and functional vision. *International Ophthalmology Clinics*, *43*(2), 5–15.

Ginsburg, A. P., & Cannon, M. W. (1983). Comparison of three methods for rapid determination of threshold contrast sensitivity. *Investigative Ophthalmology & Visual Science*, *24*(6), 798–802.

Godwin, K. E., Almeda, M. V., Seltman, H., Kai, S., Skerbetz, M. D., Baker, R. S., & Fisher, A. V. (2016). Off-task behavior in elementary school children. *Learning and Instruction*, *44*, 128–143.

Green, D. M. (1990). Stimulus selection in adaptive psychophysical procedures. *The Journal of the Acoustical Society of America*, *87*(6), 2662–2674.

Green, D. M. (1993). A maximum-likelihood method for estimating thresholds in a yes-no task. *The Journal of the Acoustical Society of America*, *93*(4), 2096–2105.

Hess, R. F., & Howell, E. R. (1977). The threshold contrast sensitivity function in strabismic amblyopia: Evidence for a two type classification. *Vision Research*, *17*(9), 1049–1055.

Hou, F., Huang, C.-B., Lesmes, L., Feng, L.-X., Tao, L., Zhou, Y.-F., & Lu, Z.-L. (2010). qCSF in clinical application: Efficient characterization and classification of contrast sensitivity functions in amblyopia. *Investigative Ophthalmology & Visual Science*, *51*(10), 5365–5377.

Hou, F., Lesmes, L. A., Kim, W., Gu, H., Pitt, M. A., Myung, J. I., & Lu, Z.-L. (2016). Evaluating the performance of the quick CSF method in detecting contrast sensitivity function changes. *Journal of Vision*, *16*(6):18, 1–19, https://doi.org/10.1167/16.6.18. [PubMed] [Article]

Jones, P. R. (2018a). QuestPlus: A Matlab Implementation of the QUEST+ adaptive Psychometric Method. *Journal of Open Research Software*, *6*(1).

Jones, P. R. (2018b). Sit still and pay attention: Using the Wii Balance-Board to detect lapses in concentration in children during psychophysical testing. *Behavior Research Methods*, *51*(1), 28–39.

Jones, P. R., Kalwarowsky, S., Braddick, O. J., Atkinson, J., & Nardini, M. (2015). Optimizing the rapid measurement of detection thresholds in infants. *Journal of Vision*, *15*(11):2, 1–17, https://doi.org/10.1167/15.11.2. [PubMed] [Article]

Kaernbach, C. (1991). Simple adaptive testing with the weighted up-down method. *Attention, Perception, & Psychophysics*, *49*(3), 227–229.

Kaunhoven, R. J., & Dorjee, D. (2017). How does mindfulness modulate self-regulation in pre-adolescent children? An integrative neurocognitive review. *Neuroscience and Biobehavioral Reviews*, *74*, 163–184.

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*(7), 885–912.

King-Smith, P. E., & Rose, D. (1997). Principles of an adaptive method for measuring the slope of the psychometric function. *Vision Research*, *37*(12), 1595–1604.

Kingdom, F. A. A., & Prins, N. (2010). *Psychophysics: A practical introduction* (1st ed.). London, UK: Elsevier Academic Press.

Downloaded from jov.arvojournals.org on 08/25/2019

Klein, S. A. (2001). Measuring, estimating, and understanding the psychometric function: A commentary. *Attention, Perception, & Psychophysics*, *63*(8), 1421–1455.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception*, *36*(14), 1–16.

Kontsevich, L. L., & Tyler, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*, *39*(16), 2729–2737.

Kujala, J. V., & Lukka, T. J. (2006). Bayesian adaptive estimation: The next dimension. *Journal of Mathematical Psychology*, *50*(4), 369–389.

Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Attention, Perception, & Psychophysics*, *63*(8), 1279–1292.

Lesmes, L. A., Lu, Z.-L., Baek, J., & Albright, T. D. (2010). Bayesian adaptive estimation of the contrast sensitivity function: The quick CSF method. *Journal of Vision*, *10*(3):17, 1–21, https://doi.org/10.1167/10.3.17. [PubMed] [Article]

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

Madigan, R., & Williams, D. (1987). Maximum-likelihood psychometric procedures in two-alternative forced-choice: Evaluation and recommendations. *Perception & Psychophysics*, *42*(3), 240–249.

Manning, C., Jones, P. R., Dekker, T. M., & Pellicano, E. (2018). Psychophysics with children: Investigating the effects of attentional lapses on threshold estimates. *Attention, Perception, and Psychophysics*, *80*(5), 1311–1324, https://doi.org/10.3758/s13414-018-1510-2.

Marmor, M. F. (1986). Contrast sensitivity versus visual acuity in retinal disease. *The British Journal of Ophthalmology*, *70*(7), 553–559.

McKendrick, A. M., & Turpin, A. (2005). Advantages of terminating Zippy Estimation by Sequential Testing (ZEST) with dynamic criteria for white-on-white perimetry. *Optometry & Vision Science*, *82*(11), 981–987.

Moore, D. R., Ferguson, M. A., Halliday, L. F., & Riley, A. (2008). Frequency discrimination in children: Perception, learning and attention. *Hearing Research*, *238*(1–2), 147–154.

Onal, S., Yenice, O., Cakir, S., & Temel, A. (2008). FACT contrast sensitivity as a diagnostic tool in glaucoma. *International Ophthalmology*, *28*(6), 407–412.

Pelli, D., & Robson, J. G. (1991). Are letters better than gratings? *Clinical Vision Sciences*, *6*, 409–411.

Pentland, A. (1980). Maximum likelihood estimation: The best PEST. *Attention, Perception, & Psychophysics*, *28*(4), 377–379.

Prins, N. (2012). The psychometric function: The lapse rate revisited. *Journal of Vision*, *12*(6):25, 1–16, https://doi.org/10.1167/12.6.25. [PubMed] [Article]

Prins, N. (2013). The psi-marginal adaptive method: How to give nuisance parameters the attention they deserve (no more, no less). *Journal of Vision*, *13*(7):3, 1–17, https://doi.org/10.1167/13.7.3. [PubMed] [Article]

Rosén, R., Lundström, L., Venkataraman, A. P., Winter, S., & Unsbo, P. (2014). Quick contrast sensitivity measurements in the periphery. *Journal of Vision*, *14*(8):3, 1–10, https://doi.org/10.1167/14.8.3. [PubMed] [Article]

Rovamo, J., Virsu, V., & Näsänen, R. (1978, January 5). Cortical magnification factor predicts the photopic contrast sensitivity of peripheral vision. *Nature*, *271*(5640), 54–56.

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

Shen, Y., Zhang, C., & Zhang, Z. (2018). Feasibility of interleaved Bayesian adaptive procedures in estimating the equal-loudness contour. *The Journal of the Acoustical Society of America*, *144*(4), 2363–2374.

Smallwood, J., Fishman, D. J., & Schooler, J. W. (2007). Counting the cost of an absent mind: Mind wandering as an underrecognized influence on educational performance. *Psychonomic Bulletin & Review*, *14*(2), 230–236.

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*(4), 431–439.

Trehub, S. E., Schneider, B. A., Thorpe, L. A., & Judge, P. (1991). Observational measures of auditory sensitivity in early infancy. *Developmental Psychology*, *27*(1), 40–49.

Treutwein, B. (1995). Adaptive psychophysical procedures. *Vision Research*, *35*(17), 2503–2522.

Turpin, A., McKendrick, A. M., Johnson, C. A., & Vingrys, A. J. (2003). Properties of perimetric threshold estimates from full threshold, ZEST, and SITA-like strategies, as determined by computer

Downloaded from jov.arvojournals.org on 08/25/2019

simulation. *Investigative Ophthalmology & Visual Science*, *44*(11), 4787–4795.

Vul, E., Bergsma, J., & MacLeod, D. I. (2010). Functional adaptive sequential testing. *Seeing and Perceiving*, *23*(5–6), 483–515.

Watson, A. B. (2000). Visual detection of spatial contrast patterns: Evaluation of five simple models. *Optics Express*, *6*(1), 12–33.

Watson, A. B. (2017). QUEST+: A general multidimensional Bayesian adaptive psychometric method. *Journal of Vision*, *17*(3):10, 1–27, https://doi.org/10.1167/17.3.10. [PubMed] [Article]

Watson, A. B., & Ahumada, A. J. (2005). A standard model for foveal detection of spatial contrast. *Journal of Vision*, *5*(9):6, 717–740, https://doi.org/10.1167/5.9.6. [PubMed] [Article]

Watson, A. B., & Pelli, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. *Perception & Psychophysics*, *33*(2), 113–120.

Werner, L. A., Marean, G. C., Halpin, C. F., Spetner, N. B., & Gillenwater, J. M. (1992). Infant auditory temporal acuity: Gap detection. *Child Development*, *63*(2), 260–272.

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

Wightman, F. L., & Allen, P. (1992). Individual differences in auditory capability among preschool children. In L. A. Werner & E. W. Rubel (Eds.), *Developmental Psychoacoustics* (pp. 113–134). Washington, DC: American Psychological Association.

Witton, C., Talcott, J. B., & Henning, G. B. (2017). Psychophysical measurements in children: Challenges, pitfalls, and considerations. *PeerJ*, *5*, e3231.

Downloaded from jov.arvojournals.org on 08/25/2019