MLP: a MATLAB toolbox for rapid and reliable auditory threshold estimation

GRASSI, Massimo and SORANZO, Alessandro <http://orcid.org/0000-0002-4445-1968>

Running head: MAXIMUM LIKELIHOOD PROCEDURE

MLP: a MATLAB toolbox for rapid and reliable auditory threshold estimation

Massimo Grassi* and Alessandro Soranzo#

*Dipartimento di Psicologia Generale - Università di Padova Via Venezia 8
35131 – Padova
Italy

#School of Social Science and Law - University of Teesside
Middlesbrough - UK

Email: massimo.grassi@unipd.it
Phone: +39 049 8277494
Fax: +39 049 8276600

* Corresponding author

Abstract

In this paper, we present MLP, a MATLAB toolbox enabling auditory thresholds estimation via the adaptive Maximum Likelihood procedure proposed by David Green (1990, 1993). This adaptive procedure is particularly appealing for those psychologists that need to estimate thresholds with a good degree of accuracy and in a short time. Together with a description of the toolbox, the current text provides an introduction to the threshold estimation theory and a theoretical explanation of the maximum likelihood adaptive procedure. MLP comes with a graphical interface and it is provided with several built-in, classic psychoacoustics experiments ready to use at a mouse click.

MLP: a MATLAB toolbox for rapid and reliable auditory threshold estimation

In this paper, we present MLP, a MATLAB toolbox enabling auditory thresholds estimation via the adaptive Maximum Likelihood procedure proposed by David Green (1990, 1993). This procedure (hereafter referred to as ML) is particularly suitable to estimate thresholds with an optimal compromise between accuracy and rapidity. For this reason, the ML procedure has been used successfully in clinical contexts (e.g., Florentine, Buus, & Geng, 2000), in studies with children (e.g., Wright et al., 1997) as well as in studies with a large number of subjects (e.g., Amitay, Irwin, & Moore 2006). For the same reason, it is suitable for those studies where subjects perform various tasks, therefore, when each task has to consume only a portion of the subject's time. The ML procedure is largely known, used and appreciated by the auditory community, it has collected more than one hundred and twenty citations and the majority of these citations come from journals specialized in the auditory research [footnote 1]. Thus, the user of this procedure can benefit of a large background literature to optimise his/hers own threshold estimation. As far as we know, MLP is the first software implementing an adaptive psychophysical procedure with a graphical interface in a freely downloadable version and it is provided with several built-in, classic psychoacoustics experiments ready to use at a mouse click.

In the next section, we give a short introduction to the threshold estimation theory. The reader familiar with these concepts may wish to skip this section. The ML procedure and the MLP toolbox will be illustrated after this section.

Sensory Threshold Estimation Theory

Sensation moves within and across two types of thresholds: *detection* and *discrimination*. The *detection* threshold is the minimum detectable stimulus *level* [footnote 2] in the absence of any other stimuli of the same sort. In other words, the detection threshold marks the beginning of the sensation of a given stimulus. The *discrimination* threshold is the minimum detectable *difference* between two stimuli levels. Therefore, for a given sensory continuum, the discrimination threshold *cuts* the steps into which the sensory continuum is divided.

The *detection* threshold can be estimated either via *yes/no* tasks or via multiple Alternative Forced Choice tasks (in brief nAFC, with n being the number of alternatives). The *discrimination* threshold, on the contrary, must be estimated exclusively via multiple nAFC tasks. In *yes/no* tasks, the subject is presented with a succession of different stimuli levels (spanning from below to above subject's detection threshold) and is asked to report whether s/he has detected the stimulus (*yes*) or not (*no*). In nAFC task, the subject is presented with a series of n stimuli differing in level. In audition, because the various stimuli have to be presented in temporal succession, tasks are often multiple intervals tasks (i.e., mI-nAFC). In a nAFC task, one stimulus (the *variable*) changes its level across the trials; whereas the level of the others (the *standards*) is fixed. The difference between standard and variables ranges from below to above subject's detection (or discrimination) threshold. After each trial, the subject is asked to report which the variable stimulus was.

Figure 1 shows the hypothetical results of a yes/no task.

FIGURE 1 ABOUT HERE

The graph shows the relation between the stimulus level and the subject's performance together with one function fitting the hypothetical data. This function is referred to as the *psychometric function.* Independently from the task type, and from the type of threshold being measured, behavioural data are fitted with a sigmoid function such as that represented in Figure 1. Different types of psychometric functions can be adopted to fit experimental data, for example, the logistic, the Weibull and the cumulative Gaussian.

The general equation of a psychometric function is the following (adapted from Wichmann & Hill, 2001) representing subject's performance as a function of the stimulus level *x*.

$$\Psi(x;\alpha,\beta,\gamma,\lambda)=\gamma+(1-\gamma-\lambda)\,f\,(x;\alpha,\beta) \qquad [1]$$

f(x) is the sigmoid function chosen by the experimenter (i.e., logistic, Weibull, cumulative Gaussian). *β* determines the function's slope, whereas *α* determines the displacement of the function along the abscissa (see Figure 2). *γ* and *λ* are "psychological" parameters and they will be discussed shortly.

FIGURE 2 ABOUT HERE

Amongst the sigmoid functions, the logistic is the most widely used, because of its computational simplicity. Its formula is the following:

$$f(x) = \frac{1}{1+e^{\beta(\alpha-x)}} \qquad [2]$$

Therefore, the corresponding psychometric function Ψ (in the toolbox, the logistic psychometric function is the Logistic.m function) is:

$$\Psi = \gamma + (1-\lambda-\gamma)\left[\frac{1}{1+e^{\beta(\alpha-x)}}\right] \qquad [3]$$

In the logistic psychometric function, *α* (often referred to as midpoint) enables the displacement of the function along the stimulus level axis. It corresponds to the average between *γ* and *λ* [*i.e., Ψ(α)=(γ+λ)/2*]. *β* is the function slope, i.e., the rate of change in the subject's performance with stimulus level. The greater the absolute value of *β,* the steeper the psychometric function will be. Moreover, for positive values of *β* the function increases whereas for negative values it decreases (see Figure 2). *γ* and *λ* come into play when adapting the function to psychological needs. *γ* assumes a different meaning depending on the task type (i.e., *yes/no* or nAFC). There is, in fact, a major difference between these tasks: in *yes/no* tasks, the subject's response criterion is not under control of the experimenter, on the contrary, it is in nAFC tasks (Green & Swets, 1966; Stanislaw & Todorov, 1999). The reason of this difference is that in a *yes/no* task, when a *yes* response is collected for a very low stimulus level, it is difficult attributing this response to a high subject sensitivity or to a bias toward the *yes* response. Biased responses are called *false alarms* and they affect the lower limit of the psychometric function, which can assume values greater than zero (see Figure 1). In other words, the probability to get a *yes* response in absence of the stimulus is greater than zero (Green, 1993). Hence, in *yes/no* tasks, γ corresponds to the subjects' false alarm rate. False alarms are absent nAFC tasks (Green & Swets, 1966; Stanislaw & Todorov, 1999). In nAFC tasks, the level of the standard(s) is always different from the level of the variable and trials have, therefore, correct and incorrect answers. When the difference in level between standard and variable is below subject's threshold (i.e., "when the stimulus level is low") the probability that the subject returns a correct answer is determined by chance, and chance will depend on the number of alternatives. For this reason, in nAFC task, γ corresponds to chance level, i.e., the reciprocal of the number of alternatives (e.g., 50% for 2AFC, 33% for 3AFC, 25% for 4AFC and so on).

The meaning of λ, on the contrary, is independent of the task and refers to another error. In both yes/no and nAFC tasks, subjects could commit errors independent from the stimulus level; they are the lapses of attention. Lapses of attention are estimated to be a small percentage of the subject's responses (i.e., 1-5% Saberi & Green, 1997; Wichmann & Hill, 2001) and they can affect the psychometric function fitting by decreasing the upper limit of the function (see Wichmann & Hill, 2001 for an extended discussion). Attentional lapses are particularly problematic at high stimuli levels - in yes/no tasks - or high differences in the level of between the standard and variable(s) - in nAFC tasks (Saberi & Green, 1997; Wichmann & Hill, 2001).

Researchers are often interested in estimating a single point of the psychometric function, which is subject's threshold. In probabilistic terms, the threshold corresponds to an arbitrary point of the psychometric function *p* (hereafter referred as to p-target) included between the lower and the upper limit of the function (i.e., γ and λ). In other words, when we estimate a threshold, we search for the stimulus level eliciting the p-target proportion of yes (or correct) responses. Treutwein (1995) proposes that the p-target should be the middle of the psychometric function (e.g, 50% for yes/no tasks, 75% for 2AFC, 66% for 3AFC, etc.). However, other authors suggest that higher values should be targeted (e.g., Green, 1990; Baker & Rosen 1998, 2001; Amitay, Irwin, Hawkey, Cowan, & Moore, 2006).

Thresholds can be estimated by means of two classes of procedures: adaptive and non adaptive. In non-adaptive procedures, for example the constant stimuli method, the stimuli levels (or differences between standard and variable level) are preset before the beginning of the experiment. The stimuli should to span from below to above subject's threshold. During the experiment, the stimuli are presented to the subject in random order and the proportion of yes (or correct) responses is calculated for each stimulus. In other words, the subject's threshold will be interpolated from a fully-sampled psychometric function making the measurement of the threshold expensive in terms of experiment's time. This represents the major drawback of this class of procedures when the experimenter needs to estimate the subject's threshold only. For the above reason, when they need to estimate a threshold, psychophysicists prefer adaptive over non adaptive procedures. In adaptive procedures, the stimuli levels are selected at the same time as the experiment is running, depending on the subject answers. Adaptive procedures maximize the ratio between number of stimuli presented at/near threshold and number of stimuli presented far from threshold. Adaptive procedures can be grossly divided in two types: nonparametric (also known as staircases) and parametric. The only assumption made by non-parametric procedures is that the psychometric function is monotonic. Parametric procedures, on the contrary, make more assumptions. For example, they assume the shape of the psychometric function. Examples of non-parametric procedures are the method of limits by Fechner (1889), the simple up-down by von Békésy (1947) and the transformed up-down by Levitt (1971). Examples of parametric procedures are the PEST, by Taylor and Creelman, (1967), the “best” PEST by Pentland (1980) and the QUEST by Watson and Pelli, (1983). The ML procedure studied by Green (1990, 1993) is a parametric procedure. Although the paternity of maximum likelihood threshold estimation cannot be attributed directly to Green (Hall, 1968 and Pentland, 1980, for example, suggested this same approach), there are no doubts that he is the author that studied this approach in more detail (Green, 1990, 1993, 1995; Gu & Green, 1994; Saberi & Green, 1997).

Nonparametric procedures are generally more used than parametric ones, even if they encompass some disadvantage. The major one is that they tend to be [footnote 3] more time consuming (e.g., Amitay et al., 2006; Leek, 2001). Nonetheless, nonparametric procedures are more used than parametric ones because they are theoretically simpler and they can be easily implemented via conventional software (e.g., MEL, E-Prime), whilst parametric procedures are theoretically more complex and require more advanced programming skills. At the state of the art, we are aware of just one parametric procedure implemented in a freely downloadable version, namely the QUEST procedure (Watson & Pelli, 1983; Brainard, 1997; Pelli, 1997).

The Maximum-Likelihood procedure

The ML procedure is composed by two independent processes: the *maximum likelihood estimation* and the *stimulus selection policy* (in the toolbox, these two processes are synthesized in the FindThreshold.m function). The algorithm used by the ML procedure differs slightly for yes/no and nAFC tasks. Because the nAFC tasks can be used to estimate all types of thresholds, in the following paragraphs we will explain the ML procedure applied this task only. Readers interested in how the procedure works in yes/no task can address to the original work by Green (i.e., Green, 1993).

Maximum likelihood-estimation. Before the beginning of the experiment, the experimenter hypothesises several psychometric functions called hypotheses. The hypotheses have the same slope β, attentional lapse rate λ, and chance level γ, but differ in midpoint α so to cover the range of stimuli levels where subject's threshold is supposed to be (see Figure 3).

FIGURE 3 ABOUT HERE

The experiment might begin by providing the subject with a stimulus level that is above threshold. The subject response is then collected and utilized to calculate the likelihood of each hypothesis. Likelihood is calculated by means of the following function:

$$L(H_j) = \prod_{i=1}^{n} H^{C}(x_i)[1 - H(x_i)]^{W} \qquad [4]$$

where L(H$_j$) is the likelihood of the jth hypothesised function, *i* is the trial number. The exponents *C* and *W* are equal to, respectively, 1 and 0 when the response is correct and 0 and 1 otherwise. The product above can be simplified into a sum by means of a logarithmic transformation as follow (CalculateLikelihood.m in the toolbox):

$$L(H_j) = \sum_{i=1}^{n} C \log H(x_i) + W \log[1 - H(x_i)] \qquad [5]$$

Once the likelihood of each hypothesis has been calculated the ML selects the highest likelihood hypothesis. This hypothesis is that having the highest likelihood to resemble to the actual subject's psychometric function. The highest likelihood hypothesis will be identified by its midpoint α.

The likelihood of the hypotheses is calculated after each trial. Hence, even after the very first trial a maximum likelihood estimate is returned by the procedure, although it may be highly inaccurate. However, the more the trials, the more the estimate becomes accurate. Therefore, the best estimate is that returned by the last trial.

Stimulus selection policy. Once the most likely hypothesis has been found, at which level the next stimulus has to be presented? The common response is to assert that we should set the stimulus level at threshold (Simpson, 1989), i.e., at p-target. Even after the very fist trial, the ML procedure has enough information to select the threshold level that is used as the stimulus level for the successive trial. The most likely hypothesis contains also the most likely subject’s threshold. The threshold will be the inverse function of the most likely hypothesis at p-target (InvLogistic.m in the toolbox), therefore:

$$\Psi^{-1}(p_t) = \alpha_j - \frac{1}{\beta}\ln\left(\frac{1-\lambda-\gamma}{p_t-\gamma} - 1\right) \qquad [6]$$

where p_t is p-target.

Green (1990, 1993) showed analytically that there is an optimal p-target that researchers should track (as well as many p-target that experimenters should avoid, see guidelines section). This particular p-target (often referred to as sweetpoint) optimises the estimate of the subject’s threshold. This is because the variance associated to the estimate of this particular p-target is smaller than the variance associated to any other possible p-target. The variance of the threshold estimate associated to any p-target is equal to the binomial variance, $\Psi(1\text{-}\Psi)$, divided by the slope of the psychometric function squared, therefore:

$$\sigma^2 = \frac{\Psi(1-\Psi)}{\Psi'^2} \qquad [7]$$

where Ψ'^2 is the derivative of the psychometric function slope squared. The best possible p-target is that minimizing the above ratio. In the case of the logistic function the sweetpoint (p_{sw}) can be calculated analytically as follows (Green, 1993):

$$p_{sw} = \frac{2\gamma + 1 + \sqrt{1 + 8\gamma}}{3 + \sqrt{1 + 8\gamma}} \qquad [8]$$

If we assume that the subject does not produce any attentional lapse (and this assumption is convenient, see guidelines section) the sweetpoint depends exclusively on *γ*. Table 1 compare sweetpoints and p-targets calculated as the arithmetic mean between γ and λ, for some γ values. When γ increases the sweetpoint becomes correspondingly greater than the average between γ and λ.

TABLE 1 ABOUT HERE

Altogether, the ML procedure can be better understood with an example (see Figure 3 and Table 2). Let us suppose that we want to estimate a detection threshold by means of a 2AFC task. We select five hypotheses whose midpoints range from ~1.5 level units to ~7.5 level units so that the step between each midpoint is equal to 1.5 level units (see Figure 3). We choose to track the sweetpoint that, in 2AFC tasks, is the stimulus eliciting the 80.9% of correct responses (as reported in Table 1). We arbitrarily set the first variable stimulus level to 11 and present the stimuli to the subject. Let us suppose that subject’s answer is *correct* (Table 2, row 1, column 3). Each hypothesis expects, for that level, a certain proportion *p* of correct responses (Table 2, row 1, columns 4-8).

The likelihood of each hypothesis is calculated by means of equation (5) and is reported in Table 2, row 1, columns 9-13. As shown in Table 2, H1 results the highest likelihood hypothesis. By means of equation (6) we calculate the subject's threshold: we take H1 and calculate the stimulus level that corresponds to 80.9% of correct responses. The subject's threshold results to be equal to 2. Therefore, the stimulus level for the second trial will be set at 2. At the second trial, the subject gives a wrong response. Each hypothesis expects, for this stimulus level, a certain proportion 1-p of wrong responses. This proportion, together with that of the first trial, can be passed to equation (5). After the likelihood calculation, the hypothesis that most likely resembles the subjects' psychometric function is now H5. The stimulus level corresponding to the sweetpoint in H5 equals 8. The stimulus level for the third trial will be set at 8. The process just described is iterated until the subject has run the number of trials that we have set at the beginning of the experiment.

TABLE 2 ABOUT HERE

The major benefit of the ML procedure is that it makes maximal use of the available data: the data of all trials are used to estimate the subject's threshold. If the experimenter has set the procedure appropriately, ML will arrive at threshold more accurately than a staircase procedure. A further, advantage of the ML is its rapidity. Green (1993) claimed that twelve trials of ML are sufficient for a reliable threshold estimate. Although, recent evidences suggest that this initial claim was too optimistic (e.g., Leek, Dubno, He, & Ahlstrom, 2000; Amitay et al., 2006) the procedure still remains a rapid one. An additional advantage of the procedure is the possibility to track whichever point of the psychometric function, i.e., p-target. The reader, however, should know that this characteristic is shared also by some non-parametric procedure (e.g., Kaernbach, 1991). This characteristic is particularly appealing, for example, when the experimenter needs replicating the results of a study targeting a specific point of the psychometric function.

Besides its advantages, the ML procedure endures also some disadvantages. The major is that both the shape and slope of the hypothesised psychometric functions are set by the experimenter before running the experiment. These parameters could be unknown and the selected ones might not coincide with the "actual" ones (see Figure 3). This disadvantage, however, does not seem to affect significantly the experiment reliability. By means of several simulations, Green evaluated whether the mismatch between the "actual" subjects' psychometric function and the psychometric function hypothesised by the procedure affects the threshold estimate. From these simulations it emerged that (i) a shape mismatch [Green (1990) compared logistic and Gaussian] does not have evident effect on the threshold estimate; and that (ii) a slope mismatch can affect the threshold estimates by increasing the estimate variance (Green, 1990, 1993). By the same token, the experimenter may not know in advance the rate of attentional lapses (λ) or false alarm tendency (γ) of the subject. In three papers, Green and colleagues (1993, 1995; Gu & Green, 1994) investigated whether the lack of knowledge of these two parameters affects the threshold estimate and whether these two parameters can be estimated by ML procedure together with the subject's threshold [footnote 4]. Lapses of attention affect the threshold estimation when they are numerous and occur in the very first trials (Green, 1995). In particular, the bias produced by the lapses of attention is large if they occur within the first five trials but becomes almost negligible in the later ones (Gu & Green, 1994). Also false alarms can bias the threshold estimate (Green, 1993, Gu & Green, 1994). However, the ML procedure can be used to have a rough estimate the subject's false alarm rate. Solutions to the problems just presented will be given in the guidelines section.

The ML Toolbox

MLP has been developed to work with MATLAB 7.0 or higher [footnote 5] and can be downloaded from the following web page: http://www.psy.unipd.it/~grassi/mlp.html/. In the web page, the user will find the complete list of the toolbox's functions and the most updated list of the available experiments together with a brief description of them. At the moment we are writing, MLP is provided with twenty four built-in experiments. MLP works with any operative system and does not require additional MATLAB toolboxes. All MLP functions are compressed in a zip archive that the user needs to expand and copy into the MATLAB "toolbox" folder. The user needs also to "add with subfolder" all the toolbox contents in the MATLAB path. All functions have a command line help. The help can be seen typing "help" followed by the function name at the MATLAB prompt.

FIGURE 4 ABOUT HERE

To start an experiment type “mlp” (i.e., the main function) at the MATLAB prompt. This call visualizes a graphical interface (see Figure 4). Now, the user has to select the experiment s/he wants to run. The majority of built-in experiments are classic psychoacoustics experiment. Some are “translations” for the ML procedure of a subset of experiments performed by Kidd, Watson and Gygi (2007). These authors run nine-teen classic psychoacoustics experiments on a large (N=340) number of adult subjects. The MLP user running these experiments can thus compare his/hers’ own results with those reported in that study [footnote 6]. Built-in experiments come with default parameters. The user can, however, change parameters at will (see guidelines section). If changes are made and saved they will be kept until the next “SAVE DEFAULTS” command will be called. When the user presses “START” the experiment begins and the stimuli are presented to the subject. In all built-in experiments the subject responds by pressing the key-numbers of the computer keyboard. In nI-nAFC experiments, the subject reports the temporal position of the variable stimulus. For example, in a 4AFC task, if the subject thinks that the variable stimulus was the third stimulus presented s/he has to press “3”. In yes/no tasks, the “1” number corresponds to the answer “yes, I perceived/detect” and any other number (e.g, “0”) corresponds to the “no, I did not perceive/detect” answer. Key pressures must be followed by the “return” key. After each block of trials the subject’s threshold is echoed on the computer screen. MLP saves two data files (tab-delimited, flat format, text files) in the MATLAB current directory. The first is an extended data file that contains all experiments’ events: i.e., subject number, name, sex, age and note, block number, trial number, level of the stimulus presented, subject’s response, threshold estimated after each trial and estimated false alarm rate γ (please note this last estimate is for yes/no tasks only). In this file the subjects' responses are coded as "1" ("yes", or correct) and "0" ("no", or wrong). The name of this file can be set by the user through the graphical interface. The second data file contains only the subject's threshold and it is saved after each subject. By default, the name of this file is the subject's name. If the user do not input the subject's name the file is called "untitled.txt".

In the case the specifics of the built-in experiments do not match the experimenter's needs s/he can edit them and adapt them to his/hers own needs. The characteristics of the sounds of each experiments are written at the beginning of the experiment.m files and can be easily changed. More advanced MATLAB users can write their own experiments by take as example any of the built-in experiments. The MLP web page provides detailed instructions on how to write custom experiments.

ML procedure guidelines

In this section we provide the reader with a set of guidelines for a fruitful use of the ML procedure by means of MLP. Users that either want to edit existing experiments (or create their own) can use these guidelines for a fruitful optimisation of the threshold search.

Before starting the a detection threshold estimation, the first decision the experimenter has to take is about the kind of task subjects will perform (we remind to the reader that discrimination thresholds must be estimated via nAFC tasks only). The choice of the specific task depends on two factors: the desired experiment duration and the desired robustness of the threshold estimation. *yes/no* experiments are usually shorter. However, as previously mentioned, in *yes/no* tasks the subject’s bias is not under control of the experimenter. For this reason, thresholds gathered with yes/no tasks can be less robust than those gathered with nAFC tasks [footnote 7]. In brief, if duration is essential we suggest to use the *yes/no* task, whereas if robustness is essential, nAFC task should to be preferred. A corollary question is about the number of alternatives in a nAFC task. Once again, there is a trade off between duration and robustness. An increase in the number of alternatives leads to an augment robustness of the threshold estimation (Schlauch & Rose, 1990) but it augments also the experiment duration. For the ML procedure, Amitay et al. (2006) suggest using three alternatives.

Once the task has been chosen, the experimenter has to set the parameters of the ML procedure such as slope, range and number of hypothesis. As we wrote previously, the ML procedure uses, for the threshold estimation, a set of hypotheses (i.e., a set of psychometric functions) with identical slope, but differing in the position along the abscissa over a certain range. Ideally, the hypotheses slope should be identical to the subjects psychometric function slope, but occasionally this parameter could be unknown. The choice of the wrong slope can affect the threshold estimate, in particular when the chosen slope is (much) less steep than the actual one (Green, 1990, 1993). When the actual slope is unknown, the experimenter might need to estimate it (at least roughly) before starting the experiment. The classic method to perform this estimation is to run a constant stimuli experiment (e.g., Saberi & Green, 1997) and successively interpolate the subject’s performance with the logistic psychometric function.

The selection of range and number of hypotheses is less problematic than the selection of the hypotheses' slope. The hypotheses range must cover abundantly the range of stimuli levels where we expect the subject's threshold is. In practice, the midpoint of the first hypothesis should be well below subject's threshold whilst that one of the last hypothesis should be well above the subject's threshold. Green (1993) showed that the number of hypothesis set to fill this range does not affect the standard deviation of the threshold estimate. The number of hypotheses affects, however, how close will be the final threshold estimate to the actual threshold. For this reason, our suggestion is to use the highest possible number of hypotheses compatible with the computer potentiality. The computational load required by the procedure is in fact proportional to the number of hypothesis. Finally, the hypotheses range can be spaced either linearly or logarithmically. We recommend the use of logarithmic spacing for physical quantities expressed on a linear scale (e.g., frequency, duration, etc.) and linear spacing for units that are already expressed logarithmically (e.g., sound pressure level when expressed in decibels).

The next hypotheses' parameter we should set pertains to the yes/no task only. It is the false alarm rate γ. In the current version of the toolbox this parameter ranges between five possible fixed values: 0%, 10,%, 20,%, 30%, and 40%. The most likely of these γ values is returned after each trial in the MLP data file. If the subject has an high false alarm rate the threshold could be underestimated. In practice, the subject reports to be able to detect very low level stimuli. Consequently, the ML procedure will look for the subject's threshold also in the very low level range. However, the ML procedure returns only a rough estimate of the subject's false alarm rate. In particular, Green (1993) showed that the ML procedure underestimates the subject’s false alarm rate. The underestimate of the false alarm rate can be reduced by introducing a number of catch trials during the threshold estimate. In catch trials, the stimulus level is either set to zero or it is set to the minimum level of the range in which we are looking for the subject’s threshold (Gu & Green, 1994; Leek et al., 2000). The answers to catch trials (which is expected to be *no* unless the subjects is producing a biased response) is included in the calculation of the hypotheses’ likelihood. With MLP, the user has the option to include catch trials in the ML procedure to reduce the underestimate of the subject’s false alarm rate. Catch trials will be presented at any moment of the threshold estimation, excluding the very first trial. The occurrence of catch trials is determined probabilistically by a proportion that the user can set in the graphical interface. During the ML procedure, when a catch trial will occur, the stimulus level will be set to the minimum level of the range in which we are looking for the subject’s threshold (i.e., the first midpoint). We recommend to keep the catch trial rate to about 20% of the total number of trials (Leek et al. 2000).

After the selection of the parameters for the ML procedure, the parameter of for the stimulus selection policy (i.e., p-target) has to be chosen. The criterion to select p-target differs slightly between yes/no and nAFC tasks. In nAFC tasks, Treutwein (1995) suggested to track the middle of the psychometric function (i.e., 75% for 2AFC, 66% for 3AFC, etc.). As we wrote previously, Green (1990, 1993) suggested tracking the sweetpoint, which is generally higher than the middle of the psychometric function (see Table 1). Baker and Rosen (1998, 2001) and Amitay, et al. (2006) tracked p-targets even higher than the sweetpoint. As a rule of thumb, we suggest to track a p-target not lower than the sweetpoint.

The selection of the p-target for the yes/no task is slightly more complex. The general tendency is to track the middle of the psychometric function (i.e., 50%). However, this p-target could be too low, especially when working with subjects that are known to have a high false alarm rate (e.g., children). Differently from the nAFC tasks, in yes/no tasks the calculation of the sweetpoint is not straightforward, because the sweetpoint depends on the subject's false alarm rate (see equation 8), which is unknown. In synthesis, if we presume that subjects are reliable, we can track the middle of the psychometric function (i.e., 50%). On the contrary, if we presume that subjects are unreliable, we have to track a higher p-target. Green (1993) tracked the average between the minimum and the maximum sweetpoint of the false alarm range we expect to observe. In other words, if we expect a false alarm rate ranging from 0% to 40%, we will track 63.1%, i.e., the average between the sweetpoint for 0% false alarm rate (i.e., 50%) and the sweetpoint for 40% false alarm rate (i.e., 76.2%). This is the option we implemented in the toolbox.

Now that all parameters of the ML procedure are set, we have to decide the length of the experiment, i.e., the number of trials. As we wrote previously, one of the major advantages of ML method is its rapidity. Reliable threshold estimates can be obtained with very few trials. Green's (1993) showed that 12 trials may be enough. That initial claim was perhaps too optimistic. Recent studies suggest that an optimal threshold estimate should require about 24 (Leek et al., 2000) or about 30 trials (Amitay et al., 2006), that still remains a small figure [footnote 8].

The very last thing the experimenter has to decide before starting the experiment is the level of the first stimulus that the procedure will present to the subject. This level is the only set by the experimenter. Green (1993) demonstrated that the starting level value has no effect on the threshold estimate. We suggest to set this level relatively high, so to offer the subject an easy first trial.

After the ML procedure has terminated there is one last thing the experimenter can do to further control the goodness of the threshold estimates, that is controlling for attentional lapses. If the subject produces attentional lapses, the threshold is likely to be overestimated. The reason is simple, the subject tells s/he is not able to detect (or to discriminate) the very high level stimulus. Therefore, the ML procedure will look for the subject's threshold also in a very high level range. However, attentional lapses affect the threshold estimate, only if they occur within the first five trials (Gu & Green, 1994). Block of trials characterized by attentional lapses are easy to spot (and remove) in the data analysis. Figure 5 compares six blocks of trials. In five of these blocks the ML procedure is "affected" by an attentional lapse that occurred either in the first, second, third, fourth or fifth trial of the block.

FIGURE 5 ABOUT HERE

References

Amitay, S., Irwin, A., & Moore D. R. (2006). Discrimination learning induced by training with identical stimuli. *Nature Neuroscience, 9,* 1446-1448.

Amitay, S., Irwin, A., Hawkey, D. J. Cowan, J. A., & Moore, D. R. (2006). A comparison of adaptive procedures for rapid and reliable threshold assessment and training in naive listeners. *Journal of the Acoustical Society of America, 119,* 1616-1625.

Baker, R. J., & Rosen, S. (1998). Minimizing the boredom by maximising likelihood – Efficient estimation of masked threshold. *British Journal of Audiology, 32,* 104-105.

Baker, R. J., & Rosen, S. (2001). Evaluation of maximum likelihood threshold estimation with tone in noise masking. *British Journal of Audiology, 35,* 43-52.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10*, 433-436.

Fechner, G. T. (1889). *Elemente der Psychophysik* (2nd ed.), Leipzig: Breitkopf & Härtel.

Florentine, M., Buus, S., & Geng, W. (2000). Toward a clinical procedure for narrowband gap detection I: A psychophysical procedure. *Audiology, 39,* 161-167.

García-Pérez, M. A., & Alcalá-Quintana, R. (2005). Sampling plans for fitting psychometric function. *The Spanish Journal of Psychology, 8*, 256-289.

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

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

Green, D. M. (1995). Maximum-likelihood procedures and the inattentive observer. *Journal of the Acoustical Society of America, 97,* 3749-3760.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Gu, X., & Green, D. M. (1994). Further studies of a maximum likelihood yes-no procedure. *Journal of the Acoustical Society of America, 96,* 93-101.

Hall, J. L. (1968). Maximum-likelihood sequential procedure for estimation of psychometric functions. *Journal of the Acoustical Society of America, 44,* 370.

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

Kidd, G. R., Watson, C. S., & Gygi, B. (2007). Individual differences in auditory abilities. *Journal of the Acoustical Society of America, 122,* 418-435.

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

Leek, M. R., Dubno, J. R., He, N. J., & Ahlstrom, J. B. (2000). Experience with a yes–no single-interval maximum-likelihood procedure. *Journal of the Acoustical Society of America, 107,* 2674-2684.

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

Pelli, D. G. (1997) The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision 10,* 437-442.

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

Saberi, K., & Green, D. M. (1997). Evaluation of maximum-likelihood estimators in nonintensive auditory psychophysics. *Perception & Psychophysics, 59,* 867-876.

Schlauch, R. S., & Rose, R. M. (1990). Two-, three-, and four-interval forced-choice staircase procedures: Estimator bias and efficiency. *Journal of the Acoustical Society of America, 88,* 732-740.

Simpson, W. A. (1989). The step method: A new adaptive psychophysical procedure. *Perception & Psychophysics, 45,* 572-576.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31,* 137-149.

Stevens, S.S., Galanter, E.H. (1957). Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology, 54,* 377-411

Stevens, S.S. (1957). On the psychophysical law. *Psychological Review, 64,* 153-181.

Taylor, M. M., & Creelman, C. D. (1967). PEST: Efficient estimates on probability functions. *Journal of the Acoustical Society of America, 41,* 782-787.

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

von Bekesy, G. (1947). A new audiometer. *Acta Otolaryngology, 35,* 411-422.

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

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

Wright, B. A., Lombardino, L. J., King, W. M., Puranik, C. S., Leonard, L. M., & Merzenich, M. M. (1997). Deficits in auditory temporal and spectral resolution in language-impaired children. *Nature, 387,* 176-178.

Acknowledgements

Authors would like to thank Stephen Neely for the incipit provided to this work, the Padova lab-meeting for the many suggestions and one anonymous reviewer for constructive criticisms. Users of MLP that want to share their own experiments can send them to us. They will be uploaded in the MLP web page for public distribution.

Footnotes

1) Source: ISI Thompson Web of Science.

2) In this paper the term level refers to both intensive (e.g., luminance, sound pressure) and non intensive (e.g., light wavelengths, acoustic frequency) physical quantities. These two classes of physical quantities are perceptually mapped into the so-called prothetic sensations (e.g., brightness, loudness) as opposed to metathetic sensations (e.g., perceived color, and pitch) (Stevens & Galanter 1957, Stevens, 1957).

3) This is the case of the most widely used non parametric procedure (i.e., transformed up-down, Levitt, 1971) that requires about twice the trials of the ML procedure (Leek et al., 2000).

4) The ML algorithm can be used to estimate all psychometric function parameters (i.e., β, γ and λ), thus, the whole subject's psychometric function. However, the estimation of the whole psychometric function requires thousands of trials (García-Pérez & Alcalá-Quintana, 2005) and it is therefore performed exclusively in Monte Carlo simulations.

5) Unfortunately, some of the graphical characteristics of the toolbox are incompatible with older versions of MATLAB.

6) Readers interested in an identical replicate of the experiments run by Kidd, Watson & Gygi (2007), thus in a direct comparison with the published results, should address to the Test of Basic Auditory Capabilities (TBAC) by the same authors (Communication Disorders Technologies Inc.).

7) There are some exceptions to this. Green (1994) found that, in some conditions, the threshold estimates gathered with yes/no tasks were more robust than those gathered with nAFC tasks.

8) Literature reports a way to further shorten the length of experiments that is to halt the ML procedure when the variance of the threshold estimation drops below a certain preset value (Leek et al, 2000). As we wrote previously, the ML procedure returns a threshold estimate after each trial. In Leek et al. (2000) the ML procedure was halted when the variance of the last 10 threshold estimates was smaller than a certain preset value. This option is not implemented in the toolbox.

Table 1

Comparison between sweetpoint and the arithmetic mean between *γ* and *λ* for the logistic function.

γ	average(γ, λ)	sweetpoint
0%	50%	50%
25% (i.e., 4AFC)	62.5%	68.3%
33% (i.e., 3AFC)	66.6%	72.9%
50% (i.e., 2AFC)	75%	80.9%

Table 2

The possible development of a ML procedure. The table represents numerically the example reported in Figure 3. In the table, “trial” is the trial number, “level” is the difference in level between standard and variable in a detection task, “answer” is the subject’s answer (“c”, correct response, “w”, wrong response). Columns H1 to H5 show the probability of correct/wrong response predicted by each of the five hypotheses set for the ML procedure. The successive columns [i.e., from L(H1) to L(H5)] are the likelihood of each hypothesis given the subject’s answers to the stimuli levels presented thus far. The highest likelihood hypothesis is written in italics. The last column (i.e., “Th”) is the estimate of the subject’s threshold calculated at the end of each trial. In the current example, we track a p-target that is the sweetpoint of a logistic function for a 2AFC task (i.e., 80.9%).

Trial	level	answer	H1	H2	H3	H4	H5	L(H1)	L(H2)	L(H3)	L(H4)	L(H5)	Th
1	11	c	1.00	1.00	1.00	1.00	0.99	*0.00*	0.00	0.00	0.00	-0.02	2
2	2	w	0.19	0.37	0.46	0.49	0.50	-1.66	-1.00	-0.77	-0.71	*-0.71*	8
3	8	c	1.00	1.00	0.99	0.94	0.81	-1.66	-1.00	-0.79	*-0.78*	-0.92	6.5
4	6.5	c	1.00	0.99	0.94	0.81	0.63	-1.66	-1.02	*-0.85*	-0.99	-1.38	5
5	5	c	0.99	0.94	0.81	0.63	0.54	-1.67	-1.08	*-1.06*	-1.45	-2.00	5
6	5	c	0.01	0.06	0.19	0.37	0.46	-1.69	*-1.15*	-1.27	-1.90	-2.62	3.5
7	3.5	c	0.94	0.81	0.63	0.54	0.51	-1.75	*-1.36*	-1.73	-2.53	-3.30	3.5
8	3.5	w	0.06	0.19	0.37	0.46	0.49	-4.56	-3.01	*-2.73*	-3.30	-4.01	5
9	5	c	0.99	0.94	0.81	0.63	0.54	-4.57	-3.08	*-2.94*	-3.75	-4.63	5
10	5	w	0.01	0.06	0.19	0.37	0.46	-8.78	-5.88	*-4.60*	-4.76	-5.40	5
11	5	c	0.99	0.94	0.81	0.63	0.54	-8.79	-5.94	*-4.81*	-5.21	-6.02	5
12	5	c	0.99	0.94	0.81	0.63	0.54	-8.81	-6.00	*-5.02*	-5.67	-6.65	5

Figure Captions

Figure 1. Results of an hypothetical yes/no task. Subject's data are fitted with a logistic function (dashed curve). Note that this subject committed some false alarms (see later in the text) because at zero stimulus level (i.e., no stimulus actually presented) we can still observe a certain number of yes responses.

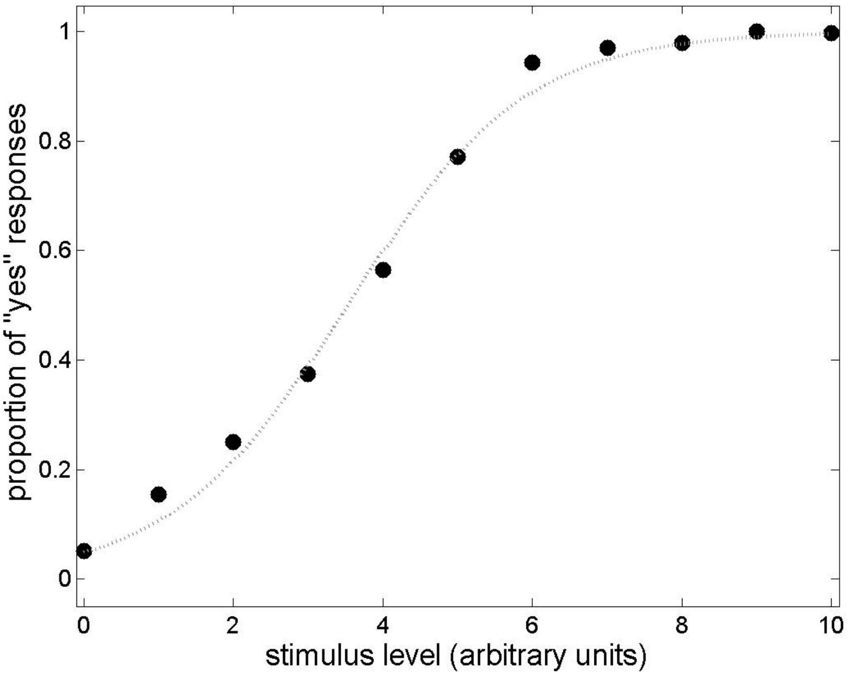
Figure 2. Five logistic psychometric functions. The grey (or black) functions have identical slope. The dashed (or dotted) functions have identical midpoint. The grey, solid function has a negative slope.

Figure 3. The solid grey function is the actual subject's psychometric function. The black dotted functions are the hypothesis we set to run the ML procedure and estimate the subject's threshold. Note that the actual subject's function is different from any of the hypothesised functions. However, within the hypothesised function, the middle one is the most similar to the subject's psychometric function.

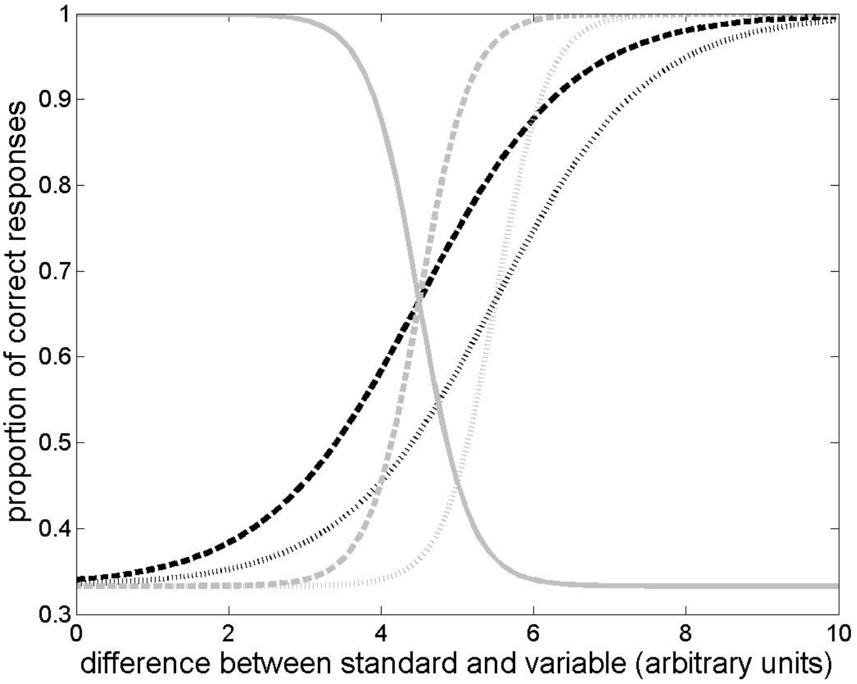
Figure 4. The MLP graphical interface.

Figure 5. Stimulus level presented by the ML procedure as a function of trial number. This figure shows the affect of attentional lapses in trials 1 to 5. The top left graph shows threshold estimation with no lapse. In the graphs, black symbols show yes (or correct) responses and white symbols show no (or wrong) responses. The dashed line shows the subject threshold.

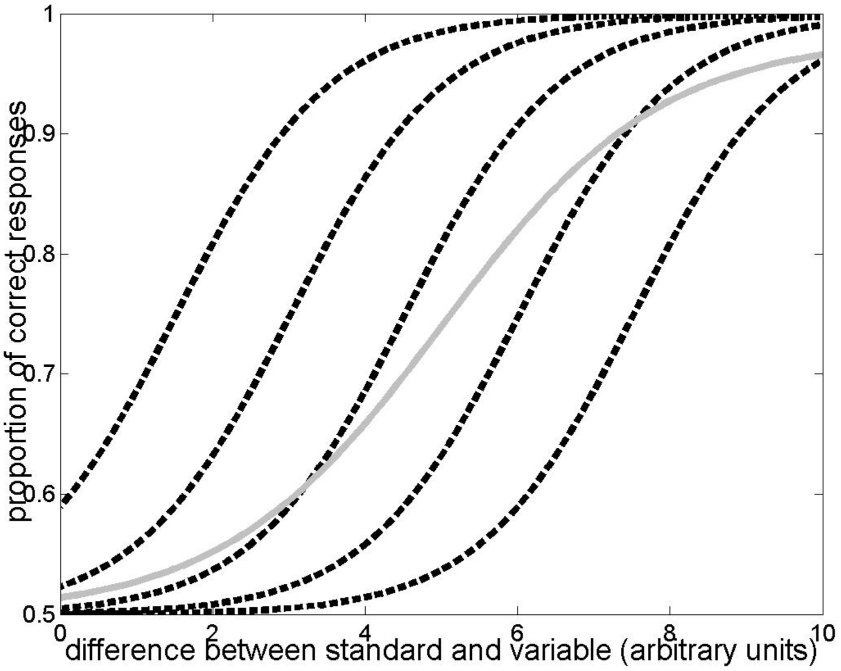
B411 - TOP


proportion of "yes" responses
0
0.2
0.4
0.6
0.8
1
0
2
4
6
8
10
stimulus level (arbitrary units)

B411 - TOP

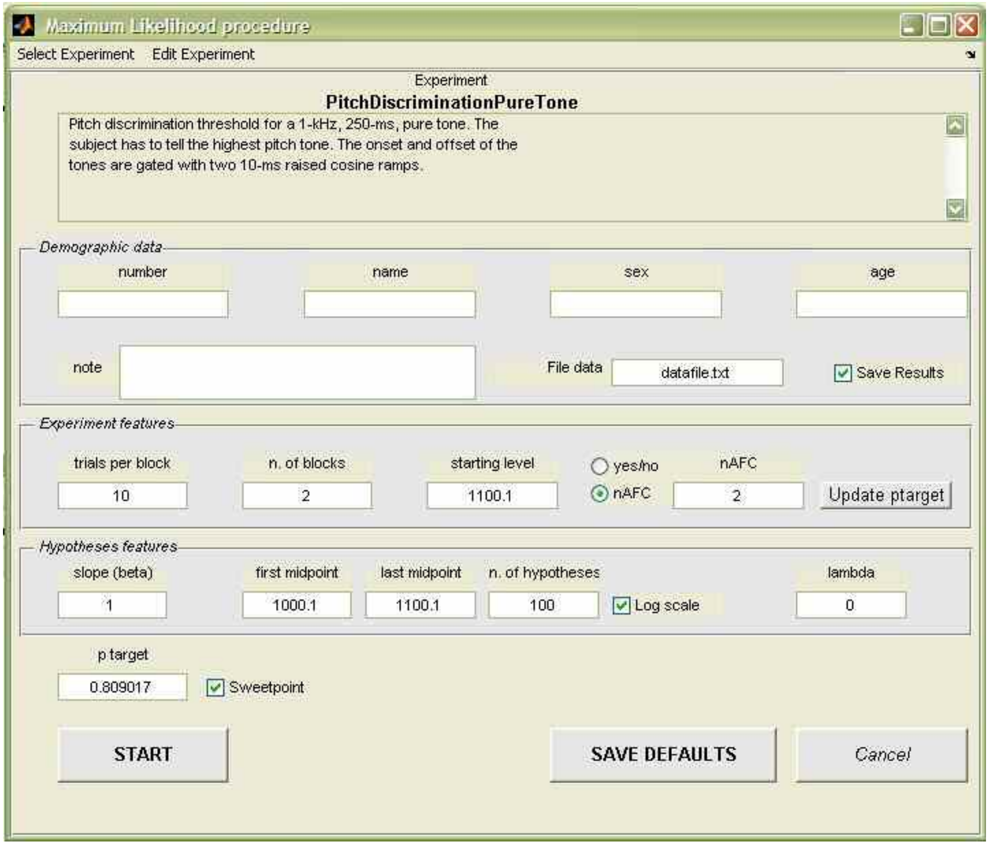

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
proportion of correct responses
0
2
4
6
8
10
difference between standard and variable (arbitrary units)

B411 - TOP



proportion of correct responses
1
0.9
0.8
0.7
0.6
0.5
0
2
4
6
8
10
difference between standard and variable (arbitrary units)

B411 - TOP



Experiment

PitchDiscriminationPureTone

Pitch discrimination threshold for a 1-kHz, 250-ms, pure tone. The subject has to tell the highest pitch tone. The onset and offset of the tones are gated with two 10-ms raised cosine ramps.

Demographic data

number	name	sex	age

note:

File data: datafile.txt

☑ Save Results

Experiment features

trials per block: 10

n. of blocks: 2

starting level: 1100.1

○ yes/no

● nAFC

nAFC: 2

Hypotheses features

slope (beta): 1

first midpoint: 1000.1

last midpoint: 1100.1

n. of hypotheses: 100

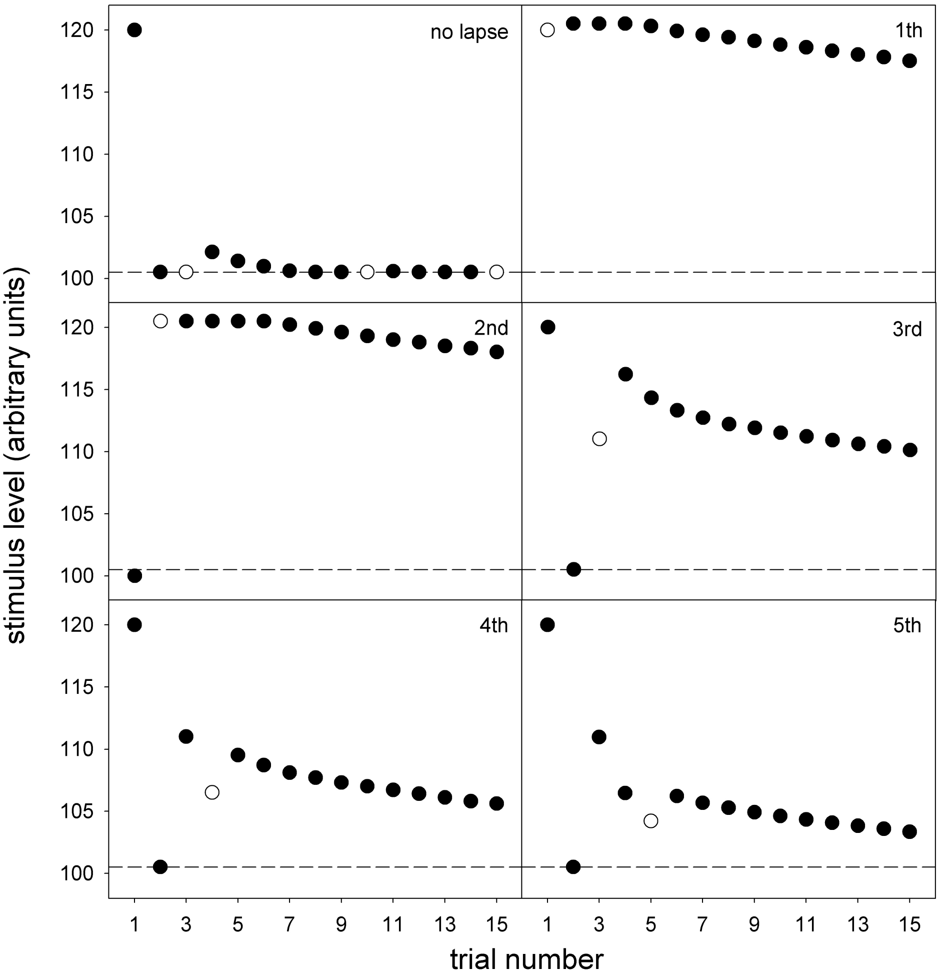
☑ Log scale

lambda: 0

p target: 0.809017

☑ Sweetpoint

B411 - TOP



no lapse
1th
2nd
3rd
4th
5th
120
115
110
105
100
stimulus level (arbitrary units)
1
3
5
7
9
11
13
15
trial number