Behav Res (2015) 47:384–394
DOI 10.3758/s13428-014-0478-7

Adaptive experiments with a multivariate Elo-type algorithm

Philipp Doebler · Mohsen Alavash · Carsten Giessing

Published online: 31 May 2014
© Psychonomic Society, Inc. 2014

Abstract The present article introduces the multivariate Elo-type algorithm (META), which is inspired by the Elo rating system, a tool for the measurement of the performance of chess players. The META is intended for adaptive experiments with correlated traits. The relationship of the META to other existing procedures is explained, and useful variants and modifications are discussed. The META was investigated within three simulation studies. The gain in efficiency of the univariate Elo-type algorithm was compared to standard univariate procedures; the impact of using correlational information in the META was quantified; and the adaptability to learning and fatigue was investigated. Our results show that the META is a powerful tool to efficiently control task performance in a short time period and to assess correlated traits. The R code of the simulations, the implementation of the META in MATLAB, and an example of how to use the META in the context of neuroscience are provided in supplemental materials.

Keywords Elo rating system · Computer adaptive testing · Psychophysics · Adaptive experiment

Electronic supplementary material The online version of this article (doi:10.3758/s13428-014-0478-7) contains supplementary material, which is available to authorized users.

P. Doebler (✉)
Department of Psychology and Sport Sciences, Westfälische Wilhelms-Universität, Fliednerstr. 21, 48149 Münster, Germany
e-mail: doebler@uni-muenster.de

M. Alavash · C. Giessing
Biological Psychology Lab, Department of Psychology, European Medical School, Carl von Ossietzky University, Oldenburg, Germany

M. Alavash
e-mail: mohsen.alavash@uni-oldenburg.de

C. Giessing
e-mail: carsten.giessing@uni-oldenburg.de

In many psychological and neuroscientific experiments, tasks—simple or complex—should match a subject's ability. Tasks that are too subtle or difficult waste time, as do tasks that are too easy or obvious. Adaptive procedures are useful in psychophysical applications, in which perception thresholds are to be determined, and in computer adaptive testing (CAT), in which tests are tailored to the examinee. Adaptive procedures can also be favorable for paired comparisons, and an example for the latter type of adaptive procedure is found in chess: here a player's performance is often assessed using the Elo rating system (ERS) so that tournament players can be matched on the basis of their Elo rating. All data for the computation of the Elo rating comes from matches against other players and after each match both players' ratings are adjusted.

Two main goals of all adaptive procedures are to *measure traits* (manifest or latent) or to create (experimental) conditions in which a subject shows a desired *performance*, often in the sense that the subject solves a given percentage of items.

The purpose of this article is to propose a new procedure that could be seen as a multivariate adaptive extension of Elo's algorithm, but that also borrows from many existing technologies. Some of these technologies are briefly reviewed so that relations to the new procedure become clear. The new procedure can be used to handle multidimensional (correlated) traits and is investigated by computer simulation. As a supplement to this article, we describe an experimental paradigm with two correlated traits that can be used to investigate fluctuations in the visual perception of near-threshold stimuli and provide MATLAB code. The supplement and the R code used in the simulations is available on the journal's homepage.

Elo's ranking system

Elo (1978) derived a rating system for chess players to estimate relative performance, which we briefly review as the

extension we propose shares a lot of features. In the so-called *Elo rating system* (ERS), each player has a rating that changes after each game. Players gain rating points if they win, and the amount gained is large if winning is unlikely according to the ratings before the game. In contrast, players' ratings decrease if they lose a game, and again the decrease is more severe if losing was unlikely according to the previous difference in ratings. Outcomes are coded as 0 for a loss, 0.5 for a draw, and 1 for a win.

One of the main advantages of the ERS is that players of similar strength can be matched on the basis of their ratings, so both players' chances of winning are predicted to be close to 50 %. Elo's system is well studied and understood, including its statistical properties (Batchelder & Bershad, 1979; Batchelder, Bershad, & Simpson, 1992; Glickman, 1995, 1999); it can be seen as a relative of Thurstonian models and Bradley–Terry–Luce models.

Before we outline how to implement the ERS, it will be instructive to discuss the underlying (implicit) statistical model. Let us assume that θ_1, $\theta_2 \in \mathbb{R}$ are the true ratings of two players. In this simple case, the outcomes are binary (0–1), and the result for Player 1 (using the coding above) is denoted by S_1. The key assumption is that a logistic curve determines the expectation of the outcome of a game. For example, the expected outcome of Player 1 is given by

$$P(S_1 = 1|\theta_1, \theta_2) = \mathrm{E}(S_1) = \frac{\exp(s(\theta_1 - \theta_2))}{1 + \exp(s(\theta_1 - \theta_2))}, \qquad (1)$$

where $s > 0$ is a scaling factor. This scaling factor is arbitrary[1] in applications with *latent* traits; also note that since the above expectation only depends on the differences between players' ratings, any number can be added to all ratings and the same model will result. Figure 1 shows the chance of winning a game according to Eq. 1.

In practice, only estimates of θ_1 and θ_2 are known, hence these estimates will be plugged into Eq. 1 to estimate $E(S_1)$. After the match, updated estimates of θ_1 and θ_2 are obtained using the following formula:

$$\widehat{\theta}_j^{\,\mathrm{new}} = \widehat{\theta}_j + K\big(S_j - \mathrm{E}\big(S_j\big)\big), \quad \text{where } j = 1, 2; \qquad (2)$$

here, $E(S_j)$ is computed from the current estimates $\widehat{\theta}_1$ and $\widehat{\theta}_2$, and $K > 0$ is a constant that determines how quickly ratings can change. The procedure outlined above adapts itself to changes in the strength of players; this can be justified intuitively, and is described in the more general results of Batchelder, Bershad, and Simpson (1992).

Recently, Klinkenberg, Straatemeier, and van der Maas (2011) described an application based on a variant of the ERS to the measurement and training of math ability in a computer based system. Here the items and the children are seen as players in the sense of the ERS, so that the ratings of items and children change over time (for the underlying psychometric model see Maris & van der Maas, 2012).

[1] Elo proposed to use the base 10 instead of e and a scaling factor of 400.

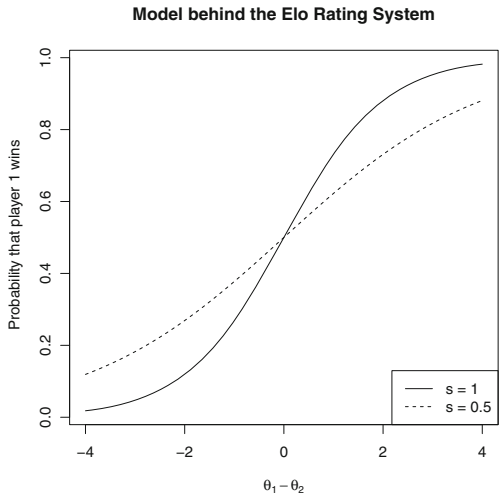

Fig. 1 Player 1's chance of winning is given by a logistic curve in the Elo rating system

Algorithms in psychophysics

Although the ERS matches players so that the chance of winning is about 50 %, psychophysical researchers match stimuli to subjects so that a certain performance is observed. Typically visual, auditory, haptic, tactile, or olfactory perception tasks are used to learn about perception thresholds—for example, what stimulus level can be detected at a specific performance level. In many cases researchers also want to learn about the underlying *psychometric function*, which maps the stimulus level to the probability of a correct answer. A range of adaptive procedures are used in psychophysics and the reader is pointed to the excellent reviews of Treutwein (1995) and Leek (2001) for comprehensive treatments. We briefly discuss the staircase and maximum likelihood methods though, as both serve as benchmarks procedures in the following.

The most simple staircase procedure works by increasing the stimulus level if a subject fails to detect the stimulus, and decreasing it otherwise. The answer is not forced and the step sizes of these decreases and increases are often fixed. In the long run, the performance of a subject will be around 50 % using this method. More complicated staircase methods exist, which lead to other targeted performances (Brown, 1996; Levitt, 1971). This method does not assume any specific psychometric function, although given enough trials, it is possible to determine some of the points of the underlying psychometric function.

Likelihood based approaches have to assume a psychometric function, often a logistic curve. Based on the psychometric function, a likelihood function for the observed responses can be calculated at any point of the procedure. This likelihood is then used to estimate the stimulus level at which the subject performs at the targeted level. Many statistical methods can be used when a likelihood is available, and maximum likelihood (ML) is probably the most well-known and the most used in psychophysics. Although likelihood based approaches are often very efficient, they crucially depend on the correct specification of the underlying psychometric function. Moreover, typical implementations of all mentioned approaches focus on measurement and thus do not take into account learning or fatigue effects during the course of an experiment—that is, changing subject traits (Leek, 2001).

Computer adaptive testing using item response theory

The algorithm we present can also be applied to latent traits; hence, we review computer adaptive testing (CAT; van der Linden & Glas, 2000, 2010; Wainer et al., 2000) so that the relationship to this state of the art technology with its information theoretic rationale becomes apparent. CAT tailors a test to an examinee. Relative to an examinee's performance, appropriate items from an item pool are presented. This performance is measured in terms of a (provisional) estimate of the person's latent ability parameter that is updated after an item is administered. CAT builds on item response theory (IRT; Baker & Kim, 2004; Embretson & Reise, 2000; Hambleton, Rogers, & Swaminathan, 1995; van der Linden & Hambleton, 1997) that uses concepts from information theory for item selection. A typical model is the two parameter logistic model (2PL; also called the Birnbaum model): Here, the probability that a person answers item i correctly is given by

$$P(X_i = 1|\theta) = \frac{\exp(\alpha_i(\theta - \beta_i))}{1 + \exp(\alpha_i(\theta - \beta_i))}, \tag{3}$$

where θ is the person parameter, α_i the ith item discrimination parameter, and β_i the difficulty parameter. Every item can be described by its item characteristic function (ICC), which is similar in concept to the psychometric functions described above; see also Fig. 2. The difference is that ICCs relate *latent traits* to observed behavior, whereas psychometric functions use *physical stimulus levels* to predict behavior.

Using likelihood methods, a person parameter is estimated (van der Linden & Hambleton, 1997). On the basis of the current estimate $\widehat{\theta}$ for the person, one now chooses the next item–for example, using the Fisher information criterion (Van der Linden & Pashley, 2010). For the 2PL model, the Fisher information for the ith item (given a person parameter θ) is

$$I(\theta) = \alpha_i^2 \frac{\exp(\alpha_i(\theta - \beta_i))}{(1 + \exp(\alpha_i(\theta - \beta_i)))^2}. \tag{4}$$

When the Fisher information criterion is used to select the next item, the Fisher information of all (unused) items in the item pool is evaluated at the current estimate $\widehat{\theta}$ and the item with the maximal Fisher information is chosen as the next item. The motivation for this approach is that the standard error of the current estimate is inversely proportional to the square root of the total information (Birnbaum, 1974). So the adaptive procedure leads to smaller standard errors than does a linear test of the same length. Typically, no specific performance is targeted, but the goal is to produce the most efficient measurement of the latent trait. Since the information function in the 2PL model peaks at the item difficulty parameter, the performance is often close to 50 %, especially if discrimination parameters are similar and the item pool covers a wide range of the latent scale. Other approaches in CAT, especially in educational applications in which frustration is an issue, control target performance by different item selection strategies (Eggen & Verschoor, 2006).

The drawback of this method is that setting up a CAT procedure involves much effort: Item design needs to make sure that the assumptions of the underlying IRT models are met. After this, nonadaptive testing is used to calibrate an item pool, typically requiring large samples for precise estimates of item parameters in the complete pool; also items that do not conform to the model have to be discarded in this phase. Finally, the adaptive procedure can really begin, typically by assuming that the estimated item parameters are the true ones.

A multivariate Elo-type algorithm

In the following section, we will speak of items, and in this way refer to any type of stimulus (simple or complex) in a setting in which a binary (0–1) answer is expected. Let us assume that M traits are of interest and that items are available to measure each of the traits. Also, we assume that the items are *pure*, in the sense that only one of the traits matters in the answering process. We will talk of success and code it with 1 if the person solves the item–that is, detects a stimulus, pushes the correct button, or answers a question correctly–and code failures with 0. We write X_{ji} for the answer of person j to item i, and assume I items and J persons. We use θ_j for the jth person parameter and since we measure M traits, the person parameter is M-dimensional, so that

$$\theta_j = (\theta_{j1}, \ldots, \theta_{jM}) \in \mathbb{R}^M.$$

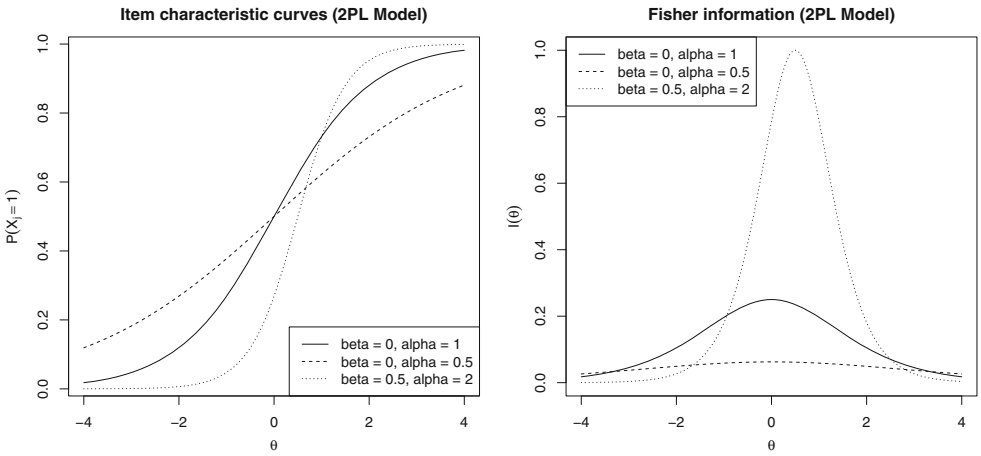

Fig. 2 Item characteristic curves of the 2PL model, and the corresponding Fisher information curves

We use β_i for the ith item parameter and since we assume items are pure, they are one-dimensional; also assume that the ith item is designed to measure the $m(i)$th trait. Our model can then be formulated in terms of the expected rate of successes:

$$\mathrm{E}(X_{ji}) = \frac{\exp\left(s\left(\theta_{jm(i)} - \beta_i\right)\right)}{1 + \exp\left(s\left(\theta_{jm(i)} - \beta_i\right)\right)}, \tag{5}$$

where s is a scaling parameter. Note that only the $m(i)$th component of θ is involved in the answering process. In addition, we assume that information on the association of the M traits is available (say, from previous research) in the form of correlations for the *population* of θ vectors; the correlation between the nth and mth trait is denoted by ρ_{nm}. In cases in which item and person parameters are latent, one can fix the scaling parameter s at 1; in other cases, the choice of s needs to be based on previous research.

There are two cases: The item parameters reflect a (known) physical quantity (say, a [log-]intensity of light, a contrast, or a sound pressure level)—in this case, we assume that the β parameters are determined by (or maybe even equal to) the physical quantities. The second case is that the item parameters are latent variables reflecting an item's difficulty. The following algorithm requires no update for the item parameters in the first case, nor does it need to be updated in the second case when estimates of the item parameters are known from previous research and can be regarded as fixed. If they cannot be regarded as fixed, an update for the item parameters is needed, which we discuss later. In both cases, each person has a (current) uncertainty ($0 \le U_j \le 1$).

We propose the following algorithm, called *META* (multivariate Elo-type algorithm), to adapt to the jth person's traits:

1. Choose an initial estimate of $\widehat{\theta}_j$, for example at a population mean (if known). Set the initial uncertainty U_j to 1.
2. Cycling through the M traits, do the following:

Item Selection: If the current trait is m we choose from items with $m(i) = m$. Among these choose the item that minimizes $|\theta_{jm} - \beta_i|$; here, θ_{jm} is the mth component of the M-dimensional person parameter θ_j, and β_i is the ith item parameter.
Present Item: Observe X_{ji}, the binary score of person j on item i.
Person Update: Calculate

$$K_j = K\left(1 + kU_j\right),$$

and set

$$\theta_{jm}^{\mathrm{new}} = \theta_{jm} + K_j\left(X_{ji} - E\left(X_{ji}\right)\right).$$

For all other components n of θ_j, use the following update:

$$\theta_{jn}^{\mathrm{new}} = \theta_{jn} + K_{\mathrm{other}} K_j \rho_{nm}\left(X_{ji} - E\left(X_{ji}\right)\right), \quad 1 \le n \le M, n \ne m,$$

here ρ_{nm} is the correlation of θ_{jm} and θ_{jn} on the population level.

Reduce Uncertainty: Set

$$U_j^{\mathrm{new}} = \max\left(U_j - u, 0\right).$$

3. Terminate the adaptive procedure once a desired number of items have been presented.

A few remarks are in order: (1) The algorithm requires some (up to a certain point arbitrary) constants: K determines the variability of the estimates, and k determines the influence of the uncertainty. K_{other} moderates the influence of the measurements of one trait on the other traits and $u > 0$ is the reduction in uncertainty after each update of the person parameter. These constants should be determined by computer simulations and small pilot experiments. (2) In each step, the algorithm focuses on updating one current component of the

person parameter (denoted θ_{jm} above) but the other components are also updated. The change in the other components is in the same direction but smaller than for the current component by a factor of $K_{\text{other}}\rho_{nm}$. (3) In the context of paired comparisons, including uncertainty into an Elo-type algorithm is also supported by Glickman (1999). The uncertainty can be seen as a quantification of the trust in the current estimate. If the META is used on two occasions, one could use the last estimate from the first experiment as the initial value for θ_j in the second, but increase the uncertainty. (4) For $K_{\text{other}} = 0$ (or $\rho_{nm} = 0$ for $m \neq n$) the measurements of one trait do not influence the measurements of the others, so M independent algorithms are run in this case. (5) If the traits at hand are at least moderately correlated, the algorithm's convergence speeds up, since it borrows information from the performance on the other dimensions. (6) The META inherits the ability to adapt to changing person parameters from the Elo algorithm; that is, it handles fatigue and/or learning. The current estimate is a kind of running estimate and is calculated from the last estimate and last response. Thus, earlier responses contribute to the estimation of current ability because they influenced the last estimate. (7) Estimates never converge to one single value. This reflects experimental conditions that do not necessarily lead to a converging performance. If convergence can nevertheless be assumed, it can be judged by several criteria (see the simulations below). (8) Although some assumptions are made with respect to the probability of a correct response, the META is not a model for change in person parameters, but a tool to keep experimental conditions constant. (9) The above item selection criterion can be seen to maximize the Fisher information for the current dimension; that is, $I(\theta)$ in Eq. 4 is maximal.

Variants of the proposed algorithm

Here, we describe some obvious and some less obvious modifications of the algorithm:

The proposed algorithm is independent of the model in the sense that other models can be used to estimate $E(X_{ji})$; especially, other item characteristic curves can be used, for example to incorporate guessing. Guessing can be handled by a (pseudo)guessing parameter c and the expectation of a correct answer becomes

$$\mathrm{E}\left(X_{ji}\right) = c + (1-c)\frac{\exp\left(s\left(\theta_{jm(i)}-\beta_i\right)\right)}{1+\exp\left(s\left(\theta_{jm(i)}-\beta_i\right)\right)},$$

where, for example, in a 2AFC paradigm, c is equal to 0.5.

The above algorithm matches items and persons on the basis of roughly a 50 % chance of success [when the item parameter is close to the person parameter, then the difference is close to 0, and hence Eq. 5 implies that $E(X_{ji})$ is 0.5]. If the desired chance of success is p, then Eq. 5 implies that β_i should be chosen to minimize

$$|\beta_i-(\theta_{jm(i)}-s^{-1}\ln\frac{p}{(1-p)})|.$$

The above algorithm does not restrict the reuse of any item. For certain applications it might be necessary to exclude items from the pool if they have been presented once, or at least wait for a while until presenting them again.

Klinkenberg et al. (2011) suggest to incorporate item uncertainty and updates for the item parameters as well. We can only recommend this modification if the item parameters are seen as latent variables. The update is similar to the person update:

$$\beta_i^{new} = \beta_i + K_i\left(E\left(X_{ji}\right)-X_{ji}\right), \text{ where } \quad K_i = K(1+kU_i).$$

Note that this small modification of the algorithm is a big step in practice: the items should be calibrated in a pilot study before testing begins when using this variant.

In the above case, one can add a tuning parameter k^*, $0 < k^* < 1$, and calculate

$$K_j = K\left(1+kU_j-k^*U_i\right) \quad \text{and} \quad K_i = K\left(1+kU_i-k^*U_j\right).$$

This tuning parameter takes item uncertainty in the person parameter update and person uncertainty in the item parameter update into account.

A possible way to force the stabilization of the estimates is to reduce K with time; so, for example, once M steps of the above algorithm have been completed, K is shrunk by 1 %. This modification can be dangerous, because estimates might stabilize too early.

One can also consider to shrink the K_{other} parameter in the same fashion: once rough estimates have been obtained, one can focus on estimation of individual traits. This is not dangerous, since even if the shrinking is aggressive (e.g., by a factor of 0.5), the algorithm will not stabilize, but the algorithm might be less efficient, as shown by the computer simulations below.

If the algorithm is used as a measurement device, one can terminate after the estimates for θ_j have become reasonably stable. Defining stability is an obvious challenge here; one way would be to use the rules terminating (transformed) staircase procedures (Levitt, 1971).

One could speculate that an extension in the spirit of multidimensional IRT models is possible; toward this end, one could discard the assumption of pure items and assumes that more than one latent trait is involved instead. One would then use multidimensional versions of the expectation in Eq. 5 (Reckase, 2009). However, a reasonable update for the person

parameter must then be based on the loadings for the different components of the person parameter. Also, item calibration usually requires large samples in the context of multidimensional IRT, putting in question the feasibility of such an approach.

The META from other perspectives

From the viewpoint of the Elo rating system, the META is an extension to correlated traits. From the perspective of psychophysics it could be described as a multivariate hybrid procedure that is similar to the simple staircase procedure, since it moves up if the subject answers correctly and down otherwise. In contrast to this the META features (diminishing) step sizes based on a likelihood approach. From a CAT viewpoint the META uses a naive estimator for the person parameter paired with a special multivariate 2PL model in which the discrimination parameter is fixed at $1/s$ for all items. In addition, all items are supposed to be pure—that is, each item only loads on one of the dimensions. If the ability is latent—that is, if there is no physical interpretation of θ—then s can be set to 1. In this case the underlying model of the META for each of the components of θ is a Rasch model. A univariate version of the META has been described as a relative of the Kalman filter (Brinkhuis & Maris, 2009). In this sense, the META could be seen as a Bayesian filter, as prior information on the population covariance is used in the algorithm, although it is not a Bayesian estimator.

Efficiency, adaptability, and robustness of Elo-type algorithms

The proposed algorithm might have intuitive appeal, but we nevertheless explored the algorithm with the help of computer simulation and investigated its robustness to the misspecification of the psychometric function analytically. The simulations addressed three questions: (1) Is an Elo-type algorithm (ETA) better suited to create an experimental setting in which subjects show a desired performance than simple staircase procedures from psychophysics? (2) How large is the gain in efficiency of the META, as compared to separate ETA algorithms? (3) To what extent can the META algorithm react to changes in the underlying person parameters? This last question is aimed at effects of within-test learning and fatigue.

First simulation experiment In the first simulation experiment, the performance of the univariate ETA was compared to two reference algorithms, a simple staircase procedure and a ML procedure. For this we simulated experimental runs aiming at a performance of 50 % with each of the three methods.

The true person parameters were sampled from a standard normal distribution and we assumed that items of arbitrary difficulty were available. In all cases a logistic psychometric function with $s = 1$ was used to model the behavior of a subject and the initial estimate of θ was 0. We use the notation from above to describe the procedures, but since the number of traits is $M = 1$, we omit the m subscripts. The ETA was run with the constants $K = 0.1$, $u = 1/40$, $k = 8$. The simple staircase used the update

$$\theta_j^{\text{new}} = \theta_j + K\left(2X_{ji} - 1\right), \tag{6}$$

and thus, θ was either decreased or increased by 0.1. The difficulty of the next item was then set to the new θ_j, so we assumed that items/stimuli were available that matched all estimated person parameters exactly. This might not be a realistic assumption as only a limited number of stimuli might be available in practice. However, as we are interested in relative efficiency gains in this simulation experiment and the following experiment, no systematic effects are expected.

After the first item the ML procedure used the following update for the estimate of θ (and, hence, to obtain the next difficulty β_i): Let $\vec{X}$, $\vec{\beta}$ denote the vectors of X_{ji} and β_i so far. Then, θ is set to the value maximizing the likelihood

$$L\left(\theta \middle| \vec{X}, \vec{\beta}\right) = \prod_i \left(1 - p_{ji}\right)^{1-X_{ji}} p_{ji}^{X_{ji}}, \tag{7}$$

where p_{ji} is given by Eq. 5. This is the ML estimate of θ. Note that if all items were answered correctly, the ML estimate is ∞. In such a case, the next θ was set to 6 (i.e., 6 standard deviations above the mean of the person parameters that were sampled from a standard normal) and if no items were answered correctly, the next θ was set to –6.

The simulated experiments were conducted for 500 updates of the person parameter. The resulting series of answers X_i, $i = 1, \ldots, 500$ ($X_i = 1$ for a correct answer, and 0 otherwise), and person parameters were then inspected for convergence according to four different criteria. Thereby, the number of items that were administered until convergence was achieved gave us a measure of efficiency for each criterion:

(omniscient) Convergence was assumed when the estimated person parameter was less than 0.1 away from the true person parameter.

(median) The median $\widehat{\theta}_{med}$ of all 500 person parameter estimates was determined and convergence was assumed when the estimated person parameter was less than 0.1 away from $\widehat{\theta}_{med}$.

(pingpong) Short series of six subsequent answers were taken and convergence was assumed when such a series contained three zeroes and three ones for the first time. The index of the last answer of this series was recorded.

(reversals) Reversals were determined–that is, indices with $X_i \neq X_{i+1}$. The mean of all person parameter estimates at these reversals was then calculated ($\widehat{\theta}_{rev}$) and convergence was assumed when the estimated person parameter was less than 0.1 away from $\widehat{\theta}_{rev}$.

All three algorithms are relatively stable once the estimated person parameter is close to the true parameter, which motivates why each criterion could be seen as appropriate to judge the efficiency. The practical value of some criteria might be limited, though. We replicated the experiment 10,000 times. The simulation was implemented in R (R Development Core Team, 2012).

Table 1 displays descriptive statistics of the length of the experiment until convergence–that is, the number of updates–and Fig. 3 presents part of the results graphically. Note that, apart from the pingpong criterion, which underestimates the length needed for convergence, all criteria lead to similar distributions of lengths. The first observation is that, regardless of the criterion, all three algorithms show greater mean than median lengths–and indeed all algorithms produce outliers–and lengths are skewed to the right. Even if we remove the top 1 % of observations before calculating the mean, this relationship does not change. In the case of the ML procedure, the estimates for θ stabilize relatively quickly (in the sense that the standard error of measurement after 30 items is ~0.2 if $s = 1$), so that additional observations only contribute little to the estimate. If the initial simulated behavior is not representative of the true θ, it takes many additional items to get close to the true θ. Also, the ETA and the staircase algorithm produce long simulated runs, but as the 95 % quantile of length until termination in Table 1 shows, these are less extreme than for the ML algorithm. The staircase algorithm needs many more items to adapt the experimental conditions than does the ETA; this is not surprising, since the variable step size of the ETA allows it to get close to the targeted performance level earlier. One should be cautious not to overemphasize the very low values for the first quartile of the distributions: After four and six items, respectively, the step sizes of the ML and ETA algorithms are still quite large, so that there is substantial variability of θ estimates. We note that conclusions drawn from this simulation experiment and the two subsequent ones are limited by the fact that some convergence criteria were artificial, and also by the assumption of a known psychometric function. Taking everything into account, the simulations indicate a clear preference for the ETA procedure when a targeted performance is to be reached.

Second simulation experiment The second simulation experiment compared the META for M traits to M univariate Elo-type algorithms. The latter procedure was obtained from the META by setting $K_{other} = 0$; since K_{other} controls the update for the other dimensions, a value of 0 leads to M independent algorithms. We varied the number of traits ($M = 2, 3, 4$), the pairwise correlation ($\rho = 0.5, 0.8$), and we also analyzed whether it makes sense to shrink the K_{other} parameter after each step of the algorithm (shrinking by factors 1 [no shrinking], 0.95, and 0.9). We used a full factorial design. Person parameters were sampled from an M-dimensional multivariate normal distribution using a covariance matrix, with 1 s on the diagonal and ρ as the off-diagonal elements. The tuning parameter had the following values for the META and the univariate algorithms: $K = 0.1$, $k = 8$, initial $K_{other} = 0.5$, $u = 1/40$.

For each of 10,000 replications, we sampled a person parameter and ran the META and the M univariate Elo-type algorithms

Table 1 Comparison of distributions of length until convergence of various univariate adaptive algorithms

Criterion	Algorithm	Quantiles				Mean		*SD*		Skewness	
		25 %	50 %	75 %	95 %	Reg	Trim	Reg	Trim	Reg	Trim
Omniscient	Staircase	8.0	22.0	38.0	68.0	26.22	25.28	22.07	20.22	1.24	0.85
	ETA	4.0	11.0	20.0	55.0	16.20	14.83	21.62	16.60	4.20	2.82
	ML	19.0	32.0	54.0	114.0	42.21	40.46	36.33	31.79	2.10	1.39
Median	Staircase	7.0	20.0	34.0	61.0	23.43	22.66	19.55	18.10	1.19	0.87
	ETA	4.0	11.0	20.0	45.0	15.37	14.19	18.99	14.84	3.78	2.59
	ML	19.0	28.0	52.0	110.0	40.38	38.83	34.11	30.46	1.82	1.34
Pingpong	Staircase	7.0	9.0	14.0	25.0	11.55	11.26	6.50	5.87	1.73	1.39
	ETA	6.0	7.0	10.0	13.0	8.22	8.11	2.59	2.37	1.50	1.18
	ML	10.0	10.0	11.0	12.0	10.34	10.29	1.03	0.95	0.61	0.02
Reversals	Staircase	8.0	21.0	35.0	60.0	23.91	23.13	19.45	18.04	1.05	0.73
	ETA	4.0	11.0	20.0	37.0	14.51	13.53	16.21	13.00	3.30	2.19
	ML	19.0	32.0	55.0	118.0	42.78	41.02	36.74	32.27	2.05	1.37

SD, standard deviation; Reg, regular statistic; Trim, trimmed statistic (top 1 % of data were discarded before calculating the statistic); ETA, univariate Elo-type algorithm; ML, maximum likelihood algorithm

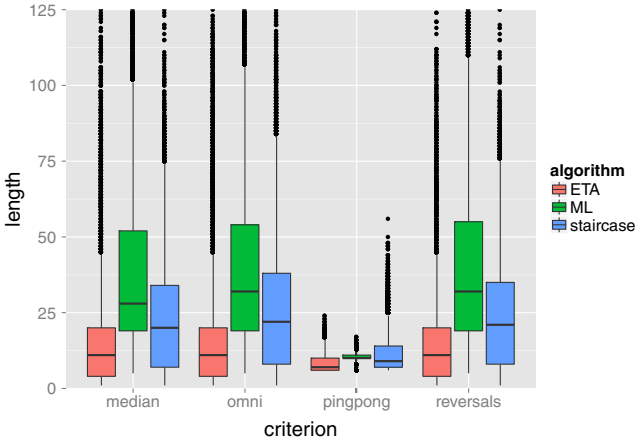

Fig. 3 Box-and-whiskers plots for the length of the experiment until convergence for the different univariate algorithms. The large number of outliers is due to the 10,000 replications of the simulation experiment

starting at $\theta = (0, \ldots, 0)$, the M-dimensional vector of zeroes. We used the omniscient convergence criterion as in the first experiment as it behaved very similar to the other criteria.

Table 2 shows the resulting mean and median test lengths, and box-and-whiskers plots for the cases without shrinking K_{other} are displayed in Fig. 4. The gain in efficiency is substantial in all cases, and the most striking results are obtained for median test lengths. Even when only studying moderate correlations of 0.5 and means, one obtains at least a gain in efficiency of 36 % in two dimensions, and 27 % in four. Table 2 also records the standard errors of the lengths, all of which are large in comparison to the mean and median lengths. This suggests that some person parameters produce substantial outliers in the distribution of lengths. Shrinking K_{other} cannot be recommended in the context of these simulations.

We finally note that badly chosen tuning parameters can substantially decrease the efficiency of any Elo-type algorithm, and that some tuning parameters would probably lead to even fewer trials in this (and also the following) experiment. However, no systematic search for the most efficient set of tuning parameters was performed.

Third simulation experiment In the third simulation experiment, we analyzed how quickly the META could adjust to changes in person parameters. We used the same tuning parameters, termination criterion, item/stimulus parameters, and way of sampling person parameters as in the second experiment. In addition, we changed the

Table 2 Gain in efficiency of META as compared with univariate Elo-type algorithm

Factors			Mean Length			Median Length			*SE* Length	
M	ρ	sf	META	ETA	*r*	META	ETA	*r*	META	ETA
2	.5	1.00	41.4	64.5	.64	16.0	27.0	.59	63.5	74.9
3	.5	1.00	111.2	161.3	.69	38.0	146.0	.26	125.1	132.8
4	.5	1.00	212.3	291.0	.73	194.0	274.5	.71	181.7	181.7
2	.8	1.00	31.1	65.2	.48	11.0	27.0	.41	55.1	75.4
3	.8	1.00	78.7	159.6	.49	21.0	143.0	.15	109.4	134.0
4	.8	1.00	138.4	280.5	.49	38.0	267.0	.14	165.2	181.0
2	.5	.95	47.1	65.0	.72	18.0	27.0	.67	66.9	74.4
3	.5	.95	122.7	161.8	.76	72.5	147.0	.49	130.6	132.7
4	.5	.95	234.2	286.0	.82	219.0	269.5	.81	184.0	180.6
2	.8	.95	38.2	64.5	.59	13.0	27.0	.48	62.1	74.9
3	.8	.95	98.2	158.7	.62	27.0	142.0	.19	122.8	133.7
4	.8	.95	178.3	278.4	.64	136.5	264.0	.52	181.7	183.4
2	.5	.90	52.3	66.7	.78	21.0	27.0	.78	70.8	76.8
3	.5	.90	132.9	161.2	.82	102.0	144.0	.71	131.6	134.1
4	.5	.90	247.3	286.9	.86	233.0	272.0	.86	183.1	180.0
2	.8	.90	43.0	64.5	.67	16.0	27.0	.59	64.9	74.4
3	.8	.90	110.7	159.8	.69	37.0	144.0	.26	127.8	133.1
4	.8	.90	204.4	280.8	.73	176.0	268.0	.66	189.1	181.2

Factors, parameters in simulation experiment; Length, number of trials until termination criterion of simulation is reached; *SE*, standard error; *M*, dimension; ρ, pairwise correlation of traits; sf, factor by which K_{other} is shrunk after each trial; META, multivariate Elo-type algorithm; ETA, separate (univariate) Elo-type algorithms derived from META by setting K_{other}, 0; *r*, ratio of means and medians, respectively

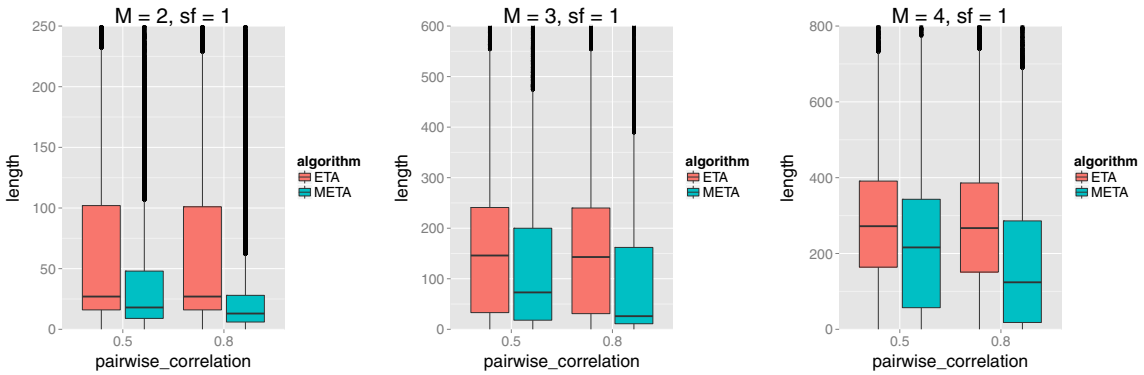

Fig. 4 Box-and-whiskers plots for the length of the experiment until convergence, for the comparison between META and the separate ETAs. The large number of outliers is due to the 10,000 replications of the simulation experiment

person parameter using one of the following scenarios after 40 trials (since $u = 1/40$, the uncertainty U has reached 0 at 40 trials; this can be seen as the worst case, since the META is least adaptive when $U = 0$):

1. (learn 1) Add 0.5 to the first component of the person parameter.
2. (learn all) Add 0.5 to all components of the person parameter.
3. (learn and fatigue) Add 0.5 to the first component of the person parameter, subtract 0.5 from the last.
4. (no change) Leave the person parameter unchanged.

Since we sampled the person parameters from a multivariate normal with standard deviations of 1 in all dimensions, the above values correspond to a change of half a standard deviation. In addition to the four scenarios, we varied the dimension ($M = 2, 3, 4$) and the pairwise correlation ($\rho = 0.5, 0.8$). After each of 10,000 replications, we recorded the total number of trials until termination. We calculated how long it took to correct for the change in the person parameter by comparing the resulting mean and median test lengths to those in the no-change conditions. Table 3 shows the resulting ratios.

Interestingly, the higher the dimension, the quicker (relative to the no-change condition) the changed parameters were reached, with the exception of the learn-and-fatigue scenario. In the latter scenario, the true covariance matrix is highly distorted by the changing person parameters, so the multivariate update cannot be expected to do well. Under none of the simulated conditions were either the median lengths or the mean lengths more than 23 % above those in the no-change condition, which we found surprising: For example, in the second simulation experiment, the mean length until convergence for $M = 2$ and $\rho = 0.5$ was 41.4, so that on average less than ten trials were needed to compensate for the learn-and-fatigue scenario. In sum, the META adapts well to changing person parameters. We mention in passing that many algorithms from psychophysics (especially the ML) can easily be modified to include only the last N trials for person parameter estimation. Such a modification also allows it to handle changing person parameters.

Robustness

The META is relatively robust to violations of the underlying model, in the following sense: If the META is to make sure a subject is exposed to items/stimuli of an appropriate difficulty, then the psychometric function can be misspecified, for which

Table 3 Adaptability of META to changing person parameter

Factors		Ratio of Lengths		
Scenario	ρ	M	Mean	Median
Learn 1	.5	2	1.14	1.18
		3	1.10	1.12
		4	1.08	1.09
	.8	2	1.19	1.23
		3	1.15	1.19
		4	1.11	1.14
Learn all	.5	2	1.19	1.22
		3	1.10	1.12
		4	1.05	1.05
	.8	2	1.17	1.24
		3	1.08	1.10
		4	1.04	1.03
Learn & fatigue	.5	2	1.18	1.23
		3	1.13	1.16
		4	1.10	1.11
	.8	2	1.16	1.22
		3	1.17	1.19
		4	1.15	1.18

All ratios of mean and median lengths have been computed relative to the *no-change* scenario; Factors, parameters in simulation experiment; Length, number of trials until termination criterion of simulation is reached; M, dimension; ρ, pairwise correlation of traits

the algorithm compensates. In this case, the META's measurements might be biased, but the observed performance will be close to the desired performance, nevertheless.

We treat this point more formally in a univariate case: Let $0 < p < 1$ denote the targeted performance, θ the true person parameter, and assume that g: $\mathbb{R}^2 \rightarrow [0, 1]$ is the true psychometric function, its first argument being the person parameter, and its second the item parameter. Assume that g is continuous in both arguments, and increases strictly monotonically in the person parameter and decreases strictly monotonically in the item parameter. Furthermore, assume that for any θ and any $x \in [0, 1]$, there is a β such that $g(\theta, \beta) = x$. Also assume that the psychometric function h used in the META also satisfies these conditions. Then, given a person estimate $\widehat{a}$, there is a difficulty β_h such that $h\left(\widehat{\theta}, \beta_h\right) = p$. In general $g(\theta, \beta_h) \neq p$. Let us discuss the two resulting cases: If $g(\theta, \beta_h) < p$, then the true chance to answer correctly is less than is assumed by the META. Hence, the person will be more likely to fail than is assumed by the META, and $\widehat{\theta}$ will decrease, on average—that is,

$$\mathrm{E}\left(\widehat{\theta}^{\mathrm{new}} \middle| \beta_h, \theta\right) < \widehat{\theta}.$$

Although this might result in estimates for $\widehat{\theta}$ that are far from θ, on average the next selected item/stimulus with parameter β_h^{next} will be easier than the previous one; that is,

$$\mathrm{E}\left(\beta_h^{\mathrm{next}}\right) < \beta_h,$$

and thus the true expected performance on this next item will increase–that is,

$$\mathrm{E}(g(\theta, \mathrm{E}\beta_h)) > g(\theta, \beta_h).$$

If $g(\theta, \beta_h) > p$, then the situation is reversed: Estimates for $\widehat{\theta}$ will increase, resulting in more difficult items/stimuli; hence, the true expected performance on the next item will now be lower. The actual performance will reach the targeted performance, but biased person parameters will result if g and h differ substantially. This observation is quite general, since almost all psychometric functions used in practice will meet the continuity and monotonicity restrictions that we demanded for g and h.

Concluding remarks

We introduced and discussed the META, a new approach for adapting experimental conditions, based on a multivariate extension of the Elo algorithm. The algorithm is relatively simple to implement and has a wide range of potential applications. If the underlying psychometric functions are well understood, our approach is also suited for measurement of (latent) traits, though traditional likelihood-based estimation techniques might be preferable in this context.

The META is superior with respect to the number of trials needed to reach the targeted performance, as compared to separate univariate Elo-type algorithms. In contrast to approaches from psychophysics or IRT, changes over time are an inherent feature of the algorithm. So the algorithm can handle within test learning and fatigue, as demonstrated by our simulations. This is in contrast to many likelihood-based procedures: ML or Bayesian estimates stabilize after a while and one has to resort to calculations of running estimates—that is, by using only the last N responses. In addition, pure likelihood-based procedures are prone to bias due to the misspecification of the likelihood. Hybrid procedures are viable alternatives here, and one such alternative, the META, also allows updates to correlated traits.

The way we presented the META does not include a parameter for the lower asymptote of the psychometric function. Such parameters have been studied in psychophysics (e.g., by Green, 1993) and in IRT, in which they are known as *pseudoguessing* parameters. Extending the META to include nontrivial lower asymptotes is possible if that asymptote is known (e.g., by the geometry of a four-alternative forced choice paradigm).

In practice, META has its potential usefulness in a wide range of experiments, in the following senses:

- META can be used for an adaptive threshold measurement of multivariate traits.
- META enables one to control the perceptual load or difficulty of experimental tasks.
- Due to its capability to estimate a subject's latent ability as reflected in the person parameter, the algorithm provides a convenient setting to make both within-subjects and between-subjects behavioral comparisons.
- Due to shorter test lengths the effects of possible confounding factors such as practice or fatigue are reduced.

As compared to staircase procedures, the META requires more preparation, since it has to be configured by the user by setting its tuning parameters to desirable values. It is advisable to use computer simulation to check the influence of the three main tuning parameters K, k, and K_{other}. Also, one has to conduct small-scale pilot experiments to determine the scale parameter s if no information from the literature is available. We remark that a range of values can result in usable algorithms, and that all of the tuning parameters have to be chosen in relation to s.

Nevertheless, the effort to set up the algorithm is low relative to computer adaptive testing, in which typically item banks (of stimuli) have to be calibrated by linear testing with

large samples. In situations with latent traits (or complex stimuli), the META handles this by pilot testing, as well; Klinkenberg et al. (2011) reported that univariate items find their positions on the latent scale quite quickly. If a physical trait such as a perception threshold is of interest, the META requires even less effort: Here, a pilot study is only needed to estimate the scaling factor s if the psychometric function can be assumed to be logistic.

In contrast to methods from psychophysics aiming to determine perception thresholds, the META can handle complex stimuli. As the large-scale implementation of a univariate Elo-type algorithm by Klinkenberg et al. (2011) has shown, these stimuli could, for example, be items measuring mathematics ability. In sum, the META, borrowing from the original ERS, CAT, and psychophysical methods, offers a powerful way to adapt experimental paradigms to subjects.

References

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Dekker.

Batchelder, W. H., & Bershad, N. J. (1979). The statistical analysis of a Thurstonian model for rating chess players. *Journal of Mathematical Psychology, 19*, 39–60.

Batchelder, W. H., Bershad, N. J., & Simpson, R. S. (1992). Dynamic paired-comparison scaling. *Journal of Mathematical Psychology, 36*, 185–212.

Birnbaum, A. (1974). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (2nd ed.). Reading, MA: Addison-Wesley.

Brinkhuis, M. J. S., & Maris, G. (2009). *Dynamic parameter estimation in student monitoring systems* (Measurement and Research Department Report No. 2009-1). Arnhem, The Netherlands: Cito.

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

Chica, A. B., Paz-Alonso, P. M., Valero-Cabré, A., & Bartolomeo, P. (2013). Neural bases of the interactions between spatial attention and conscious perception. *Cerebral Cortex, 23*, 1269–1279. doi:10.1093/cercor/bhs087

Eggen, T. J. H. M., & Verschoor, A. J. (2006). Optimal testing with easy or difficult items in computerized adaptive testing. *Applied Psychological Measurement, 30*, 379–393. doi:10.1177/0146621606288890

Elo, A. E. (1978). *The rating of chessplayers, past and present*. New York, NY: Arco.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Glickman, M. E. (1995). Chess rating systems. *American Chess Journal, 3*, 59–102.

Glickman, M. E. (1999). Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society, Series C, 48*, 377–394.

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*, 2096–2105.

Hambleton, R. K., Rogers, H. J., & Swaminathan, H. (1995). *Fundamentals of item response theory*. New York, NY: Sage.

Klinkenberg, S., Straatemeier, M., & van der Maas, H. L. J. (2011). Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers and Education, 57*, 1813–1824.

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

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

Maris, G. K. J., & van der Maas, H. L. J. (2012). Speed–accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika, 77*, 615–633.

R Development Core Team. (2012). R: A language and environment for statistical computing (ISBN 3-900051-07-0). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from www.R-project.org

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

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

van der Linden, W. J., & Glas, C. A. W. (2000). *Computerized adaptive testing: Theory and practice*. New York, NY: Kluwer.

van der Linden, W. J., & Glas, C. A. W. (2010). *Elements of adaptive testing*. New York, NY: Springer.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. Berlin, Germany: Springer.

Van der Linden, W. J., & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. In *Elements of adaptive testing* (pp. 3–30). New York, NY: Springer.

Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., & Steinberg, L. (2000). *Computerized adaptive testing: A primer*. Mahwah, NJ: Erlbaum.