Journal of Mathematical Psychology 95 (2020) 102326

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Sequential hypothesis tests for multinomial processing tree models[☆]

Martin Schnuerch[a,*], Edgar Erdfelder[a,*], Daniel W. Heck[b]

[a] *Department of Psychology, School of Social Sciences, University of Mannheim, Germany*
[b] *Department of Psychology, Philipps-Universität Marburg, Germany*

ARTICLE INFO

Article history:
Received 1 July 2019
Received in revised form 21 December 2019
Available online xxxx

Keywords:
Multinomial processing tree models
Hypothesis tests
Efficiency
Sequential analysis
Sequential probability ratio test

ABSTRACT

Stimulated by William H. Batchelder's seminal contributions in the 1980s and 1990s, multinomial processing tree (MPT) modeling has become a powerful and frequently used method in various research fields, most prominently in cognitive psychology and social cognition research. MPT models allow for estimation of, and statistical tests on, parameters that represent psychological processes underlying responses to cognitive tasks. Therefore, their use has also been proposed repeatedly for purposes of psychological assessment, for example, in clinical settings to identify specific cognitive deficits in individuals. However, a considerable drawback of individual MPT analyses emerges from the limited number of data points per individual, resulting in estimation bias, large standard errors, and low power of statistical tests. Classical test procedures such as Neyman–Pearson tests often require very large sample sizes to ensure sufficiently low Type 1 and Type 2 error probabilities. Herein, we propose sequential probability ratio tests (SPRTs) as an efficient alternative. Unlike Neyman–Pearson tests, sequential tests continuously monitor the data and terminate when a predefined criterion is met. As a consequence, SPRTs typically require only about half of the Neyman–Pearson sample size without compromising error probability control. We illustrate the SPRT approach to statistical inference for simple hypotheses in single-parameter MPT models. Moreover, a large-sample approximation, based on ML theory, is presented for typical MPT models with more than one unknown parameter. We evaluate the properties of the proposed test procedures by means of simulations. Finally, we discuss benefits and limitations of sequential MPT analysis.

1. Multinomial processing tree models

Among a multitude of outstanding contributions to the field of psychology, one of the arguably most prominent instances of William H. Batchelder's (1940–2018) scientific impact is the development of a class of stochastic models for the measurement of cognitive processes, known as multinomial processing tree (MPT) models. In what is now considered a classical article, Riefer and Batchelder (1988) introduced and promoted the use of MPT models which, in contrast to other scientific areas, had received but little attention in psychology at the time (Erdfelder, Auer, Hilbig, Aßfalg, Moshagen, & Nadarevic, 2009). Stimulated by this pioneering work and Batchelder's ongoing effort in the following years (e.g., Batchelder & Riefer, 1999), MPT models have become a powerful instrument to measure and disentangle the contribution of latent processes underlying observed behavior.

MPT models are substantively motivated stochastic models for categorical data (but see Heck, Erdfelder, & Kieslich, 2018, for an extension to continuous data). They are based on the assumption that each observable response in a specific paradigm originates from a finite set of sequences of discrete processing states. These sequences are conceptualized as branches in a processing tree. Nodes along these branches denote latent cognitive states and the links between the nodes represent (conditional) probabilities of entering the respective states. The product of these link probabilities determines the branch probability. Each category probability, in turn, is defined as the sum of probabilities of all branches terminating in this category. Based on the assumption that observed category frequencies follow a multinomial distribution, the (conditional) link probabilities can be estimated and, thus, the contribution of each latent processing state can be measured and tested statistically (Erdfelder et al., 2009; Hu & Batchelder, 1994).

Nowadays, MPT models are widely used in various branches of psychology, particularly in (social-)cognitive research. Even though the primary context of MPT applications is experimental psychology, Batchelder himself repeatedly promoted the use of

[☆] This research was supported by a grant from the Deutsche Forschungsgemeinschaft, Germany (DFG, GRK 2277) to the Research Training Group "Statistical Modeling in Psychology" (SMiP).
Reproducible scripts for simulations and analysis as well as all simulated raw data can be downloaded from the Open Science Framework (https://osf.io/98erb/).

[*] Correspondence to: Cognition and Individual Differences Lab, Department of Psychology, University of Mannheim, 68131 Mannheim, Germany.
E-mail addresses: martin.schnuerch@psychologie.uni-mannheim.de (M. Schnuerch), erdfelder@psychologie.uni-mannheim.de (E. Erdfelder).

https://doi.org/10.1016/j.jmp.2020.102326
0022-2496/© 2020 Elsevier Inc. All rights reserved.

MPT models for psychometric purposes (e.g., Batchelder, 1998; Batchelder & Riefer, 1999). Unlike item response models, for example, MPT models are based on explicit assumptions about the latent cognitive processes underlying observed responses and aim at measuring and disentangling these processes. Thus, Batchelder (1998) identified an "untapped potential" (p. 331) of what he referred to as "cognitive psychometrics" for individual assessments of specific cognitive processes, for example in clinical settings.

Despite the apparent appeal of MPT models for individual assessment, there is a notable limitation of this type of cognitive psychometrics. In experimental settings, MPT analyses typically make use of group data, either in a pooled or a hierarchical fashion (Chechile, 2009; Heck, Arnold, & Arnold, 2018; Klauer, 2006, 2010; Smith & Batchelder, 2010). As a consequence, parameter estimates and statistical tests are based on many data points, often resulting in high precision of estimates and high statistical power of tests. Individual analyses, in contrast, are typically based on far fewer observations. Thus, parameter estimates may be biased and will necessarily be less precise, resulting in large standard errors and low statistical power (Batchelder, 1998).

To remedy the problem of few observations in individual parameter estimation, Batchelder suggested to make use of Bayesian methods. Drawing on data from other individuals that are matched to the testee based on theoretical considerations (e.g., a reference group similar in age and educational background), one can construct a prior distribution for the parameters of interest. Using Bayes' theorem, this prior is then combined with the testee's data to obtain the posterior distribution. The mean (or mode) of this distribution serves as a point estimate while its variance or other measures of dispersion denote estimation uncertainty. When there is little variance in the prior, this empirical Bayes estimation procedure will result in a more precise estimate than classical maximum likelihood estimation without prior information (Batchelder, 1998). Of course, there is also a danger of considerable bias if the testee deviates systematically from other individuals (i.e., if the prior is misspecified).

A frequent goal in individual clinical assessment, however, goes beyond mere estimation of model parameters: To identify a specific cognitive deficit or to decide on a particular intervention for the individual testee, statistical tests on model parameters are required. For example, to assess whether or not an individual is able to utilize a certain cognitive process, one might want to test whether the corresponding parameter is substantially different from zero. In MPT modeling, tests of parameter constraints typically rely on null-hypothesis significance testing (NHST) based on the asymptotic distribution of some fit statistic under the null hypothesis (Batchelder & Riefer, 1999). In MPT models, these fit statistics denote the distance between model-implied and observed category frequencies. They can be characterized as a power divergence family (Read & Cressie, 1988), the most well-known special cases of which are Pearson's χ^2 or the log likelihood ratio G^2 (Hu, 1999; Hu & Phillips, 1999).

Standard applications of NHST to MPT models typically ignore statistical power, that is, the probability of rejecting a set of parameter constraints if the constraints do indeed not hold in the population. However, both in basic research and in clinical settings, sufficient statistical power is necessary for unbiased inference (Batchelder & Riefer, 1990, 1999). To this end, classical methods to control statistical error probabilities based on the seminal theory by Neyman and Pearson (1933) require an a priori power analysis. Given a certain expected effect size and a predefined Type 1 error probability α, the Type 2 error probability β (and the power of the test, $1 - \beta$) is a function of the sample size. Although power analyses are easily carried out with MPT software (e.g., multiTree; Moshagen, 2010) or general-purpose software for power analysis (e.g., G*Power; Faul, Erdfelder, Buchner, & Lang, 2009), there are two major drawbacks in the context of MPT analyses: First, a power analysis not only requires assumptions concerning the test-relevant parameters as specified by the null and the alternative hypothesis but depends on all other model parameters as well. This poses a problem whenever the model contains parameters for which the population values are unknown, so-called *nuisance* parameters. The second major limitation of classical power analyses in the Neyman–Pearson framework concerns scenarios in which the expected effect size is small. In this case, classical Neyman–Pearson tests require extremely large numbers of observations, often much larger than realistically feasible.

The problem of achieving a sufficiently powered hypothesis test is particularly pressing when data collection is costly: either when assessing a single participant with as few trials as possible or when each participant provides only a single data point (e.g., Batchelder, 1998; Heck, Thielmann, Moshagen, & Hilbig, 2018; Klauer, Stahl, & Erdfelder, 2007; Moshagen, Hilbig, Erdfelder, & Moritz, 2014; Moshagen, Musch, & Erdfelder, 2012; Schild, Heck, Ścigała, & Zettler, 2019). However, it potentially applies to any MPT model analysis (Batchelder & Riefer, 1990, 1999; Riefer & Batchelder, 1988). Therefore, in this article we introduce a sequential statistical method for hypothesis testing in MPT models that (1) allows to control both α and β error probabilities (unlike NHST), (2) requires on average much less observations than classical power analyses, and (3) does not rest on explicit assumptions about the population values of nuisance parameters of the model.

The approach we promote herein is based on Abraham Wald's sequential probability ratio test (Wald, 1947). In the following, we introduce the basic idea as well as an extension of Wald's method by Cox (1963). We then show how sequential tests can be used for efficient hypothesis tests in MPT models and how this may improve the applicability of MPT models for purposes of individual assessment. Overall, with the present article we hope to increase efficiency not only of typical experimental applications of MPT models, but also for applications to individuals in the context of cognitive psychometrics.

2. Sequential analysis

2.1. Sequential probability ratio tests

Classical statistical methods rely on fixed samples of an a priori defined size. Sequential statistics, in contrast, are based on the continuous monitoring of the data throughout the sampling process. This process continues until some predefined criterion is met, at which point sampling is terminated (*optional stopping*) and a statistical decision is made. Crucially, unlike the recursive application strategy of classical methods known as *p*-hacking (Simmons, Nelson, & Simonsohn, 2011), sequential methods do not compromise control of long-term error rates (Wetherill, 1975).

Due to their characteristic to terminate early whenever the data strongly support a hypothesis, statistical analysis may substantially reduce the required sample size. For a decision between two simple hypotheses, Wald's (1947) sequential probability ratio test (SPRT) has been proven to be the most efficient test (Matthes, 1963; Wald & Wolfowitz, 1948). That is, for given long-term error rates α and β, there is no test procedure that requires less observations than the SPRT on average.

To illustrate the SPRT, consider a random variable $\mathbf{X}$, $X \sim f(x|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes the true parameter vector of the underlying population. The random variable may be discrete or continuous, in which case the function $f(.)$ refers to the probability mass or the probability density, respectively. Assume a test of the two simple hypotheses $\mathcal{H}_0$: $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $\mathcal{H}_1$: $\boldsymbol{\theta} = \boldsymbol{\theta}_1$. A hypothesis is *simple* when all parameters of the statistical model that define the probability distribution of the data are either known or specified by the hypothesis. If at least one parameter is not known or restricted to a specific value, the hypothesis is *composite*. For the given example, the parameter vector $\boldsymbol{\theta}$ is completely specified under each hypothesis. Thus, the hypotheses are simple and the SPRT is the optimal test to decide between them with a given strength (α, β).

In the SPRT, the ratio of the probabilities of the observed data after any nth observation, $x^n = (x_1, \ldots, x_n)$, under each hypothesis i is computed. As the probability density is proportional to the likelihood, that is, $f(x^n|\boldsymbol{\theta}_i) \propto \mathcal{L}(\boldsymbol{\theta}_i;\ x^n)$, this ratio is typically referred to as a likelihood ratio (LR):

$$LR_n = \frac{f(x^n|\boldsymbol{\theta}_1)}{f(x^n|\boldsymbol{\theta}_0)} = \frac{\mathcal{L}(\boldsymbol{\theta}_1;\ x^n)}{\mathcal{L}(\boldsymbol{\theta}_0;\ x^n)}. \tag{1}$$

Sampling continues by adding independent observations x_{n+1} as long as

$$B < LR_n < A. \tag{2}$$

If $LR_n \geq A$, sampling is terminated and $\mathcal{H}_1$ is accepted. By definition, any sample x^n which leads to the acceptance of $\mathcal{H}_1$ is thus at least A times as likely under $\mathcal{H}_1$ as under $\mathcal{H}_0$. This implies that the probability to accept $\mathcal{H}_1$ is at least A times larger under $\mathcal{H}_1$ than under $\mathcal{H}_0$. In the usual notation based on the Neyman–Pearson theory, the former probability is defined as $1-\beta$, whereas the latter is denoted by α. Hence, $1 - \beta \geq A\alpha$, which can be written as

$$\frac{1-\beta}{\alpha} \geq A. \tag{3}$$

In contrast, if $LR_n \leq B$, sampling is terminated and $\mathcal{H}_0$ is accepted. Following the same logic as for A, we see that

$$\frac{\beta}{1-\alpha} \leq B, \tag{4}$$

which implies that upper/lower limits for A and B are given by $(1 - \beta)/\alpha$ and $\beta/(1 - \alpha)$, respectively. In practical applications, however, the inequalities in (3) and (4) can be treated as equalities defining threshold values A and B for the LR that satisfy pre-specified statistical error probabilities α and β. More precisely, the resulting sequential test procedure provides an approximate control of error probabilities with α and β serving as upper bounds to the actual error rates (Wald, 1947; Wetherill, 1975).

Functions describing the test procedure's properties (i.e., power and expected sample size at termination) can be approximated analytically by formulae derived by Wald (1947). Moreover, as mentioned before, the SPRT has been proven to be the most efficient test for given error rates. As soon as the statistical model defining the probability distribution of the data contains nuisance parameters, however, the general theory of the SPRT no longer applies.

This constitutes a practically relevant limitation since composite hypotheses due to nuisance parameters occur frequently in many common MPT models (e.g., in models for memory paradigms, which typically comprise guessing parameters). Different methods have been proposed to remedy this problem: For example, Wald (1947) suggested to integrate out nuisance parameters by means of weight functions (that resemble prior distributions in Bayesian inference). In a different approach, the likelihood ratio is constructed based on simple sufficient statistics (Barnard, 1952; Cox, 1952; Rushton, 1950). Although this approach provides an adequate solution for certain problems such as the classical t test (Schnuerch & Erdfelder, 2019), its applicability is restricted to specific situations. In the following, we consider a more general method introduced by Cox (1963), building on Bartlett's (1946) idea to construct a sequential test based on asymptotic maximum likelihood (ML) theory.

2.2. Sequential maximum likelihood ratio tests

Let $\mathbf{X}$ be a random variable denoting the observed data, with $X \sim f(x|\boldsymbol{\theta}, \boldsymbol{\phi})$. Similar as in the SPRT above, we consider a test of the hypotheses $\mathcal{H}_i$: $\boldsymbol{\theta} = \boldsymbol{\theta}_i$ $(i = 0, 1)$, $\boldsymbol{\phi}$ denoting nuisance parameters of the statistical model. The method developed by Cox (1963) and outlined in this section applies to any $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ regardless of their dimensionalities. Therefore, without loss of generality, we will assume that both parameters are single-valued in what follows. A detailed mathematical justification of Cox's method can be found in Breslow (1969).

In the SPRT, it is straightforward to consider the log likelihood ratio rather than the likelihood ratio. Assume the true value of ϕ was known, then the SPRT as defined in the previous section would require to continue sampling as long as

$$\log\left(\frac{\beta}{1-\alpha}\right) < \ell(\theta_1, \phi;\ x^n) - \ell(\theta_0, \phi;\ x^n) < \log\left(\frac{1-\beta}{\alpha}\right), \tag{5}$$

where $\ell(\theta_i, \phi;\ x^n)$ denotes the log likelihood for hypothesis i after n observations. Calculations involving exact log-likelihood functions are often difficult or even infeasible. As a remedy, based on large-sample theory, the exact log likelihood can be replaced by a second-order Taylor series expansion about the true parameter value θ, treating the difference $\theta_i - \theta$ $(i = 0, 1)$ as of order $1/\sqrt{n}$ (cf. Joanes, 1972):

$$\ell(\theta_i, \phi;\ x^n) = \ell(\theta, \phi;\ x^n) + (\theta_i - \theta)\frac{\partial \ell(\theta, \phi;\ x^n)}{\partial \theta} + \tfrac{1}{2}(\theta_i - \theta)^2 \frac{\partial^2 \ell(\theta, \phi;\ x^n)}{\partial \theta^2}\ , \tag{6}$$

such that the log likelihood ratio in (5) becomes

$$(\theta_1 - \theta_0)\frac{\partial \ell(\theta, \phi;\ x^n)}{\partial \theta} + \tfrac{1}{2}(\theta_1 - \theta_0)(\theta_1 + \theta_0 - 2\theta)\frac{\partial^2 \ell(\theta, \phi;\ x^n)}{\partial \theta^2}. \tag{7}$$

If, in contrast, ϕ is not known, the log likelihood ratio can be constructed using the ML estimate $\hat{\phi}$ based on x^n, that is,

$$\ell(\theta_1, \hat{\phi};\ x^n) - \ell(\theta_0, \hat{\phi};\ x^n)\ . \tag{8}$$

Note that Bartlett (1946) suggested separate ML estimates for the nuisance parameter ϕ conditional on $\mathcal{H}_1$ and $\mathcal{H}_0$ (i.e., the estimates $\hat{\phi}_1$ and $\hat{\phi}_0$ assuming $\theta = \theta_1$ or $\theta = \theta_0$, respectively). In contrast, Cox's (1963) method involves the use of a single estimate $\hat{\phi}$ for both terms in (8), conditional on a model without restrictions on θ or ϕ. Expanding about the true parameter values (θ, ϕ) analogously to (6), (8) becomes

$$(\theta_1 - \theta_0)\frac{\partial \ell(\theta, \phi;\ x^n)}{\partial \theta} + \tfrac{1}{2}(\theta_1 - \theta_0)(\theta_1 + \theta_0 - 2\theta)\frac{\partial^2 \ell(\theta, \phi;\ x^n)}{\partial \theta^2} + (\theta_1 - \theta_0)(\hat{\phi} - \phi)\frac{\partial^2 \ell(\theta, \phi;\ x^n)}{\partial \theta \partial \phi}. \tag{9}$$

It is easy to see that (9) is equivalent to (7) if the last term becomes 0, that is, if θ and ϕ are independent and, thus,

$$E\left[\frac{1}{n}\frac{\partial^2 \ell(\theta, \phi;\ x^n)}{\partial \theta \partial \phi}\right] \xrightarrow[n\to\infty]{} 0\ . \tag{10}$$

In this case, the ML estimates $\hat{\theta}$ and $\hat{\phi}$ are asymptotically independent as well. A simple SPRT as defined in (5) where ϕ is replaced by $\hat{\phi}$ is then asymptotically equivalent to that when ϕ is known. If $\hat{\theta}$ and $\hat{\phi}$ are not asymptotically independent, however, the test procedure will not satisfy the long-run error rates implied by α and β. As a remedy, the sampling error of $\hat{\phi}$ must be taken into account.

Eq. (9) can be further simplified based on large-sample ML theory, showing that it is asymptotically equivalent to the following expression (see Appendix A and Cox, 1963, for details):

$$n\mathcal{I}_{\theta\theta}(\theta_1 - \theta_0)\left[\hat{\theta} - \tfrac{1}{2}(\theta_0 + \theta_1)\right], \tag{11}$$

where $\mathcal{I}_{\theta\theta}$ denotes the (θ, θ) element or submatrix of the expected Fisher information matrix $\mathcal{I}(\theta, \phi)$ for sample size $n = 1$, assuming observations to be independent and identically distributed.

For simplification, Cox (1963) suggested to base the sequential test procedure on a monotonic transformation of (11) obtained by dropping the multiplicative constant $\mathcal{I}_{\theta\theta}(\theta_1 - \theta_0)$ (see also Wetherill, 1975, p. 60),

$$T_n = n\left[\hat{\theta} - \tfrac{1}{2}(\theta_0 + \theta_1)\right], \tag{12}$$

where $\hat{\theta}$ is the ML estimate of θ based on x^n. This test statistic has to be computed after any nth observation, and stopping boundaries corresponding to the constant likelihood-ratio boundaries of the SPRT (Eq. (2)) are given by

$$\frac{\mathcal{V}_{\theta\theta}}{\theta_1 - \theta_0}\log\left(\frac{\beta}{1-\alpha}\right) < T_n < \frac{\mathcal{V}_{\theta\theta}}{\theta_1 - \theta_0}\log\left(\frac{1-\beta}{\alpha}\right). \tag{13}$$

In (13), $\mathcal{V}_{\theta\theta}$ denotes the (θ, θ) element of the inverse of the expected unit Fisher information, that is, $\mathcal{V} = \mathcal{I}(\theta, \phi)^{-1}$. In many cases, the analytical derivation of the expected Fisher information is infeasible. Thus, for practical purposes, it can be replaced by the observed Fisher information $\mathbf{I}(\hat{\theta}, \hat{\phi})$, that is,

$$\mathbf{I}(\hat{\theta}, \hat{\phi}) = -\frac{1}{n}\mathbf{H}(\hat{\theta},\ \hat{\phi}), \tag{14}$$

where $\mathbf{H}(\hat{\theta},\ \hat{\phi})$ is the Hessian matrix of second-order partial derivatives of the log likelihood function, evaluated at the ML estimates.

As an element of the inverse of the unit Fisher information, $\mathcal{V}_{\theta\theta}$ (or, when using the observed information matrix, $\mathbf{V}_{\theta\theta}$) denotes the variance of the ML estimate $\hat{\theta}$ based on a single observation (cf. Ly, Marsman, Verhagen, Grasman, & Wagenmakers, 2017). Thus, the threshold values in (13) are adjusted based on the precision with which the test-relevant parameter is estimated, thereby correcting for the uncertainty that results from the necessity to estimate the unknown nuisance parameter ϕ.

Cox's test, henceforth referred to as sequential maximum likelihood ratio test (SMLRT), satisfies the long-run error rates α and β for testing hypotheses about θ without making explicit assumptions about the nuisance parameters ϕ (Cox, 1963; Wetherill, 1975). As it is based on asymptotic ML theory, however, the approximations involved in the derivation of the formulae cannot be expected to work sufficiently well for small samples. Hence, the proposed sequential test requires a sufficiently large initial sample (Cox & Roseberry, 1966). Otherwise, it is not ensured that the Taylor series expansion in (9) is valid or that the observed Fisher information $\mathbf{I}(\hat{\theta}, \hat{\phi})$ provides a good approximation of the expected Fisher information $\mathcal{I}(\theta, \phi)$ (Hu & Phillips, 1999). Nevertheless, even though the initial sample size needs to be sufficiently large, we show below that the SMLRT still requires on average much smaller sample sizes than Neyman–Pearson tests without compromising error probability control.

The practical implementation of the SMLRT for MPT models is straightforward with MPT software such as, for example, multiTree (Moshagen, 2010) or MPTinR (Singmann & Kellen, 2013). After any nth observation, the ML estimate $\hat{\theta}$ can easily be computed with these software packages. Additionally, $\mathbf{V}_{\theta\theta}$ can be computed from software output based on the estimated standard error of $\hat{\theta}$, $SE_{\hat{\theta}}$. Since $SE_{\hat{\theta}}$ is the (θ, θ) element of $\left[n\mathbf{I}(\hat{\theta}, \hat{\phi})\right]^{-1/2}$, it follows that $\mathbf{V}_{\theta\theta} = n(SE_{\hat{\theta}})^2$.

3. Sequential MPT analysis

3.1. Case 1: Simple hypothesis

As a running example, consider a psychometric experiment administered to an individual participant in a clinical assessment situation. Assume we are interested in the individual's perceptual abilities. Specifically, we want to assess whether or not the participant is able to detect a visual stimulus of a given intensity.

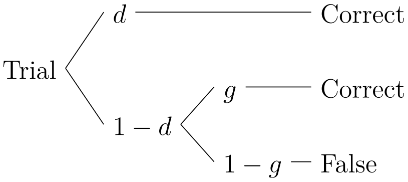

Fig. 1. A simple multinomial processing tree model for a perception experiment with a two-alternative forced-choice test. d = probability to detect the stimulus; g = probability to guess correctly.

The experiment is carried out as follows: In the style of classical experiments on visual thresholds (Blackwell, Pritchard, & Ohmart, 1954) and decision processes underlying visual perception (Swets, Tanner, & Birdsall, 1961), the participant is presented with a visual stimulus in one of two defined temporal intervals in each trial. A stimulus typically used in such experiments is a flash of light displayed on a screen (for 100 ms, say) with a certain diameter and magnitude (that is, luminous intensity). Following each trial, the participant is prompted to decide in which of the two intervals the stimulus was presented. Thus, the perceptual performance is measured in a two-alternative forced-choice test (2AFC).

If the participant detects the stimulus, they will answer correctly. If they do not detect the stimulus, however, they might still give a correct answer by guessing the interval in which the stimulus was presented. Thus, the performance in the 2AFC is diluted by guessing processes which do not represent actual perceptual abilities (Swets et al., 1961). In order to assess these directly, the processes underlying response behavior in the 2AFC can be disentangled by means of an MPT model.

Fig. 1 displays the simplest instance of an MPT model for the paradigm under consideration. In each trial, participants either enter a state of detection (with probability d) and choose the correct answer, or they do not detect the stimulus $(1-d)$. In this state of uncertainty, they have to guess which of the intervals contained the stimulus. Thus, they can either guess correctly (with conditional probability g) or incorrectly $(1-g)$.

Formally, the probability of each branch j $(j = 1, \ldots, J)$ leading to response category k $(k = 1, \ldots, K)$ in a binary MPT model is defined as

$$p_{jk}(\boldsymbol{\Theta}) = c_{jk}\prod_{s=1}^{S}\theta_s^{a_{jks}}(1-\theta_s)^{b_{jks}}, \tag{15}$$

where $\boldsymbol{\Theta} = (\theta_1, \ldots, \theta_S)$ represents the vector of parameters in the model denoting the (conditional) link probabilities along the branches, with $\boldsymbol{\Theta} \in \boldsymbol{\Omega} = [0,\ 1]^S$. The count variables a_{jks} and b_{jks} indicate how often a parameter θ_s (or its complement $1-\theta_s$, respectively) occurs in a branch, while c_{jk} denotes the product of fixed parameter values along each branch (Hu & Batchelder, 1994).

The probability of each category k as a function of the model parameters is the sum of all branch probabilities ending in category k,

$$p_k(\boldsymbol{\Theta}) = \sum_{j=1}^{J}p_{jk}(\boldsymbol{\Theta}). \tag{16}$$

For the observed category frequencies $n^K = (n_1, \ldots, n_K)$, $\sum_{k=1}^{K}n_k = N$, the resulting likelihood function is then given by

$$\mathcal{L}(\boldsymbol{\Theta};\ n^K) = N!\prod_{k=1}^{K}\frac{[p_k(\boldsymbol{\Theta})]^{n_k}}{n_k!}. \tag{17}$$

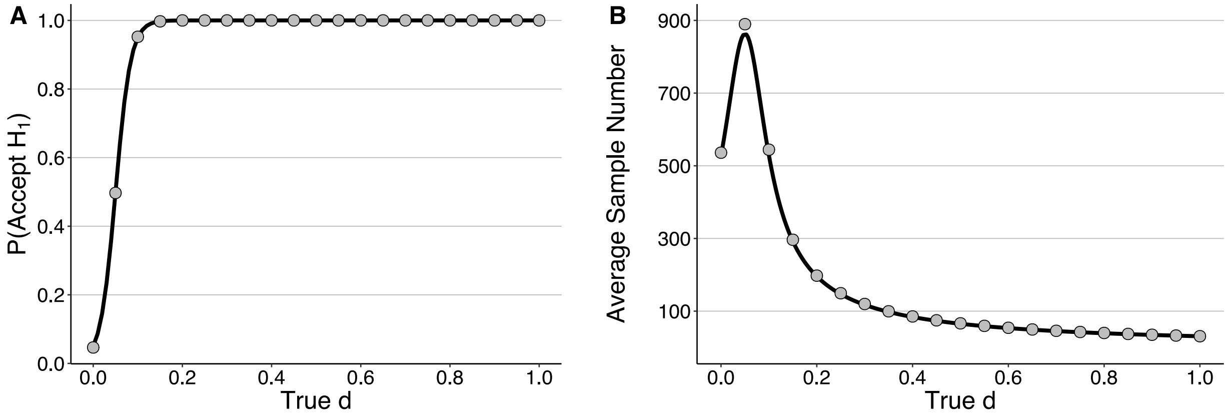

Fig. 2. A: The black line denotes the Operating Characteristic (OC) function for a sequential probability ratio test (SPRT) on $\mathcal{H}_0$: $d = 0$ versus $\mathcal{H}_1$: $d = .10$ with $\alpha = \beta = .05$. **B**: The black line denotes the Average Sample Number (ASN) function of the respective SPRT. Grey dots denote simulated estimates of OC and ASN for the given test, based on 10,000 replications per estimate.

For parameter interpretation, any statistical modeling requires the model fitted to the data to be identifiable. In case of an MPT model, this means that $\boldsymbol{\Theta} \neq \boldsymbol{\Theta}'$ implies that $\mathbf{p}(\boldsymbol{\Theta}) \neq \mathbf{p}(\boldsymbol{\Theta}')$, for all $\boldsymbol{\Theta}, \boldsymbol{\Theta}' \in \boldsymbol{\Omega}$. In other words, a model is globally identifiable if any specific set of model-consistent category probabilities corresponds to a unique set of parameter values (Bamber & van Santen, 2000).

In our example, the MPT model contains two parameters: $\boldsymbol{\Theta} = (d,\ g)$. In a balanced and completely randomized design, it is reasonable to assume that guessing in a 2AFC cannot be systematically "biased" towards a correct or incorrect response. Therefore, we fix the guessing parameter a priori, $g = .50$. Thus, according to (16) the probability of a correct response is given by

$$p_c(d) = d + (1 - d) \cdot .50\ , \tag{18}$$

while the probability of an incorrect response is given by $1 - p_c(d)$, since there are only two observed response categories. The restricted model is identifiable, but since $K' = S'$, with K' denoting the number of independent categories and S' the number of free parameters, it is saturated and does not allow for tests of goodness of fit. It is still possible, however, to test hypotheses about free parameters in a saturated model.

To assess the participant's ability to detect the visual stimulus, we want to test the following hypotheses on the detection parameter d in our MPT model: $\mathcal{H}_0$: $d = 0$ versus $\mathcal{H}_1$: $d > 0$. In other words, is the response behavior based entirely on guessing or can the participant detect the stimulus at least sometimes? To control the probabilities of decision errors, we request that the test accepts $\mathcal{H}_1$ with probability $\alpha = .05$ if $d = 0$, and $\mathcal{H}_0$ with probability $\beta \leq .05$ if $d \geq .10$. To this end, we test the simple hypothesis that $d = d_0 := 0$ versus the simple alternative that $d = d_1 := .10$.

In a classical analysis, we would sample N observations from the participant and test whether our MPT model with two restricted parameters $\boldsymbol{\Theta}_{R_2} = (d = 0,\ g = .50)$ fits the data worse than a model where d remains unrestricted, that is, $\boldsymbol{\Theta}_{R_1} = (d,\ g = .50)$. As the two models are nested, the test is based on the difference of the respective fit statistics, $\Delta\mathrm{PD}^{\lambda}$, where PD^{λ} denotes any power divergence statistic defined by λ, for example, the log-likelihood ratio statistic G^2 if $\lambda = 0$ or Pearson's χ^2 statistic if $\lambda = 1$. Under the null hypothesis defined above, $\Delta\mathrm{PD}^{\lambda} \sim \chi^2(1)$ holds asymptotically, irrespective of the PD^{λ} statistic chosen (Read & Cressie, 1988). Thus, if $P(\chi^2(1) \geq \Delta\mathrm{PD}^{\lambda}) < \alpha$, we decide in favor of the hypothesis $d \geq .10$.

In this example, a power analysis is straightforward. The models under $\mathcal{H}_0$ and $\mathcal{H}_1$ imply certain category probabilities. A common standardized effect size measure for the discrepancy between expected proportions under two hypotheses is Cohen's w (Cohen, 1992). In a single-tree MPT model, w is given by

$$w = \sqrt{\sum_{k=1}^{K} \frac{(p_{1k} - p_{0k})^2}{p_{0k}}}\ , \tag{19}$$

where p_{ik} denotes the probability of category k under hypothesis $i = 0, 1$. Based on (18) and (19), the expected effect size in our example with $d_0 = 0$ and $d_1 = .10$ is $w = 0.10$, denoting a small effect. Thus, a one-tailed asymptotic test of the hypothesis that $d = d_0$ versus $d = d_1$ with error probabilities $\alpha = \beta = .05$ requires approximately $N = 1083$ observations (Faul et al., 2009).

Since we are dealing with simple hypotheses, the SPRT provides a most efficient alternative. Let $p_i = p_c(d_i)$, then the likelihood given hypothesis i, according to (17), is

$$\mathcal{L}(d_i;\ n_c) = \binom{N}{n_c} p_i^{n_c} (1 - p_i)^{N - n_c}\ , \tag{20}$$

where n_c denotes the observed number of correct responses. Thus, our hypotheses can be tested by means of an SPRT by continuing to sample observations from the participant as long as

$$\frac{\beta}{1 - \alpha} < \frac{p_1^{n_c}(1 - p_1)^{N - n_c}}{p_0^{n_c}(1 - p_0)^{N - n_c}} < \frac{1 - \beta}{\alpha} \tag{21}$$

and terminating as soon as one of the inequalities is violated, thus accepting either $\mathcal{H}_0$ or $\mathcal{H}_1$.

Based on formulae derived by Wald (1947), it is straightforward to approximate functions describing the test procedure's properties (see Appendix B for details). Specifically, we can analytically determine the procedure's probability to accept the alternative hypothesis (the so-called Operating Characteristic, OC) as well as the expected sample size at termination (the so-called Average Sample Number, ASN) as a function of the true value of the parameter d. The respective functions of the SPRT in this example are depicted in Fig. 2. Additionally, we simulated the SPRT for the given hypotheses to demonstrate how well the procedure's properties are approximated in practice.[1] The results are denoted by the grey dots in Fig. 2. Except for a slight underestimation of the ASN when the true value lies between d_0 and d_1, the analytical functions approximate the simulated estimates almost perfectly.

[1] R scripts for this and all following simulations as well as all simulated raw data are available from https://osf.io/98erb/.

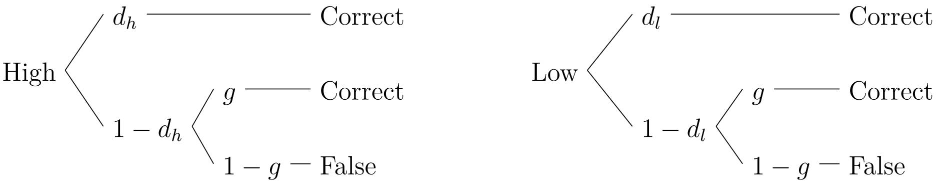

Fig. 3. A multinomial processing tree model for a perception experiment with two stimulus magnitudes (high versus low luminous intensity) and a two-alternative forced-choice test. d_h = probability to detect the stimulus with high magnitude; d_l = probability to detect the stimulus with low magnitude; g = probability to guess correctly.

As the results show, the SPRT not only controls error probabilities as accurately as Neyman–Pearson tests do, it does so notably more efficiently. For any true value d, the expected sample size at termination is substantially smaller than the sample size determined by an a priori power analysis for the given hypotheses ($N = 1083$). When d equals d_0 or d_1, the expected sample size of the SPRT is approximately $N = 545$, that is, almost 50% smaller. Moreover, if the true parameter value is notably larger than specified by the hypothesis, the test will require even lower sample sizes to make a decision. Classical analysis, in contrast, requires the a priori defined sample size irrespective of the true value. Thus, for the test of a simple hypothesis in a single-parameter MPT model, the SPRT is a highly efficient alternative to classical inference procedures.

3.2. Case 2: Composite hypothesis with a single nuisance parameter

In practical applications of cognitive psychometrics as well as in experimental settings, parameter tests in MPT models will rarely be on absolute parameter values as in Case 1. It is much more common to test equality or order constraints on model parameters to compare cognitive processes under different conditions or with different stimulus material. The challenge with this kind of parameter tests, however, is that they typically imply tests of composite hypotheses.

Consider the following extension of the simple psychometric experiment introduced in Case 1. Instead of the absolute perceptual ability, we now want to assess the testee's perceptual sensitivity. Specifically, we manipulate physical features of the visual stimulus presented and assess whether the participant's ability to detect the stimulus differs between conditions (see Blackwell et al., 1954, for a similar experimental procedure). To this end, the stimulus is now presented in two different magnitudes (low versus high luminous intensity). As in Case 1, we want to test the detection processes directly by means of an MPT analysis of the individual's performance in the 2AFC.

Fig. 3 depicts the extended MPT model for Case 2. The model now comprises two processing trees, one for each stimulus magnitude. We still assume unbiased guessing of the correct response in the 2AFC, that is, $g = .50$ for each stimulus type. However, to test whether the manipulation of stimulus magnitude affects the detection probability, the model now contains two detection parameters, d_h (high magnitude) and d_l (low magnitude). We want to test the hypotheses $\mathcal{H}_0$: $d_h = d_l$ versus $\mathcal{H}_1$: $d_h > d_l$, as the probability to detect the stimulus should be higher for high stimulus magnitude than for low magnitude.

To incorporate parametric order constraints into a binary MPT model, it is straightforward to reparameterize the model such that the new model satisfies all assumptions of binary MPT models and is statistically equivalent to the original model (Knapp & Batchelder, 2004): By restructuring the processing tree for low-intensity stimuli and introducing a new parameter ξ (Fig. 4), we can express d_l in terms of d_h:

$$d_l = \xi d_h \ . \tag{22}$$

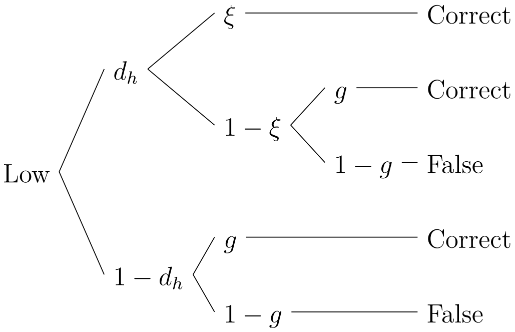

Fig. 4. Reparameterization of the second processing tree depicted in Fig. 3 for the order constraint $d_h > d_l$. d_h = probability to detect the stimulus with high magnitude; ξ = ratio of the probability to detect the stimulus with low magnitude to d_h; g = probability to guess correctly.

The reparameterized model, just as the original model, contains two unknown parameters, $\Theta = (d_h,\ \xi)$, both of which are free to vary in the entire parameter space $\Omega = [0,\ 1]$. Our hypotheses are then reformulated in terms of ξ, that is, $\mathcal{H}_0$: $\xi = \xi_0$ ($\xi_0 = 1$) and $\mathcal{H}_1$: $\xi = \xi_1$ ($\xi_1 < 1$). Thus, our hypotheses are about the ratio of detection probabilities for low and high stimulus magnitude.

It is easy to see that these hypotheses are composite as the probability distribution of our data depends both on ξ, which is specified by the hypotheses, and d_h, which is unknown. This is a particular problem for a Neyman–Pearson test of the hypotheses, as the effect size and, in turn, the power of the test also depend on both parameters.

If an MPT model includes more than one tree, the model becomes a joint MPT model. For $T > 1$ trees, Cohen's effect size measure w generalizes to

$$w = \sqrt{\sum_{t=1}^{T} \pi_t \cdot \sum_{k_t=1}^{K_t} \frac{(p_{1kt} - p_{0kt})^2}{p_{0kt}}}\ , \tag{23}$$

where K_t denotes the total number of categories in tree t ($t = 1, \ldots, T$) and π_t the proportion of the total sample size N assigned to tree t. Resembling Case 1, p_{1kt} denotes the predicted category probabilities for category k of tree t according to $\mathcal{H}_1$. However, since $\mathcal{H}_0$ is composite, the corresponding p_{0kt} category probabilities are now obtained by fitting the $\mathcal{H}_0$ model (with d_h free) to these $\mathcal{H}_1$ probabilities such that w becomes a minimum (Erdfelder, Faul, & Buchner, 2005). Note that (23) reduces to (19) iff $T = 1$.

Assume $\xi_0 = 1.00$ and $\xi_1 = .75$. Then the expected effect size for $d_h = .70$ in a balanced design with $\pi_{high} = \pi_{low} = .50$ is approximately $w = 0.11$ according to (23). An a priori power analysis for this effect size reveals a required sample size of $N = 892$ observations for a one-tailed asymptotic test with $\alpha = \beta = .05$. If, however, $d_h = .50$, the expected effect size is only $w = 0.07$ and the required sample size for the same test is more than twice as large, that is, $N = 2248$.

To ensure a sufficiently powered test in the context of a composite hypothesis, a rational strategy would be to assume a conservative value of d_h such that the resulting test has power $1 - \beta \geq .95$ for any d_h in a reasonable range. However, this can be inefficient and demand very large sample sizes.

Instead, we can analyze the data sequentially by means of the SMLRT. Let $p_h = d_h \cdot (1 - d_h) \cdot .50$ denote the probability of a correct response in a trial with high stimulus magnitude under both hypotheses, and $p_{li} = \xi_i d_h + (1 - \xi_i d_h) \cdot .50$ the corresponding probability for low stimulus magnitude under hypothesis i. The likelihood function is then given by

$$\mathcal{L}(\xi_i, d_h;\ n_1, n_2, n_3, n_4) = \frac{N!}{\prod_{k=1}^{4} n_k!} p_h^{n_1} (1 - p_h)^{n_2} p_{li}^{n_3} (1 - p_{li})^{n_4} \ , \tag{24}$$

where the n_k ($k = 1, 2, 3, 4$; $\sum_{k=1}^{4} n_k = N$) denote observed frequencies of correct versus false responses for high versus low stimulus magnitude, respectively.

If Eq. (10) is satisfied in our case (with $\theta = \xi$ and $\phi = d_h$), that is, if $\hat{d}_h$ and $\hat{\xi}$ are asymptotically independent, the SMLRT reduces to a simple SPRT where d_h is replaced by $\hat{d}_h$ at each step. However, since

$$\frac{\partial^2 \ell(\xi, d_h;\ n_1, n_2, n_3, n_4)}{\partial \xi \partial d_h} = \frac{n_3}{(\xi d_h + 1)^2} - \frac{n_4}{(\xi d_h - 1)^2} \tag{25}$$

and considering that in a balanced design, $E(n_3) = N/2 \cdot [\xi d_h + (1 - \xi d_h) \cdot .50]$ and $E(n_4) = N/2 - E(n_3)$, we see that

$$E\left[\frac{1}{N} \frac{\partial^2 \ell(\xi, d_h;\ n_1, n_2, n_3, n_4)}{\partial \xi \partial d_h}\right] = -\frac{\xi d_h}{2(1 - \xi^2 d_h^2)} \ . \tag{26}$$

The term in (26) no longer depends on N and, thus, (10) is not satisfied if neither ξ nor d_h are equal to 0. Hence, as suggested by Cox (1963), the test should be based on (12) with stooping boundaries (13), where θ is replaced by ξ. To calculate the expected Fisher information in order to obtain $\mathcal{V}_{\xi\xi}$ at each step, observed cell frequencies in the Hessian matrix $\mathbf{H}(\xi, d_h)$ are replaced by the expected cell frequencies, as was done in (26). Additionally, when $\xi_1 < \xi_0$, the inequalities in (13) must be inverted. This will be the case for order constraints in MPT models, where the null hypothesis typically denotes $\xi_0 = 1$, such as in our example.

Unlike in the SPRT for simple hypotheses, there are no analytical formulae for the SMLRT's properties for composite hypotheses. Therefore, we simulated the SMLRT for the perception experiment in Case 2 (1) to assess whether long-run error rate control works as expected and (2) to compare the expected sample size required by the SMLRT with that of the classical Neyman–Pearson test.

The simulations were carried out in the statistical computing environment R (R Core Team, 2019). We generated participants' responses according to the model depicted in Figs. 3 and 4 and analyzed them sequentially by means of the SMLRT defined by (12) and (13) with inverted boundaries. Estimates of ξ were computed with the R package MPTinR (Singmann & Kellen, 2013).

We simulated data for different true values of d_h ($d_h = .70, .50$) and ξ ($\xi = 1.00, .75, .50$). Under the null hypothesis, ξ_0 was always equal to 1, while under the alternative hypothesis, ξ_1 was equal to .75 or .50. Furthermore, we varied the initial sample size of the sequential procedure (N_{min}). As the SMLRT is based on large-sample approximations, a too small sample size might negatively affect the procedure and compromise its error rates (Cox & Roseberry, 1966). As a simple strategy to find a suitable number, the initial sample was therefore defined to be 25%, 40%, or 50% of the sample size required by a corresponding Neyman–Pearson test (N_{NP}).[2]

In each step, the sample size was increased by +2, one observation for each stimulus magnitude, until a threshold was reached. Threshold values were chosen such that $\alpha = \beta = .05$. For each parameter combination, we replicated the test procedure 1000 times.

The results are displayed in Fig. 5. It contains the empirical error rates (α' and β') and the required sample sizes as a function of d_h, the true value of ξ, that is, $\xi = \xi_0$ or $\xi = \xi_1$, and the initial sample size. Error bars for the error rates denote 95% exact confidence intervals (Clopper & Pearson, 1934). The sample size distributions are displayed as boxplots. Black dots denote outliers (data points further than 1.5 times the inter-quartile range below or above the first or the third quartile, respectively), grey dots represent the means of the distributions, that is, the ASN. Dashed lines denote the nominal error rates and the sample sizes required by a corresponding Neyman–Pearson test.

The left part of Fig. 5 shows the results for $\xi_1 = .50$, the right part displays results for $\xi_1 = .75$. For all parameter combinations, β' substantially undercuts the nominal level. At the same time, except for a slight upward deviation when $d_h = .50$ and the initial sample size is small, α' adheres to the nominal level. Moreover, the SMLRT controls error probabilities notably more efficiently than a corresponding Neyman–Pearson test: The ASN is on average 45% smaller. Across all parameter combinations, the test terminates with a sample size smaller than N_{NP} in 94% of the cases.

In almost all of the simulated scenarios, the SMLRT shows satisfying results for an initial sample size of $N_{min} = .25 \cdot N_{NP}$. With increasing N_{min}, the test procedure becomes more conservative and less efficient. However, the increase in ASN is only slight and still below the sample size required by the Neyman–Person test. Concluding from our results, an initial sample size of 25% of a corresponding Neyman–Pearson test is a reasonable starting point to efficiently control long-run rates of statistical decision errors for parameter tests in MPT models with a single unknown nuisance parameter.

3.3. Case 3: Composite hypothesis with several nuisance parameters

Commonly, MPT models contain several unknown parameters. Thus, hypotheses about single parameters typically involve more than one nuisance parameter. To illustrate that the SMLRT naturally extends to this case, consider the following variation of our psychometric experiment.

To assess potential biases involved in the decision process as well as perceptual processes, the experiment is now based on a Yes/No test. That is, in each trial either a stimulus (target) or no stimulus (lure) is presented. For each trial, the participant has to indicate whether they detected a stimulus ("Yes") or not ("No"). As in Case 2, stimuli are light flashes presented in two different magnitudes (high versus low luminous intensity).

Fig. 6 displays the MPT model for Case 3. It contains three detection parameters denoting the probability to detect a stimulus with high magnitude (d_h), a stimulus with low magnitude (d_l), or the absence of a stimulus (d_n). Additionally, it contains the parameter g, which represents the conditional probability to guess "Yes" in a state of uncertainty.

With $K' < S'$, the model is not identifiable. Thus, we need to restrict at least one of the parameters. As g no longer refers to guessing correctly but rather guessing that a stimulus was

[2] To increase computational efficiency, each simulated trajectory started with $N_{min} = .25 \cdot N_{NP}$ and was then reanalyzed with $N_{min} = .40 \cdot N_{NP}$ and $N_{min} = .50 \cdot N_{NP}$.

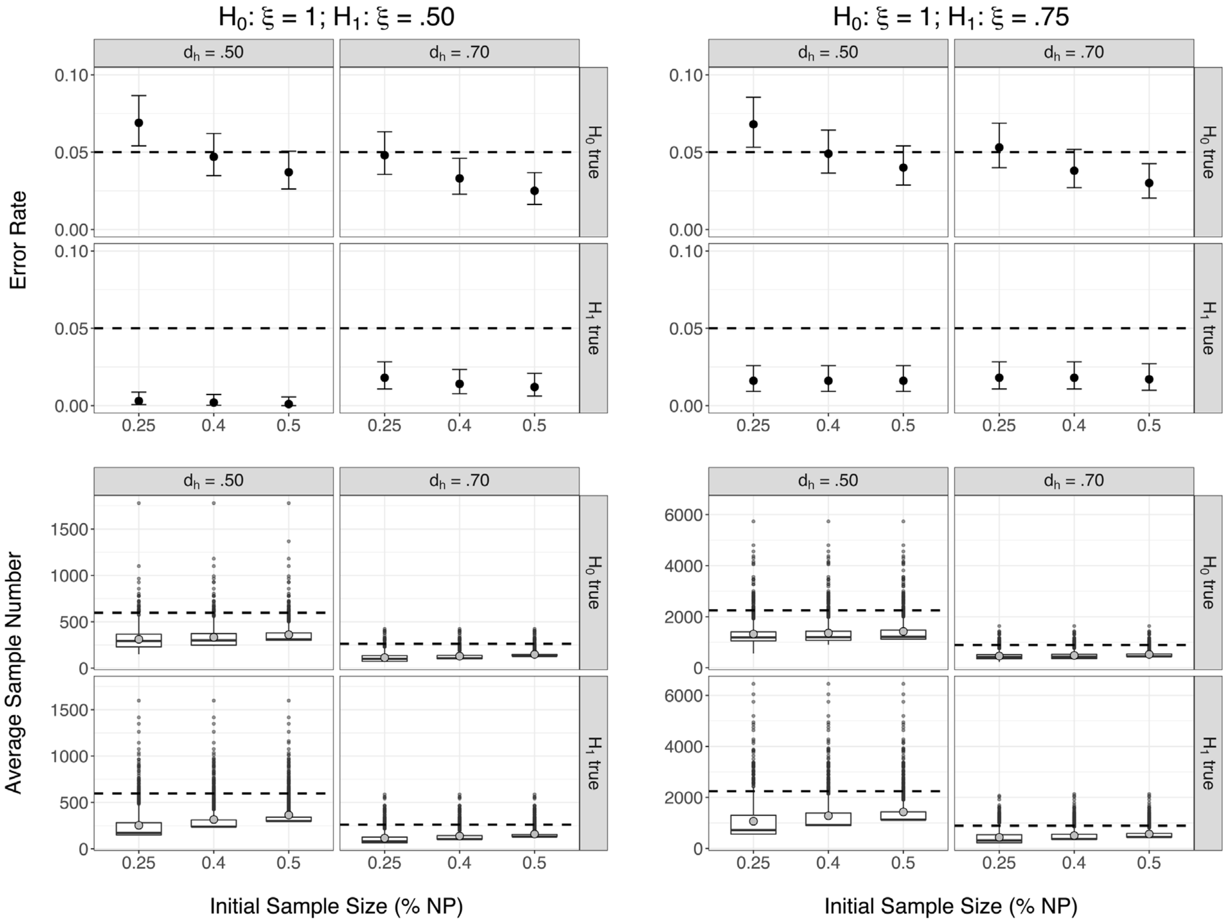

Fig. 5. Empirical error rates and sample size distributions of the sequential maximum likelihood ratio test (SMLRT) as a function of the hypothesis tested, the true detection parameter d_h, and the data-generating scenario. Error bars denote 95% Clopper–Pearson exact confidence intervals (Clopper & Pearson, 1934). Black dots in the boxplots denote outliers (data points more than 1.5 times the inter-quartile range below or above 1st or 3rd quantile). Grey dots denote mean sample sizes. Dashed lines represent nominal error rates and sample sizes required by a corresponding Neyman–Pearson (NP) test.

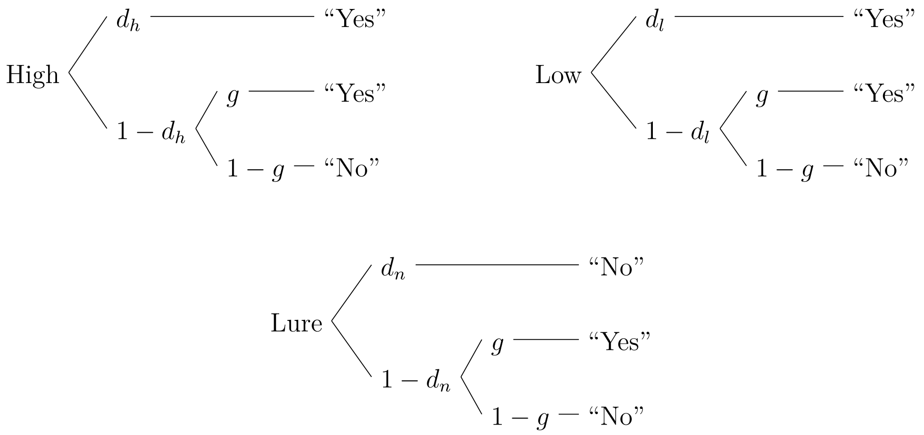

Fig. 6. A multinomial processing tree (MPT) model for a perception experiment with two stimulus magnitudes (high versus low) and a Yes/No test. d_h = probability to detect the stimulus with high magnitude; d_l = probability to detect the stimulus with low magnitude; d_n = probability to detect a lure trial in which no stimulus was presented; g = probability to guess "Yes".

presented, it seems reasonable not to restrict it a priori. For the given experiment, we rather assume that the absence of a stimulus should be equally salient and detectable as the presence of a high-magnitude stimulus. Thus, we will assume that $d_n = d_h$. The restricted model is identifiable and saturated.

To test whether the participant is sensitive to the manipulation of stimulus magnitude in the new paradigm, we will reparameterize the model as we did in Case 2 (see Fig. 4), such that $d_l = \xi d_h$. Again, we test the hypotheses $\mathcal{H}_0$: $\xi = \xi_0$ ($\xi_0 = 1$) versus $\mathcal{H}_1$: $\xi = \xi_1$ ($\xi_1 < 1$). This time, the power of a hypothesis test on ξ not only depends on d_h but also on the bias to respond "Yes", g.

Similar to Case 2, the effect size for this case can be calculated based on (23), this time with $T = 3$ and $\pi_{high} = \pi_{low} = \pi_{lure} =$

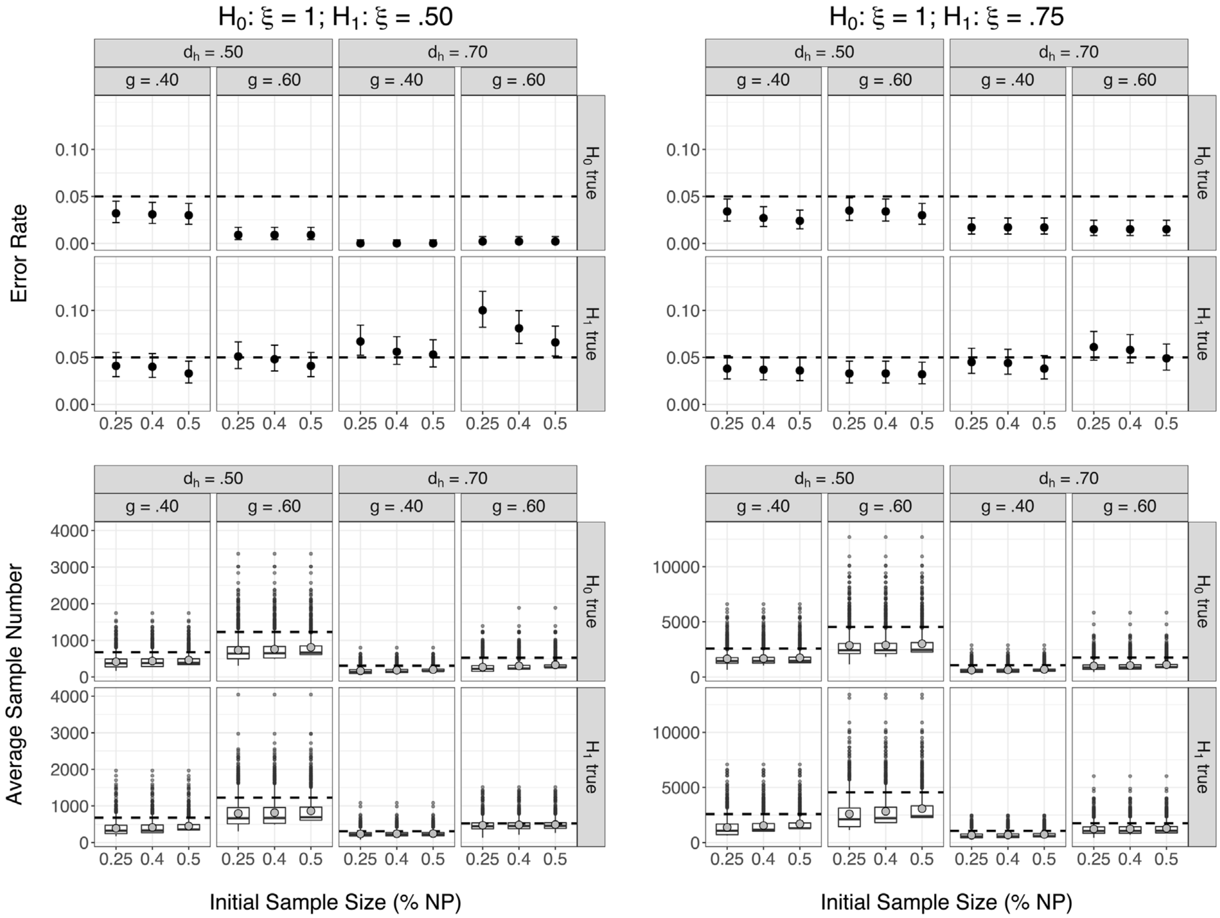

Fig. 7. Empirical error rates and sample size distributions of the sequential maximum likelihood ratio test as a function of the hypothesis tested, the true detection parameter d_h, the true guessing parameter g, and the data-generating scenario. Error bars denote 95% Clopper–Pearson exact confidence intervals (Clopper & Pearson, 1934). Black dots in the boxplots denote outliers (data points more than 1.5 times the inter-quartile range below or above 1st or 3rd quantile). Grey dots denote mean sample sizes. Dashed lines represent nominal error rates and sample sizes required by a corresponding Neyman–Pearson (NP) test.

.33. When testing $\xi_0 = 1.00$ versus $\xi_1 = .75$ while assuming $d_h = .70$ and $g = .50$ for the nuisance parameters under $\mathcal{H}_1$ (while treating d_h and g as free parameters under $\mathcal{H}_0$), the effect size is $w = 0.09$. A classical one-tailed asymptotic test with $\alpha = \beta = .05$ would thus require $N = 1335$ observations. However, if the participant has a slight bias to respond with "Yes" under uncertainty, $g = .60$, the effect size is reduced to $w = 0.08$ and the same test would require about $N = 1752$ observations. For $g = .40$, in contrast, the required sample size reduces to about $N = 1059$. Also taking into account different possible values of d_h would further increase the number of possible power analyses, thus illustrating the difficulty of determining a reasonable sample size for classical hypothesis tests in MPT models with more than one unknown nuisance parameter.

In the SMLRT, in contrast, we only need $\hat{\xi}$ and $\mathcal{V}_{\xi\xi}$ (or $SE_{\hat{\xi}}$), as the test is based on (12) and (13). The uncertainty with respect to the actual values of the nuisance parameters is taken into account implicitly, since $\mathcal{V}_{\xi\xi}$ and, correspondingly, the standard error of $\hat{\xi}$ depend in general on the precision and values of all parameter estimates. In the same vein, further increasing the complexity of the model in terms of the number of nuisance parameters or experimental conditions would not alter the general procedure for testing hypotheses on ξ.

As shown in the previous simulation, however, the SMLRT requires a sufficiently large initial sample size. If the sample size is too low, error rates may be inflated. If it is too large, the test may be less efficient. To illustrate that the required initial sample size may depend on the values of the nuisance parameters, we simulated the SMLRT for Case 3. The settings of the simulation were essentially identical to those in the previous simulation. Additionally, we varied the guessing parameter $g = .40$ versus $g = .60$. As the experiment in Case 3 comprises three stimulus categories (high versus low magnitude targets and lures), the sample size was increased by $+3$ in each step, one observation per stimulus category. For each parameter combination, 1000 replications were simulated.

The results are displayed in Fig. 7. For all simulated parameter combinations, the test shows very low rates of Type 1 errors. At the same time, however, the ASN in this case is still on average 38% smaller than the Neyman–Pearson sample size. The empirical β' closely approximate the nominal error rate for almost all parameter combinations. Only when $d_h = .70$ and $g = .60$, the test of $\xi_1 = .50$ yields too large β' when the initial sample size is smaller than $.50 \cdot N_{NP}$. Across all parameter combinations, the SMLRT is on average 34% more efficient than a Neyman–Pearson test and terminates with a smaller sample in 88% of the cases.

As our simulations show, the general procedure of the SMLRT extends to models with more than one unknown nuisance parameter. However, we also see the importance of a sufficiently large initial sample size in this case. When both d_h and g are large, the model predicts very low probabilities of "No" responses. In case of a large expected effect such as $\xi_1 = .50$, the classical Neyman–Pearson test is already quite efficient. Consequently, an initial sample based on 25% or 40% of N_{NP} is so small that the risk of extremely small cell frequencies is high. In such a case, the asymptotic approximations upon which the SMLRT is based cannot be expected to hold (cf. Cox & Roseberry, 1966).

For example, if $d_h = .70$ and $g = .60$, a classical one-tailed test requires $N = 174$ observations per tree to test $\xi_0 = 1$ versus $\xi_1 = .50$ with $\alpha = \beta = .05$. Thus, an initial sample size of 25% of the Neyman–Pearson sample size would comprise $N = 44$ observations per tree only. Conditional on the assumed values of d_h and g, the expected number of incorrect responses for high-magnitude targets in this case is only $44{\cdot}(1-.70){\cdot}(1-.60) = 5.28$. Not surprisingly, the large-sample approximations on which the SMLRT is based do not hold in such a situation. This could be remedied by further increasing the initial sample size. In that case, however, the test would no longer be more efficient than a classical test procedure.

It is important to note, however, that a case in which the classical test is already so efficient that the SMLRT cannot satisfy the nominal error rates with smaller samples is of practical relevance only if we can place high confidence in the parameter assumptions we make. Under uncertainty, we would rather rely on conservative assumptions to ensure sufficient power. If we follow this advice, the SMLRT will in general be more efficient.

4. Discussion

Hypothesis tests on parameter constraints in MPT models often rely on NHST, thus ignoring statistical power. Although power analyses have been worked out for categorical data (Erdfelder et al., 2005) and are readily available in existing software (e.g., Faul et al., 2009; Moshagen, 2010), practitioners typically face two challenges. First, to determine the effect size for a hypothesis test on a single parameter in a multi-parameter MPT (or other) model, the population values of all other parameters must be known or specified a priori based on theoretical considerations that may or may not hold. As a remedy, one can perform multiple power analyses for a range of reasonable parameter values and then choose the most conservative one. However, this strategy often fosters a second challenge for practitioners, namely, that the required sample sizes may become very large and practically infeasible.

As a remedy, in the present article we suggest to rely on sequential tests, an efficient alternative to classical statistical methods for hypothesis tests in MPT models. Sequential hypothesis tests control error probabilities of statistical decisions just as classical Neyman–Pearson tests do. Yet, at the same time, they are based on continuous monitoring of the data as they are sampled and terminate as soon as the data contain sufficient evidence for one hypothesis vis-à-vis the other. Thus, on average, sequential tests require notably smaller samples than classical methods that are based on a priori defined sample sizes (Schnuerch & Erdfelder, 2019).

We introduced the SPRT (Wald, 1947) and demonstrated how it is easily applied to analysis of MPT models with a single free parameter. We showed that it is substantially more efficient than classical Neyman–Pearson tests, requiring about 50% smaller samples on average. However, although there are applications of single-parameter MPT models in the literature (e.g., models for the randomized response technique; see Ulrich, Schröter, Striegel, & Simon, 2012), MPT models that are commonly used in cognitive psychology typically contain more than one parameter, many of which are nuisance parameters that need to be estimated.

Therefore, we introduced an extension of the SPRT suggested by Cox (1963) for sequential tests of composite hypotheses. In the SMLRT, the likelihood ratio is constructed based on ML estimates of both the test-relevant and the nuisance parameters. The sequential procedure is then corrected for the additional estimation uncertainty, such that the resulting test does not exceed long-run error rates α and β. Hence, the test procedure controls error probabilities without requiring knowledge or a specification of the exact values for the unknown nuisance parameters in the statistical model.

We illustrated how the SMLRT can be used to test hypotheses on MPT model parameters with existing MPT software. Essentially, the procedure merely requires the ML estimate $\hat{\theta}$ of the test-relevant parameter and the expected Fisher information (or the standard error of the estimate). Moreover, the SMLRT does not only remedy the problem of unknown nuisance parameters, it also increases efficiency of hypothesis testing. We demonstrated by means of simulations that the SMLRT requires on average 34% (Case 3) to 45% (Case 2) smaller samples to satisfy the same or even lower error rates compared with classical Neyman–Pearson tests even when these are based on the true, data-generating values of the nuisance parameter (which is an unlikely assumption in practice).

The sequential approach can be particularly useful in individual assessments (e.g., clinical diagnosis). As part of Bill Batchelder's proposal of cognitive psychometrics – that is, building a bridge between the fields of mathematical psychology and psychometrics – he strongly promoted the use of MPT models in the context of psychological assessment (Batchelder, 1998). For instance, he identified the great potential for substantive MPT model applications as diagnostic tools in clinical settings (Batchelder & Riefer, 1999; Riefer, Knapp, Batchelder, Bamber, & Manifold, 2002). However, Batchelder also acknowledged the obvious drawback of reduced estimation precision and low statistical power as a consequence of the small number of data points on the individual level. The sequential approach we promote in this article may facilitate the application of MPT models in individual assessments whenever it is necessary to make decisions about the presence or absence of specific cognitive symptoms while controlling error probabilities. More generally, we hope that the SMLRT for MPT models will further contribute to the increasing number of substantive applications in cognitive psychometrics.

Apart from individual assessment, sequential analysis is also particularly useful for efficient MPT modeling of data on the group level when each participant provides only a single data point (e.g., Heck, Thielmann, et al., 2018; Klauer et al., 2007; Moshagen et al., 2014, 2012; Schild et al., 2019). This setting has the advantage that the MPT analysis need not be built into the experimental procedure or software because the data can be analyzed after data collection for each participant. It is thus easily implemented in practice, thereby providing an attractive alternative to classical methods in terms of a more efficient and less costly control of error probabilities.

4.1. Limitations

The approaches presented in this article are so-called unrestricted sequential procedures that do not have a definite upper bound of sample size. Hence, although the test is on average more efficient than classical procedures, there is a potential risk that the data provide inconclusive evidence in single cases, meaning that the test will continue for a long time without reaching one of the two boundaries. Concluding from our simulation results, this risk is small (approximately 6% in Case 2, 12% in Case 3). Nevertheless, this risk potentially limits its applicability in individual analysis to situations in which the number of data points is not restricted a priori. Think, for example, of an experimental paradigm assessing long-term episodic memory processes (e.g., Batchelder & Riefer, 1986). Such a paradigm typically includes a learning phase and a test phase. The number of possible data points in the test phase is limited by the number of items learned during the first phase. Thus, sequential analysis during the test phase will make sense only if it requires no more than the number of learned items. As this obviously cannot be guaranteed, the unrestricted sequential approach is not appropriate for such applications.

Second, the SMLRT for composite hypotheses is based on large-sample approximations (Cox, 1963). Therefore, as the simulation results in Case 3 showed, the method may fail when initial sample sizes are too small (see also Cox & Roseberry, 1966; Wetherill, 1975). The relevance of a sufficiently large initial sample size increases with model complexity, as does the required sample size at termination. The practical challenge is of course to determine a suitable initial sample size for the sequential procedure on a priori grounds. If the sample size is too small, error rates might be seriously inflated. If it is too large, on the other hand, the test's efficiency is reduced (although our simulations demonstrated that the increase in ASN due to larger initial sample sizes is only slight).

As a remedy, we suggest to search for a model-specific minimum sample size by means of an a priori power analysis and Monte Carlo simulations. Of course, this will again entail assumptions about reasonable true parameter values of the nuisance parameters. However, the consequences of an overly conservative assumption in the context of a sequential test are much less severe than for a standard test procedure. If the initial sample size is chosen too large, the evidence provided by the data may already be compelling very early during data collection, meaning that the test procedure will stop immediately. Thus, the SMLRT will be more efficient than a correspondingly conservative classical test, in which one cannot use optional stopping even if the data clearly speak in favor of one of the hypotheses.

Third, sequential approaches assume that observations are independent and identically distributed (i.i.d.). This assumption is reasonable for sequential analyses of data generated by an individual provided that the experimental design prevents contaminations of the data by exercise effects, fatigue effects, or order effects. The i.i.d. assumption is also plausible in model applications where each participant provides a single data point only. If, however, MPT models are applied to aggregate data of repeated observations of multiple individuals, the i.i.d. assumption may be questioned and is often implausible (Smith & Batchelder, 2008). If there is heterogeneity in items or participants, ignoring the hierarchical structure might bias parameter estimates and statistical tests (Heck, Arnold, & Arnold, 2018). Thus, if parameter tests are performed at the group level based on data aggregated across items and participants, the sequential approaches promoted herein may not be suitable. This issue is especially critical for sequential tests if the data are collected in a batch-wise fashion. For instance, if one first collects 100 observations from a person that does perform extraordinarily well, the sequential test may already indicate a decision, thereby ignoring data from other participants that perform worse.

Finally, it is important to keep in mind that both SPRT and SMLRT address problems of hypothesis testing, not estimation. In this article, we focused on efficiency in statistical decision making exclusively. If this is the primary concern, SPRT and SMLRT are appropriate alternatives to classical methods. However, if the aim is to estimate a parameter as precisely as possible, these sequential procedures are not suitable. Whereas efficiency requires to make a decision with as few observations as possible, high precision of parameter estimates is achieved with as many observations as possible (without optional stopping depending on the current value of the estimates). In fact, parameter estimates following a sequential hypothesis test may be biased (Whitehead, 1986). Thus, the sequential approach promoted herein should only be used if the aim is in fact to make an efficient statistical decision, for example, in psychological assessments.

4.2. Conclusion

Multinomial processing tree models have proven useful in many areas of cognitive and social psychology as tools to measure and disentangle latent cognitive processes. As repeatedly argued and demonstrated by Bill Batchelder, they have great potential especially for psychometric purposes, for example, in the context of individual diagnostics in clinical settings (e.g., Batchelder, 1998; Batchelder & Riefer, 1999; Riefer et al., 2002). We introduced sequential test procedures proposed by Cox (1963) and Wald (1947) and illustrated how they can be adapted to MPT model analysis. By means of simulations, we demonstrated the excellent properties of the sequential approach for testing hypotheses on MPT model parameters both in the absence and presence of nuisance parameters. Thereby, we hope to improve efficiency of statistical inference in MPT modeling, particularly in the context of individual assessments (i.e., cognitive psychometrics) and other settings with scarce resources.

Appendix A. Sequential maximum likelihood ratio tests

To show that (11) is asymptotically equivalent to (9), consider that according to ML theory, the expected Fisher information matrix for a sample of size n is given by

$$n\mathcal{I}(\theta,\phi) = E\left[-\mathbf{H}(\theta,\phi)\right] = E\left[-\begin{pmatrix} \dfrac{\partial^2\ell(\theta,\phi;\ x^n)}{\partial\theta^2} & \dfrac{\partial^2\ell(\theta,\phi;\ x^n)}{\partial\theta\partial\phi} \\ \dfrac{\partial^2\ell(\theta,\phi;\ x^n)}{\partial\phi\partial\theta} & \dfrac{\partial^2\ell(\theta,\phi;\ x^n)}{\partial\phi^2} \end{pmatrix}\right] \tag{27}$$

where $\mathbf{H}(\theta,\phi)$ denotes the Hessian matrix of second-order partial derivatives. Accordingly, $n\mathcal{I}_{\theta\theta}$ and $n\mathcal{I}_{\theta\phi}$ denote the (θ,θ) and (θ,ϕ) element (or submatrix) of this matrix. Moreover, $\hat{\theta}$, $\hat{\phi}$ asymptotically satisfy the following equation (Cox, 1963):

$$n\left[\mathcal{I}_{\theta\theta}(\hat{\theta}-\theta) + \mathcal{I}_{\theta\phi}(\hat{\phi}-\phi)\right] = \frac{\partial\ell(\theta,\phi;\ x^n)}{\partial\theta}\ . \tag{28}$$

Thus, writing (9) in terms of (27) and (28) gives

$$n(\theta_1-\theta_0)\mathcal{I}_{\theta\theta}(\hat{\theta}-\theta)\ +\ n(\theta_1-\theta_0)\mathcal{I}_{\theta\phi}(\hat{\phi}-\phi)\ -\ \tfrac{1}{2}(\theta_1-\theta_0)(\theta_1+\theta_0-2\theta)n\mathcal{I}_{\theta\theta} - (\theta_1-\theta_0)(\hat{\phi}-\phi)n\mathcal{I}_{\theta\phi} \tag{29}$$

which by application of simple calculus yields

$$\begin{aligned} &n(\theta_1-\theta_0)\Big[\mathcal{I}_{\theta\theta}(\hat{\theta}-\theta)\ +\mathcal{I}_{\theta\phi}(\hat{\phi}-\phi) \\ &\quad-\ \tfrac{1}{2}\mathcal{I}_{\theta\theta}(\theta_1+\theta_0-2\theta) - \mathcal{I}_{\theta\phi}(\hat{\phi}-\phi)\Big] \\ &= n(\theta_1-\theta_0)\mathcal{I}_{\theta\theta}(\hat{\theta}-\theta\ -\tfrac{1}{2}\theta_1-\tfrac{1}{2}\theta_0+\theta) \\ &= n\mathcal{I}_{\theta\theta}(\theta_1-\theta_0)\left[\hat{\theta}-\tfrac{1}{2}(\theta_1+\theta_0)\right]\ . \end{aligned} \tag{30}$$

Appendix B. Properties of the sequential probability ratio test

To approximate the functions describing power and expected sample size of the sequential probability ratio test (SPRT) for a test of hypotheses about d in the MPT model displayed in Fig. 1 (with $g = .50$), we can use formulae derived by Wald (1947). For any given d_0, d_1, α, and β, the power of the SPRT is a function of the true value d. Let Ψ_d denote the probability to accept $\mathcal{H}_1$ given a certain true value d, then

$$\Psi_d \approx \frac{1-\left(\dfrac{\beta}{1-\alpha}\right)^h}{\left(\dfrac{1-\beta}{\alpha}\right)^h - \left(\dfrac{\beta}{1-\alpha}\right)^h}\ , \tag{31}$$

where h is the non-zero root of the equation

$$p\left(\frac{p_1}{p_0}\right)^h + (1-p)\left(\frac{1-p_1}{1-p_0}\right)^h = 1 \tag{32}$$

with p and p_i denoting the true and predicted probability of a correct response under hypothesis i, respectively, $p_i = d_i + (1 - d_i) \cdot .50$.

It is easy to see that if $d = d_1$, which means that $p = p_1$, the non-zero root of (32) is $h = -1$,

$$\begin{aligned}&p_1\frac{p_0}{p_1} + (1-p_1)\frac{(1-p_0)}{(1-p_1)} - 1\\ &= p_0 + (1-p_0) - 1\\ &= 0\ ,\end{aligned} \tag{33}$$

which, as expected, yields

$$\begin{aligned}\Psi_{d=d_1} &= \frac{1 - \left(\dfrac{1-\alpha}{\beta}\right)}{\left(\dfrac{\alpha}{1-\beta}\right) - \left(\dfrac{1-\alpha}{\beta}\right)}\\ &= \frac{\alpha+\beta-1}{\beta} \cdot \frac{\beta(1-\beta)}{\beta\alpha - (1-\beta)(1-\alpha)}\\ &= 1-\beta.\end{aligned} \tag{34}$$

In the same vein, if $d = d_0$ the non-zero root of (32) is $h = 1$,

$$\begin{aligned}&p_0\frac{p_1}{p_0} + (1-p_0)\frac{(1-p_1)}{(1-p_0)} - 1\\ &= p_1 + (1-p_1) - 1\\ &= 0\ ,\end{aligned} \tag{35}$$

which yields

$$\begin{aligned}\Psi_{d=d_0} &= \frac{1 - \left(\dfrac{\beta}{1-\alpha}\right)}{\left(\dfrac{1-\beta}{\alpha}\right) - \left(\dfrac{\beta}{1-\alpha}\right)}\\ &= \frac{1-\alpha-\beta}{1-\alpha} \cdot \frac{\alpha(1-\alpha)}{(1-\alpha)(1-\beta) - \alpha\beta}\\ &= \alpha\ .\end{aligned} \tag{36}$$

In a second step, the expected sample size at termination as a function of the true value d can be approximated by

$$E_d(N) \approx \frac{\Psi_d \log\left(\dfrac{1-\beta}{\alpha}\right) + (1-\Psi_d)\log\left(\dfrac{\beta}{1-\alpha}\right)}{p \log\left(\dfrac{p_1}{p_0}\right) + (1-p)\log\left(\dfrac{1-p_1}{1-p_0}\right)}\ , \tag{37}$$

where Ψ_d is given by (31).

References

Bamber, D., & van Santen, J. P. H. (2000). How to assess a model's testability and identifiability. *Journal of Mathematical Psychology*, *44*, 20–40. http://dx.doi.org/10.1006/jmps.1999.1275.

Barnard, G. A. (1952). The frequency justification of certain sequential tests. *Biometrika*, *39*, 144–150. http://dx.doi.org/10.2307/2332473.

Bartlett, M. S. (1946). The large-sample theory of sequential tests. *Mathematical Proceedings of the Cambridge Philosophical Society*, *42*, 239–244. http://dx.doi.org/10.1017/S0305004100022994.

Batchelder, W. H. (1998). Multinomial processing tree models and psychological assessment. *Psychological Assessment*, *10*, 331–344. http://dx.doi.org/10.1037/1040-3590.10.4.331.

Batchelder, W. H., & Riefer, D. M. (1986). The statistical analysis of a model for storage and retrieval processes in human memory. *British Journal of Mathematical and Statistical Psychology*, *39*, 129–149. http://dx.doi.org/10.1111/j.2044-8317.1986.tb00852.x.

Batchelder, W. H., & Riefer, D. M. (1990). Multinomial processing models of source monitoring.. *Psychological Review*, *97*, 548–564. http://dx.doi.org/10.1037/0033-295X.97.4.548.

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, *6*, 57–86. http://dx.doi.org/10.3758/BF03210812.

Blackwell, H. R., Pritchard, B. S., & Ohmart, J. G. (1954). Automatic apparatus for stimulus presentation and recording in visual threshold experiments. *Journal of the Optical Society of America*, *44*, 322–326. http://dx.doi.org/10.1364/JOSA.44.000322.

Breslow, N. (1969). On large sample sequential analysis with applications to survivorship data. *Journal of Applied Probability*, *6*, 261–274. http://dx.doi.org/10.2307/3211997.

Chechile, R. A. (2009). Pooling data versus averaging model fits for some prototypical multinomial processing tree models. *Journal of Mathematical Psychology*, *53*, 562–576. http://dx.doi.org/10.1016/j.jmp.2009.06.005.

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, *26*, 404–413. http://dx.doi.org/10.2307/2331986.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. http://dx.doi.org/10.1037/0033-2909.112.1.155.

Cox, D. R. (1952). Sequential tests for composite hypotheses. *Mathematical Proceedings of the Cambridge Philosophical Society*, *48*, 290–299. http://dx.doi.org/10.1017/S030500410002764X.

Cox, D. R. (1963). Large sample sequential tests for composite hypotheses. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, *25*, 5–12.

Cox, C. P., & Roseberry, T. D. (1966). A large sample sequential test, using concomitant information, for discrimination between two composite hypotheses. *Journal of the American Statistical Association*, *61*, 357–367. http://dx.doi.org/10.2307/2282824.

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Zeitschrift für Psychologie / Journal of Psychology*, *217*, 108–124. http://dx.doi.org/10.1027/0044-3409.217.3.108.

Erdfelder, E., Faul, F., & Buchner, A. (2005). Power analysis for categorical methods. In B. S. Everitt, & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1565–1570). Chichester, UK: Wiley.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160. http://dx.doi.org/10.3758/BRM.41.4.1149.

Heck, D. W., Arnold, N. R., & Arnold, D. (2018). TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods*, *50*, 264–284. http://dx.doi.org/10.3758/s13428-017-0869-7.

Heck, D. W., Erdfelder, E., & Kieslich, P. J. (2018). Generalized processing tree models: Jointly modeling discrete and continuous variables. *Psychometrika*, *83*, 893–918. http://dx.doi.org/10.1007/s11336-018-9622-0.

Heck, D. W., Thielmann, I., Moshagen, M., & Hilbig, B. E. (2018). Who lies? A large-scale reanalysis linking basic personality traits to unethical decision making. *Judgment and Decision Making*, *13*, 356–371.

Hu, X. (1999). Multinomial processing tree models: An implementation. *Behavior Research Methods, Instruments, & Computers*, *31*, 689–695. http://dx.doi.org/10.3758/BF03200747.

Hu, X., & Batchelder, W. H. (1994). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika*, *59*, 21–47. http://dx.doi.org/10.1007/BF02294263.

Hu, X., & Phillips, G. A. (1999). GPT.EXE: A powerful tool for the visualization and analysis of general processing tree models. *Behavior Research Methods, Instruments, & Computers*, *31*, 220–234. http://dx.doi.org/10.3758/BF03207714.

Joanes, D. N. (1972). Sequential tests of composite hypotheses. *Biometrika*, *59*, 633–637. http://dx.doi.org/10.1093/biomet/59.3.633.

Klauer, K. C. (2006). Hierarchical multinomial processing tree models: A latent-class approach. *Psychometrika*, *71*, 7–31. http://dx.doi.org/10.1007/s11336-004-1188-3.

Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika*, *75*, 70–98. http://dx.doi.org/10.1007/s11336-009-9141-0.

Klauer, K. C., Stahl, C., & Erdfelder, E. (2007). The abstract selection task: New data and an almost comprehensive model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 680–703. http://dx.doi.org/10.1037/0278-7393.33.4.680.

Knapp, B. R., & Batchelder, W. H. (2004). Representing parametric order constraints in multi-trial applications of multinomial processing tree models. *Journal of Mathematical Psychology*, *48*, 215–229. http://dx.doi.org/10.1016/j.jmp.2004.03.002.

Ly, A., Marsman, M., Verhagen, J., Grasman, R. P., & Wagenmakers, E.-J. (2017). A tutorial on Fisher information. *Journal of Mathematical Psychology*, *80*, 40–55. http://dx.doi.org/10.1016/j.jmp.2017.05.006.

Matthes, T. K. (1963). On the optimality of sequential probability ratio tests. *The Annals of Mathematical Statistics*, *34*, 18–21. http://dx.doi.org/10.1214/aoms/1177704239.

Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods*, *42*, 42–54. http://dx.doi.org/10.3758/BRM.42.1.42.

Moshagen, M., Hilbig, B. E., Erdfelder, E., & Moritz, A. (2014). An experimental validation method for questioning techniques that assess sensitive issues. *Experimental Psychology*, *61*, 48–54. http://dx.doi.org/10.1027/1618-3169/a000226.

Moshagen, M., Musch, J., & Erdfelder, E. (2012). A stochastic lie detector. *Behavior Research Methods*, *44*, 222–231. http://dx.doi.org/10.3758/s13428-011-0144-2.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London A (Mathematical and Physical Sciences)*, *231*, 289–337. http://dx.doi.org/10.1098/rsta.1933.0009.

R Core Team (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, http://www.R-project.org/.

Read, T. R. C., & Cressie, N. A. C. (1988). *Goodness-of-fit statistics for discrete multivariate data*. New York, NY: Springer.

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, *95*, 318–339. http://dx.doi.org/10.1037/0033-295X.95.3.318.

Riefer, D. M., Knapp, B. R., Batchelder, W. H., Bamber, D., & Manifold, V. (2002). Cognitive psychometrics: Assessing storage and retrieval deficits in special populations with multinomial processing tree models. *Psychological Assessment*, *14*, 184–201. http://dx.doi.org/10.1037/1040-3590.14.2.184.

Rushton, S. (1950). On a sequential t-test. *Biometrika*, *37*, 326–333. http://dx.doi.org/10.2307/2332385.

Schild, C., Heck, D. W., Ścigała, K. A., & Zettler, I. (2019). Revisiting REVISE: (Re)Testing unique and combined effects of REminding, VIsibility, and SElf-engagement manipulations on cheating behavior. *Journal of Economic Psychology*, *75*, 102161. http://dx.doi.org/10.1016/j.joep.2019.04.001.

Schnuerch, M., & Erdfelder, E. (2019). Controlling decision errors with minimal costs: The sequential probability ratio t test. *Psychological Methods.* Advance online publication, http://dx.doi.org/10.1037/met0000234.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. http://dx.doi.org/10.1177/0956797611417632.

Singmann, H., & Kellen, D. (2013). MPTinR: Analysis of multinomial processing tree models in R. *Behavior Research Methods*, *45*, 560–575. http://dx.doi.org/10.3758/s13428-012-0259-0.

Smith, J. B., & Batchelder, W. H. (2008). Assessing individual differences in categorical data. *Psychonomic Bulletin & Review*, *15*, 713–731. http://dx.doi.org/10.3758/PBR.15.4.713.

Smith, J. B., & Batchelder, W. H. (2010). Beta-MPT: Multinomial processing tree models for addressing individual differences. *Journal of Mathematical Psychology*, *54*, 167–183. http://dx.doi.org/10.1016/j.jmp.2009.06.007.

Swets, J. A., Tanner, W. P., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review*, *68*, 301–340. http://dx.doi.org/10.1037/h0040547.

Ulrich, R., Schröter, H., Striegel, H., & Simon, P. (2012). Asking sensitive questions: A statistical power analysis of randomized response models. *Psychological Methods*, *17*, 623–641. http://dx.doi.org/10.1037/a0029314.

Wald, A. (1947). *Sequential analysis*. New York: Wiley.

Wald, A., & Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, *19*, 326–339. http://dx.doi.org/10.1214/aoms/1177730197.

Wetherill, G. B. (1975). *Sequential methods in statistics* (2nd ed.). London: Chapman and Hall.

Whitehead, J. (1986). On the bias of maximum likelihood estimation following a sequential test. *Biometrika*, *73*, 573–581. http://dx.doi.org/10.1093/biomet/73.3.573.