Journal of Mathematical Psychology 92 (2019) 102278

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

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

Review

Adaptive design for systems factorial technology experiments☆

Joseph J. Glavan [a,*], Elizabeth L. Fox [a], Mario Fifić [b], Joseph W. Houpt [a]

[a] Department of Psychology, Wright State University, Dayton, OH, United States
[b] Psychology Department, Grand Valley State University, Allendale, MI, United States

ARTICLE INFO

Article history:
Received 2 August 2018
Received in revised form 5 August 2019
Available online 18 September 2019

Keywords:
Systems factorial technology
Adaptive experimental design

ABSTRACT

Systems factorial technology (SFT) is a powerful framework for examining how people use multiple sources of information together. Unfortunately, it is often difficult to apply. Appropriate manipulation of the salience of each source of information is critical to assessing processing characteristics so a significant amount of time can be spent piloting to determine the correct levels. Even with piloting, some participants' data ends up unusable due to individual differences. We first examine the use of an accuracy-only adaptation for stimulus levels, based on the Psi method. In some cases a focus entirely on accuracy may be insufficient, particularly given that response time (RT) is the primary measure with SFT. Hence, we also introduce an approach to adapting stimulus levels for each individual participant's joint accuracy and RT. This will increase the likelihood that salience manipulations will be effective and that a participant's data will be usable.

Contents

1. Introduction

In many situations, choices and actions depend on multiple sources of perceptual and cognitive information. As such, one of the fundamental endeavors in cognitive science is determining the qualitative properties associated with using those multiple sources of information. Systems factorial technology (SFT;

☆ This research was supported in part by the Consortium Research Fellowship Program.

* Corresponding author.
E-mail address: glavan.3@wright.edu (J.J. Glavan).

https://doi.org/10.1016/j.jmp.2019.102278
0022-2496/

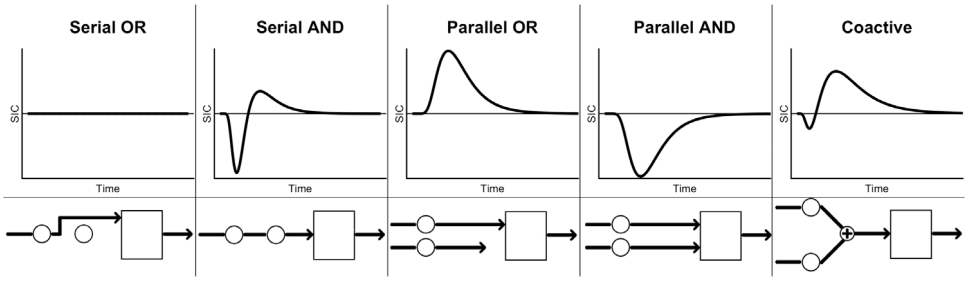

Fig. 1. Predicted survivor interaction contrast for serial, parallel, and coactive models with both OR and AND stopping rules. Corresponding processing schematics for each class of model are depicted below the SICs.

Townsend & Nozawa, 1995) is a framework based on rigorous experimental design and nonparametric statistical tools that can evaluate the underlying processing mechanisms of a system operating on multiple sources of information. While the theory is quite general and gives clear diagnostic information, SFT is used by a relatively small number of researchers. One of the reasons for the limited employment of SFT is that it is often difficult to implement. In this paper, we address one source of difficulty in applying SFT, the determination of appropriate stimulus salience levels. Here, salience refers to the perceptual strength of a stimulus. SFT experiments commonly leverage salience manipulations to discern how multiple sources of information are processed together. We endeavor to ameliorate this difficulty by proposing two approaches for adapting salience levels for each source of information to the individual subject: one based on task accuracy and one based on a joint model of accuracy and response time (RT). Before detailing these approaches, we first outline the characteristics of cognitive systems that are of interest to SFT.

Consider a detection task where the operator is asked to make a button press whenever they hear a tone (auditory signal) or see a light flash on a display (visual signal). There are various ways in which the cognitive system may turn the information from the two signals into a single decision whether to respond. The system may process both signals simultaneously. Alternatively, it may first decide whether the auditory signal is present before it processes the visual signal (or vice versa). These examples of parallel and serial processing, respectively, represent two general classes of models based on their temporal organization or architecture. Orthogonal to a system's processing architecture is its stopping rule. In the current example, the cognitive system may terminate as soon as it has finished processing either the auditory signal or the visual signal; this is known as a self-terminating stopping rule. When only two sources of information are used, it may also be called a first-terminating (OR) rule. While the stopping rule is often constrained by task instructions, the task does not always force a specific stopping rule. In the present example, the operator may still fully process both the auditory and visual signals using an exhaustive (AND) stopping rule, even if the detection of a single signal is sufficient to respond accurately.

In addition to architecture and stopping rule, we may also be concerned with the independence of each processing channel. In the previous example, evidence for the presence of the auditory signal may cause the visual signal to be processed faster (or slower). In the extreme, evidence for each signal could be pooled into a common processing channel. This special case of parallel processing, known as coactive processing, makes only a single decision so that the question of stopping rule becomes undefined. Schematic diagrams of independent parallel, serial, self-terminating, and exhaustive models as well as coactive models are shown in the second row of Fig. 1.

To infer these qualities in cognitive systems, SFT includes a series of measures. The measures of interest to us in the current work require the processing speed of each source of information to be selectively increased and decreased. This is commonly achieved by factorially manipulating the salience of each source of information, colloquially referred to as the double factorial paradigm (DFP). The assumption is that manipulating the saliency of each source of information (e.g., loudness, brightness) influences how fast it is processed. If the salience manipulations effectively and selectively influence the processing rate of each corresponding source of information, the resulting response time distributions will be ordered such that responses on trials where all sources of information are highly salient will be faster than responses on trials where some of the sources of information are less salient. Responses to trials where all sources of information are relatively less salient should be slower than responses to trials where one or more sources of information are more highly salient. Formally, we desire to find stimuli that yield response time distributions such that $S_{HH}(t) < \{S_{HL}(t), S_{LH}(t)\} < S_{LL}(t)$, where 'H' reflects a high salience manipulation and 'L' reflects a low salience manipulation. The $S(t)$ are survivor functions,[1] which in the present context describe the probability that a response has not been made by time t. In the earlier example, $S_{HL}(t)$ is the survivor function estimated for the experimental condition in which the auditory signal is fairly loud (H) but the visual signal is relatively dim (L). For more information about the DFP, its associated trial rates, and how it can be applied to different kinds of experiments see Fifić and Little (2017) and Houpt, Blaha, McIntire, Havig, and Townsend (2014).

SFT relies on two measures for inferring the architecture and stopping rule of a cognitive system. The first, the Mean Interaction Contrast (MIC; Eq. (1)), can discriminate between parallel and serial processing architectures, and for parallel architectures, it is able to diagnose stopping rules. The logic behind the MIC is the same as that of the interaction term in a typical ANOVA where mean RTs are factorially contrasted. It requires that the processing rate of each source of information be manipulated without affecting the processing of the other(s) (i.e. *selective influence*).

$$\mathrm{MIC} = \left[\overline{\mathrm{RT}}_{LL} - \overline{\mathrm{RT}}_{LH}\right] - \left[\overline{\mathrm{RT}}_{HL} - \overline{\mathrm{RT}}_{HH}\right] \quad (1)$$

In a serial model, the effect of the selective influence manipulations on the mean RTs are additive such that MIC $= 0$. A parallel model predicts a nonzero MIC with self-terminating models being over-additive (MIC > 0) and exhaustive models being under-additive (MIC < 0).

In the present paper, we are primarily concerned with the second measure from SFT, which is related to the first.[2] The Survivor Interaction Contrast (SIC; Eq. (2)) is also based on a double difference of RTs; however, whereas the MIC is a value, the SIC is a functional measure, which makes it more informative. The SIC can be used to discriminate between self-terminating

[1] $S(t) = 1 - F(t)$ where $F(t)$ is the cumulative distribution function of t.
[2] $\mathrm{MIC} = \int_0^\infty \mathrm{SIC}(t)\,dt$.

and exhaustive serial models in addition to the models that can be discriminated with the MIC. The canonical combinations of architecture and stopping rule predict distinctive SIC shapes (Fig. 1). A serial-OR model produces a flat SIC for all t, and a serial-AND model produces a first negative and then positive SIC where the area under the curve equals zero. A parallel-OR model results in an all positive SIC, and a parallel-AND model results in an all negative SIC. The SIC also conveys some information about stochastic independence between processing channels (Houpt & Townsend, 2011; Townsend & Nozawa, 1995). While the previously described models all assume stochastic independence, a coactive model produces a first negative and then positive SIC where the sum under the curve is positive (MIC $>$ 0; see also Eidels, Houpt, Altieri, Pei, & Townsend, 2011).

$$\mathrm{SIC}(t) = [S_{LL}(t) - S_{LH}(t)] - [S_{HL}(t) - S_{HH}(t)] \quad (2)$$

Statistical tests for interpreting an individual's MIC and SIC exist (Houpt & Fifić, 2017; Houpt, MacEachern, Peruggia, & Townsend, 2016; Houpt & Townsend, 2010) and are most powerful when salience levels are used that maximally separate the marginal response time distributions for each subject. However, in practice this is difficult to achieve. Instead, it is more common to specify a single set of salience levels to use throughout an SFT experiment for all subjects. The problem with this approach is that there are low-level perceptual (e.g. acuity) and higher-level cognitive (e.g. strategy) effects that vary across individuals such that even with extensive pilot testing one cannot always determine a single set of intensities that are appropriate for every subject. This inevitably leads to weakened or inconclusive results because the differences between response time distributions will be smaller or undetectable. Worse yet, resources may be wasted if the experimenter chooses to run additional subjects and discard the data from subjects who were insensitive to the salience manipulations. This may even bias the study's conclusions toward subpopulations that are sensitive to the chosen salience levels.

To demonstrate the dangers of using a single set of group-level salience levels when individual differences are present, we collected some pilot data from a simple visual search task where participants had to find a target stimulus amongst homogeneous distractors. We focus on detection accuracy instead of response times for the sake of brevity, but the lesson still holds. In the left column of Fig. 2 we examine manipulations of the color difference between the target and distractors while in the right column we examine manipulations of the difference in orientation (i.e. rotation) between the target and distractors. In the top row we estimate individuals' probability of responding correctly as the respective stimulus dimension varies. These psychometric functions vary not only between people (different colored lines), but also within the same person on a separate day (same colored lines). Notice how certain intensities (e.g. distractors rotated 60° from target stimulus) are simultaneously too low to be detected by some individuals yet too high to be missed by others. If we average the curves in the top row across individuals and session for each respective stimulus dimension, we can pick high and low salience levels for the group (depicted by black vertical lines in the top row of Fig. 2). In the bar graphs in the bottom of Fig. 2, we plot the expected accuracy of each subject in each session. For some subjects, these salience levels lead to high accuracy, but for others they do not. Clearly, a single set of salience levels will not be effective for this group of subjects. While these results illustrate the potentially catastrophic effect individual differences can have in terms of accuracy, the same type of problem can occur in RTs. The goal of our current effort is therefore to present a method for choosing individualized salience levels that yield clear RT differences without sacrificing accuracy. This will highlight variation across participants while increasing experimental efficiency.

2. Accuracy focused approaches to determining salience levels

When choosing stimulus intensities for a DFP, we want to manipulate salience such that the previously discussed survivor function ordering, $S_{HH}(t) < \{S_{HL}(t), S_{LH}(t)\} < S_{LL}(t)$, holds. Naively, one may wish to choose the greatest disparities in salience possible so as to maximize the chances of statistically detecting the needed differences; however, one must also be aware of how the salience manipulations affect accuracy because we can only use the RTs from correct responses when calculating SICs. Assuming that accuracy and response times are negatively correlated, if we choose a very low intensity for a particular condition, we may reliably increase processing time, but we may also incur many incorrect responses, which waste experimental resources because we cannot use these trials. On the other hand, choosing a higher intensity stimulus may ensure most responses are correct, but we may still waste experimental resources if processing is not slowed enough to yield detectable ordering among the RT distributions. One approach for using accuracy to determine appropriate salience levels is to choose an intensity that maximizes accuracy for the high salience condition and then find a low salience value that compromises on accuracy. For example, we have previously used 90% (Fox & Houpt, 2016). Other researchers could choose to use alternative levels depending on their needs and resources.

A psychometric function maps the intensity of a stimulus to the probability of responding appropriately. Choosing a fixed level of accuracy and estimating the corresponding stimulus intensity for an individual provides a way to obtain a similar level of perceptual difficulty across participants. Accuracy thresholds are a standard way for controlling individual differences in psychophysics and other areas of experimental psychology.

There are a number of classical methods for determining thresholds that date back to the earliest days of psychophysics. One example is the method of constant stimuli, where subjects make repeated judgments over a discrete number of intensities. The resulting response rates for each intensity are then treated as point-estimates of the underlying psychometric function, and a parametric model such as a Weibull distribution can be fit to the points so that a continuum of thresholds may be extrapolated. The method of constant stimuli allows one to be very confident in the thresholds obtained because there is direct correspondence between how the psychometric function is estimated from past responses and how it is used to predict future responses: the subject is likely to respond at the same rate to a given intensity regardless of whether the intensity is being used to estimate or elicit the response rate. A significant shortcoming of the method of constant stimuli is that it requires many trials to confidently measure the response rate at a particular stimulus intensity because responses must be averaged. Furthermore, multiple stimulus intensities are needed to fit a full psychometric function, but not all of the intensities sampled will be equally informative. In short, the method of constant stimuli is very inefficient.

An alternative approach is to use any one of various adaptive psychophysical methods that use an algorithm to intelligently sample informative stimulus intensities. Probably the most well-known nonparametric adaptive method is the truncated staircase. The standard staircase (called a 1-up 1-down staircase) targets the 50% threshold. On each trial if the subject responds positively (negatively) then the next trial is placed at one step lower (higher) in stimulus intensity. When a reversal occurs (a change from responding positively to negatively or vice versa), the step size is reduced, usually by half. The experimenter determines step sizes, and the most common stopping rule for staircases is the number of reversals. The advantage to staircase methods is

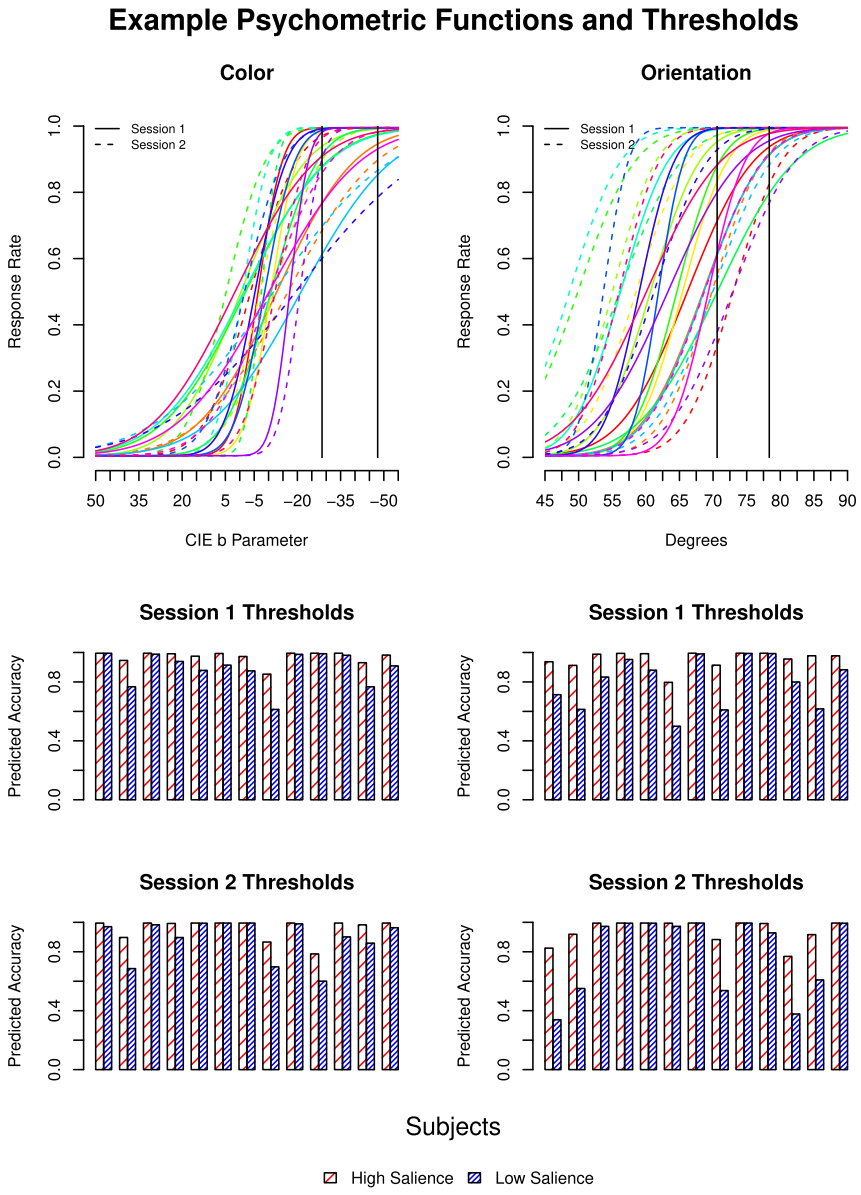

Fig. 2. Psychometric functions (top row) by subject and session for a visual search task where participants utilize either color (left column) or orientation (right column) to detect a target among distractors. Vertical black lines indicate the 99% threshold (high salience) and 90% threshold (low salience) obtained by averaging over psychometric function location and slope parameters for each subject and session. The bottom bar plots depict the accuracy predicted for each subject in each session when these high and low salience intensities are used. Notice that the single set of group level intensities causes poor performance in one or both salience conditions for some subjects, while for others it produces ceiling performance in both salience conditions.

that they are relatively easy to set up and run. However, merely targeting the 50% threshold is not ideal for estimating a subject's SIC as half of the low saliency trials (on average) will be incorrect and discarded.

Methods for expanding the staircase method to target different thresholds have been developed and used with success. Levitt (1971) proposed a transformed staircase that uses a set of previous trials to determine when to change the stimulus intensity. By triggering reversals after a particular sequence of repeated responses, the transformed staircase method is able to target a limited, discrete set of thresholds. A common example of this, the 1-up 2-down staircase, targets 70.7% accuracy. However, each targeted level of accuracy requires more trials to estimate it the further it is from 50%. In order for a transformed staircase to converge on 90.6% accuracy, reversals would need to occur following every negative response and after a sequence of seven positive responses.[3] If one wishes to target any arbitrary threshold they can use a method proposed by Kaernbach (1991). This weighted up–down method transforms the step sizes themselves instead of transforming the number of responses needed before taking a step. The step sizes are adjusted according to the formula $\delta_U = \delta_D(1-\varphi)/\varphi$ where δ_U is the up-step size, δ_D is the down-step size, and φ is the targeted threshold. The weighted staircase method works well for thresholds near 50%, but like the transformed method it too becomes unwieldy at extreme threshold values: in order to obtain a 90% threshold the down-step size becomes nine times the up-step size. One can see how these sporadic jumps in step size might alert the subject to the purpose of the psychophysical experiment, compromising the validity of the

[3] In Levitt (1971)'s notation, the probability of a sequence for the down group is $[P(X)]^7 = 0.5$, which corresponds to the probability of a positive response: $P(X) = 0.906$.

threshold obtained. Rammsayer (1992) suggested that interleaving multiple such staircases might diminish the chance of their realization, but he was also quick to point out that doing so would eliminate any efficiency the experimenter would gain by using adaptive methods.

A popular alternative adaptive method, based on parametric assumptions about the psychometric function, is QUEST (Watson & Pelli, 1983). QUEST uses Bayes' rule to combine the information from all previous trials with any prior information the experimenter may have about the psychometric function (e.g., from previous research, pilot subjects, etc.). It then places the next trial intensity at the most likely threshold value. Watson and Pelli (1983) outline procedures for terminating QUEST when a certain confidence interval has been achieved; they also suggest that the procedure may be terminated after a set number of trials. QUEST converges quickly, and Watson and Pelli (1983) report an efficiency of 84% over 128 trials.[4]

QUEST relies on a prespecified functional form for the psychometric function (the Weibull distribution is a common choice). Parameterizing the psychometric function like this requires the experimenter to assume the values of some terms, making the overall implementation more difficult than nonparametric procedures like the staircase method. There are four parameters needed to specify the psychophysical model: a guess parameter γ determines the chance level of performance and is typically defined by the type of task (e.g. $\gamma = 0.5$ for two alternative forced choice); a lapse parameter δ defines the negative response rate at maximum intensity, and its complement $(1 - \delta)$ determines the accuracy ceiling; a location parameter α determines the placement of the psychometric function along the stimulus intensity axis; and a slope parameter β determines how quickly performance goes from chance to ceiling. QUEST estimates the location parameter and requires the experimenter to choose slope and lapse parameters from other sources. This is a problem for us because we wish to use high accuracy thresholds, which will be strongly affected by the choice of slope parameter. Furthermore, the slope parameter may vary by subject. Because QUEST does not provide a way for estimating individualized slopes, it may not be the ideal candidate for pairing with SFT.

Another alternative is the Psi method (Kontsevich & Tyler, 1999). Like QUEST, Psi is a parametric Bayesian method that is able to incorporate prior knowledge about the psychometric function from previous research and information from all previous trials. Unlike QUEST, Psi provides estimates of both location and slope parameters. It does so by constructing a matrix of possible location and slope values the experimenter believes to include the true location and slope pair, then estimating the probability that each pair is the true pair using Bayes' rule. On each trial, the expected entropy of each possible stimulus intensity is calculated and the intensity corresponding to the minimal entropy is selected for testing. This is different from QUEST because instead of the most likely threshold being tested next, the intensity that will provide the most information about the psychometric function is tested next. The resultant location and slope pair is finally determined using expected a posteriori estimation once a termination criterion is reached. Kontsevich and Tyler (1999) acknowledge that confidence level testing can be used to determine when to terminate the Psi method, but they recommend terminating after a set number of trials to both ensure that the subject's effort is evenly distributed and provide certainty to the experimenter of when the search will end. This last assurance is important to the overall efficiency of the adaptive SFT experiments we wish to conduct because obtaining the most precise threshold possible is not the focus; the focus is to quickly determine useful values for the saliency conditions of the SIC experiment. Kontsevich and Tyler (1999) report that Psi typically requires less than 30 trials to accurately estimate location and around 300 trials to estimate both location and slope.

Like QUEST, Psi leaves the lapse parameter to be set by the experimenter; however, this is a much smaller problem than unaccounted for individual variability in the slope parameter. The majority of lapses typically occur in the first few trials as the subject settles into the task, and Kontsevich and Tyler (1999) recommend excluding these trials to reduce the lapse rate. One could think of these early trials as practice. After the reduction is performed, lapse rates should be small even at conservative values, and thus any variance in δ across individuals is most likely negligible. For this reason, it should not be a problem to specify a single (even conservative) error rate for the experiment population. Because of Psi's ability to quickly and efficiently estimate the parameters of the psychometric function that will most greatly vary by individual, it seems to be suitable for pairing with SFT.[5]

3. Joint RT–accuracy approaches to determining salience levels

While the Psi method can be used to determine accuracy-based salience levels, it does not take into account RT, which is ultimately the measure required for the SIC. One clear approach to using RT is to mimic the approach taken for accuracy: assume a parametric form for the relationship between salience and RT (i.e., the chronometric curve) then estimate the parameters of that function using either an offline method (method of constant stimuli) or adaptive procedure. The range of candidate salience levels from the accuracy-focused and RT focused results can then be combined. This approach was used for group level salience by Blaha, Houpt, McIntire, Havig, and Morris (manuscript in preparation). RT and accuracy are rarely assumed to be independent measures, so the relationship between stimulus salience and accuracy and the relationship between salience and RT are likely mutually informative. Furthermore, looking at the combination of response time and accuracy eliminates the bias due to variation in speed–accuracy trade-off across individuals and conditions. This combination also allows for a joint criteria threshold of speed and accuracy such that the model estimates what the best stimulus intensity level is for an individual to obtain a desired level of accuracy and response time. This can be particularly useful in the SFT paradigm as accuracy levels must remain high (e.g. $\geq$ 90%) while response times are manipulated.

There are multiple well-known models for jointly evaluating RT and accuracy. Perhaps the most well known is the drift–diffusion model (DDM) originally proposed by Ratcliff (1978, see also Stone, 1960). This model assumes the choice process can be represented by a biased diffusion process that leads to a response whenever the diffusion hits a boundary of a prespecified region. The choice is specified by the particular boundary reached by the diffusion and the RT is specified by the amount of time taken to reach the boundary (an additional random variable meant to indicate perceptual and motor processes is usually included). The magnitude of the bias of the diffusion represents the relative amount of information indicating a response in the biased direction. The distance the diffusion must travel to reach a boundary represents cautiousness. Indeed, this model has even

[4] Compare to 40%–50% for the PEST procedure (Taylor & Creelman, 1967), which is not discussed in this paper because it accomplishes the same as QUEST but with lower efficiency.

[5] It is worth noting that Psi has recently been generalized to handle multidimensional psychometric functions and additional trial outcomes. The advanced method, known as QUEST+ (Watson, 2017), may prove useful in future applications of our methodology.

been applied in similar circumstances to parametrically map between signal intensity and the strength of the perceptual information (e.g., Mulder, Wagenmakers, Ratcliff, Boekel, & Forstmann, 2012; Murphy, Vandekerckhove, & Nieuwenhuis, 2014).[6]

Because our goal is to simplify the data collection process for SFT, we would ideally avoid a large number of trials and additional selective influence manipulations tangential to the estimation of the SIC. We chose to use a more streamlined joint RT–accuracy model, the log-normal race model (LNRM; Rouder, Province, Morey, Gomez, & Heathcote, 2015). Many joint RT–accuracy models are directly concerned with relative variation in specific parameters. For this application, we are not interested in the latent parameters per se, thus the distinction among effects on average drift rate, threshold, and bias are less of a concern. The LNRM sacrifices these distinctions in favor of reduced computational complexity (in contrast with Mulder et al., 2012). The LNRM models the response process as a race between random variables representing the time that it would take to choose each of the available options. Whichever option is the fastest on a trial is the observed choice. For example, with two choices, a and b, the LNRM would associate a random variable with the time to respond to each, T_a and T_b. On any trial on which an observation of T_a is smaller than T_b, the model responds a; on any trial on which an observation of T_b is smaller than T_a, the model responds b. The response time is modeled as the time taken by the fastest choice plus an additional variable (ψ) that does not depend on the choice which represents the time taken by non-decision processes (e.g., executing the response once a choice is made). Given this setup, the joint distribution of choice a and response time t is given by

$$f(a, t; \mu_a, \mu_b, \sigma_a, \sigma_b) = g(t - \psi; \mu_a, \sigma_a)\left[1 - G\left(t - \psi; \mu_b, \sigma_b\right)\right].$$

The final core assumption of the LNRM is that the choice duration random variables have log-normal distributions. With $\Phi()$ indicating the standard normal cumulative distribution function,

$$g(t; \mu, \sigma) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln t - \mu)^2}{2\sigma^2}\right]$$
$$G(t; \mu, \sigma) = \Phi\left(\frac{\ln t - \mu}{\sigma}\right).$$

One of the strengths of the LNRM indicated by Rouder et al. (2015) is that the μ parameter can be used to build additional complexity into the model, particularly with respect to the relationship between the stimulus and the choice and response time probabilities.[7] For example, $\mu_a - \mu_b$ could be modeled with an increasing function of the evidence in favor of a to represent the increasing probability of choosing a and the higher probability of faster responses.

For our purposes, we would like to approximate the relationship between salience and the difference between the correct and incorrect μ parameters. Given this relationship, we can then use the inverse of that relationship to estimate the stimulus level that would lead to a targeted level of performance.

[6] A frequently used alternative accumulator model is the linear ballistic accumulator model (LBA; Brown & Heathcote, 2008), which has the same general form, but replaces the diffusion process with a linear information change over time. This linear form leads to an analytically simpler model. Because neither RT nor accuracy directly inform the instantaneous properties of the information accumulation process, the LBA and drift–diffusion models often make similar predictions and lead to similar qualitative conclusions (Donkin, Brown, & Heathcote, 2011). Unfortunately, both the drift–diffusion model and the LBA tend to require a large number of trials and parameter-specific selective influence manipulations to identify parameter values.

[7] The σ parameter could also be used to model the influence of experimental variables; however, we do not explore that option herein.

4. Summary of our approach

Experiments utilizing the SFT methodology are inherently time consuming because the functional statistics employed operate on distributions of RTs, and many observations (trials) need to be collected in order to sufficiently estimate these distributions. As highlighted previously, this presents a considerable sunk cost when subjects must be excluded or replaced before analysis because of variability in their sensitivity to the experimental manipulations. Our proposed solution is to first fit a psychophysical model to each subject and then use the thresholds recommended by the model as salience levels in the SFT experiment; however, a similar concern regarding the duration of an experimental session arises when one considers that this approach requires additional observations to be collected on top of the many needed to calculate the SFT statistics. We have attempted to address this catch by proposing the use of powerful adaptive methods that are able to quickly estimate a subject's psychometric function from relatively few trials.

In the following sections, we demonstrate the use of two different methods that a researcher may consider taking within our general procedure. The first is the Psi method (Kontsevich & Tyler, 1999), which has been in use for nearly twenty years and is available in multiple open-source and commercial software packages (e.g., Python, MATLAB, etc.). The second is the lognormal race model (Rouder et al., 2015), which is based on a more complex model than the Psi method, but has the advantage of using RT and accuracy jointly.

We first conduct a series of simulations to show the asymptotic performance of these approaches. By tracing the convergence of each psychophysical models' parameters as a function of the number of trials, we identify recommended practices and characterize the efficacy a practitioner may expect if they were to use fewer (or more) trials. We then use our recommendations to conduct a simulated DFP study and examine the sensitivity of the methods to individual differences.

We next perform an experiment with human subjects as a proof of concept. At the beginning of each experimental session, participants complete a psychophysical block and then complete a visual search task in a DFP using salience values extracted from the respective psychophysical model. Whereas the simulation study represents the ideal way to administer our proposed method, we recognize that SFT studies are typically conducted in environments where participants are recruited and compensated based on one hour sessions. We thus design the human study to include these practical limitations and use a minimally reasonable number of trials in order to describe the robustness of our approach to the sub-optimal conditions in which it is more likely to be used.

5. Simulations

We use simulation data to demonstrate the application of the accuracy-only and the joint RT–accuracy approaches to a SFT study. To stand in for human subjects, we use an R implementation of the DDM (Molenaar, Tuerlinkcx, & van der Maas, 2015; Ratcliff, 1978) to generate choice and RT data. We explicitly do not want the generating model to match the measurement models underlying the analyses because it is quite likely that human performance is not in perfect agreement with the measurement models. We want to show that even with a mismatched model, the recommended salience levels lead to improved SFT testing. All simulation code is available at https://github.com/jhoupt/adaptiveSFT.

For the DDM, we set the mean drift rate ξ to be a deterministic function of the stimulus intensity and a scaling parameter z.

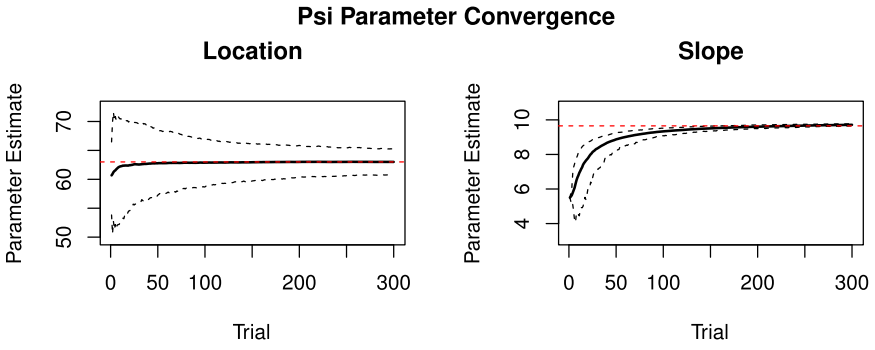

Fig. 3. Convergence of the Psi psychometric function parameters across trials. The upper dashed line indicates the upper ninety-fifth percentile and the lower dashed line indicates the fifth percentile. The solid line is the mean estimate across models.

Specifically, $\xi = z \times$ intensity. All other model parameters are fixed. We sample stimulus intensities differently for each approach during the psychophysical blocks.

The application of the Psi approximation to the DDM used the following sequence:

1. Psi generated an initial stimulus intensity x_0 based on the specified prior.
2. A single trial from the DDM was sampled with the fixed parameters and drift rate $\xi = zx_i$.
3. The accuracy of that single trial was fed into the Psi algorithm, which in turn updated the posterior distribution over the Psi parameters accordingly.
4. The updated posterior was used to determine the most informative new intensity level
5. Steps 2 through 4 were repeated for the specified number of trials.

The application of the LNRM followed a slightly different pattern:

1. For each sample stimulus intensity x_i, n response times and choices were simulated from the DDM with the fixed parameters and $\xi = zx_i$.
2. The posterior distribution of the LNRM parameters was then estimated using Stan (Carpenter et al., 2017).

We first examine the rate of parameter convergence for each of the approaches. These values can help to guide researchers on the number of trials needed to calibrate the stimuli for an individual to be used in a DFP study. Next, we use the salience levels estimated by the psychophysical approaches to simulate a DFP for each of the combinations of architecture and stopping rule discussed before (Fig. 1). We also use random perturbations of the DDM parameters to simulate variation across individuals. Lastly, we provide guidelines to the practitioner for using each method.

5.1. Parameter convergence

5.1.1. Psi

We initialized the DDM with the following parameter values: starting point $= 0$, drift scaling parameter ($z = 1.60$), variability of the accumulation process ($s = 0.25$), non-decision time ($t_{er} = 0.10$), and symmetrical upper and lower decision boundaries $(a, b; (a - b)/2 = 1.45)$. We verified that the DDM with these parameters generated RTs that were qualitatively similar to typical human RT distributions.

We ran 500 simulations of the DDM through 300 trials of a psychophysical task using Psi with a fixed lapse parameter[8] ($\delta = .01$). On each trial of each simulation we extracted Psi location (α) and slope (β) parameters and estimated a psychometric function. We evaluated the rate that the location and slope parameters stabilized in Fig. 3. The left plot indicates Psi overcame the initialization bias after about 25 trials and obtained stable estimates of the true location parameter (red dashed line) after 75 trials. The right plot in Fig. 3 illustrates that Psi took longer ($\sim$40 trials) to recover from initialization bias in the slope parameter such that stabilization occurred after about 150 trials.[9] Fig. 4 shows the joint convergence of the psychometric function. As expected, estimates were more accurate as the number of sample trials increased.

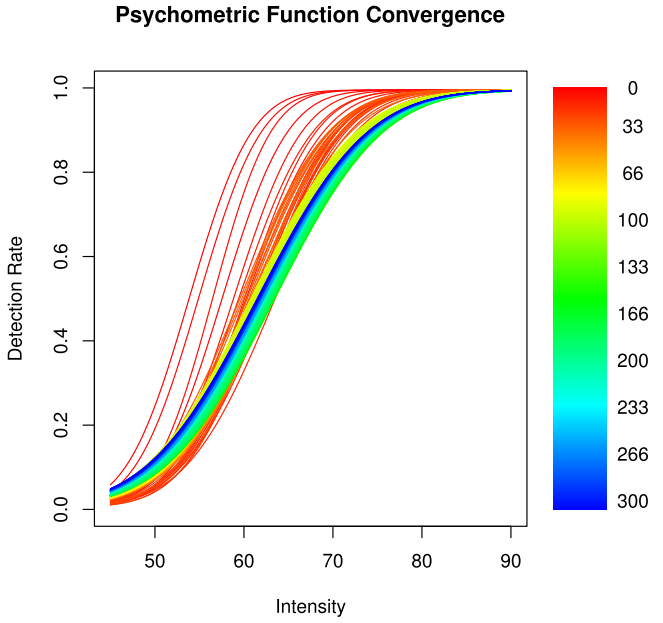

Fig. 4. Convergence of the Psi psychometric function across trials. This function maps stimulus intensity to the probability of a correct response. The estimate better approximates the true probability of a correct response for a given intensity as the number of trials increases from 1 (red) to 300 (blue).

5.1.2. LNRM

We generated data from the DDM using the method of constant stimuli, then fit the LNRM to these data to examine the relationship between precision and number of trials. To determine stimulus salience levels for the DFP, the most important

[8] See the earlier discussion of assuming conservative lapse rates in Section 2.

[9] Kontsevich and Tyler (1999) originally found the slope parameter to require 300 trials to be accurately estimated, and they suggest that this efficiency may depend on the validity of the assumed lapse rate.

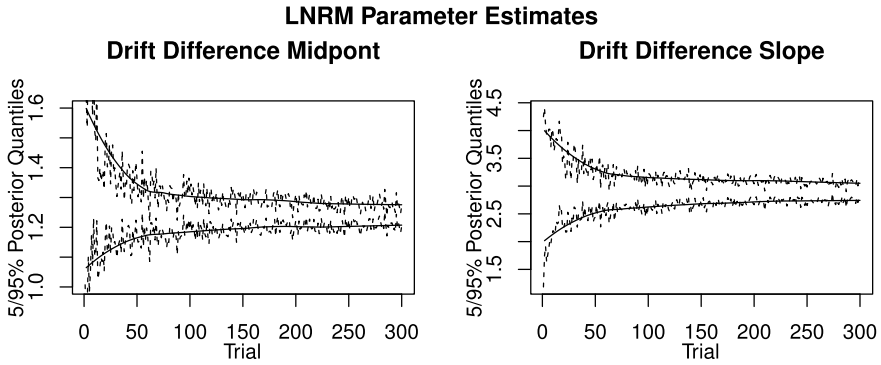

Fig. 5. Convergence across trials of the latent parameters of the LNRM function mapping the difference between average completion time of the competing response processes. The upper line indicates the ninety-fifth percentile and the lower line indicates the fifth percentile of the posterior distribution. Dashed lines give quantities from each simulation. The solid line is a smooth curve fit to those values.

parameters are those that map between the input salience and the difference between the mean completion times of correct and incorrect racers. For this paper, we assume this mapping follows a logistic function,

$$z_{\text{correct}}(x) - z_{\text{incorrect}}(x) = \frac{L}{1 + \exp\left(-k(x - x_0)\right)}. \tag{3}$$

Here L represents the maximum difference between the racers which should be set based on the scale of the response times in the task. The DDM parameters we used led to response times on the order of a few seconds, so we fixed L to 10. The midpoint parameter, x_0, determines the salience level at which the rate difference is half-way to its maximum (i.e., when $L = 10$, $z_{\text{correct}}(x_0) - z_{\text{incorrect}}(x) = 5$). This parameter and the precision of the prior should depend on the scale of the stimulus intensity that will be used. When the stimulus intensity is normalized, a standard normal prior is appropriate for the midpoint. The slope parameter, k, determines the rate of increase of the difference between the correct and incorrect racers, where higher k corresponds to faster increases. As a rate parameter, this depends on the ratio of the response time scale to the stimulus intensity scale, and thus the priors precision should be set correspondingly. We used a half-normal prior with mean 0 and standard deviation 2.

Fig. 5 indicates the rate of convergence of the posterior distribution over the parameters mapping stimulus salience to drift rate separation. Each plot shows the fifth and ninety-fifth percentile of the posterior distribution for simulated data with one to 300 trials at each of ten stimulus intensity levels. Note that this implies the total number of trials is ten times the number indicated on the x-axis, meaning the axes are not directly comparable to the accuracy-only approach. Because distinct simulated data was used for each number of trials, the functions are not smooth, although we overlay the quantiles with a smoothed function to indicate the general trend. For both parameters, the posterior precision, indicated by the distance between the quantiles, improves rapidly at first, then stabilizes around 50 trials. This indicates that, when a researcher has only a vague idea of what the parameter values will be and generic priors are used, that 50 to 100 trials per stimulus level (500 to 1000 total) should be targeted. In situations where more prior information is available, such as from pilot testing or other subjects, fewer trials are likely sufficient. Other factors, such as the overall variability of the participants' response times may also influence the convergence, so this number should be considered a heuristic estimate.

5.2. DFP simulation study

5.2.1. Psi

In order to simulate a DFP, we need to obtain high and low salience levels for each of the two stimulus dimensions of interest. For the low salience levels, we used the 90% thresholds recommended by Psi. For the high salience levels, we chose to use the maximum physical intensity of the stimulus. This promotes peak performance in the case that accuracy plateaus before RT has been minimized. Of course, one could always set the high salience level based on the psychometric function estimated by Psi at no additional cost, but this intensity would be limited to the threshold corresponding to $(1 - \delta)$.

Using the DDM from the previous section as our observer model, we first simulated 100 psychophysical trials using the Psi method for each of the two stimulus dimensions and extracted thresholds to use as low salience intensities. We then simulated 100 trials for each of the critical DFP conditions: HH, HL, LH, and LL; for each potential combination of architecture and stopping-rule: parallel-OR, parallel-AND, serial-OR, and serial-AND.

Before attempting to interpret a SIC, we must first confirm that the salience manipulations effectively and selectively influenced the corresponding cognitive processes. While this is necessarily the case for our simulations, direct tests of selective influence can be difficult with humans. The standard approach is to test for a necessary condition of selective influence, the aforementioned survivor distribution ordering: $S_{HH}(t) < \{S_{HL}(t), S_{LH}(t)\} < S_{LL}(t)$. Because our accuracy-based approach relies on an assumed relationship between accuracy and RT, it is even more crucial that we verify that the salience manipulations led to the expected differences in RT distributions. We used the Kolmogorov–Smirnov null hypothesis test to compare each pair of survivor functions and found significant statistics indicating the appropriate ordering for all models. Accuracy also correlated with salience as expected. We plot the survivor functions and SICs corresponding to each model in Fig. 6. Compare these results to the examples in Fig. 1.

Next, we conducted statistical tests of the SIC. Following the recommendation of Houpt and Burns (2017), we use $\alpha = .33$. Parallel-OR and serial-AND models produced significantly positive deviations from zero, parallel-AND and serial-AND models produced significantly negative deviations from zero, and the SIC generated from the serial-OR model did not significantly deviate from zero. Hence, the simulations produce the characteristic SIC functions for each standard model as expected (compare to Fig. 1).

Now confident that our approach will recover the architecture and stopping rule of a single parameterization of the generating model, we varied the previously fixed parameters of the DDM to simulate variation across individuals and repeated the previous procedure. We generated data from ten distinct sets of DDM parameters, representing ten unique subjects, for each combination of architecture and stopping rule. DDM parameters varied such that: a, b; $a - b \in [2.77, 3.90]$, $z \in [1.72, 2.79]$, $t_{er} \in [0.07, 0.13]$). We fixed the variability of the accumulation process ($s = 0.20$).

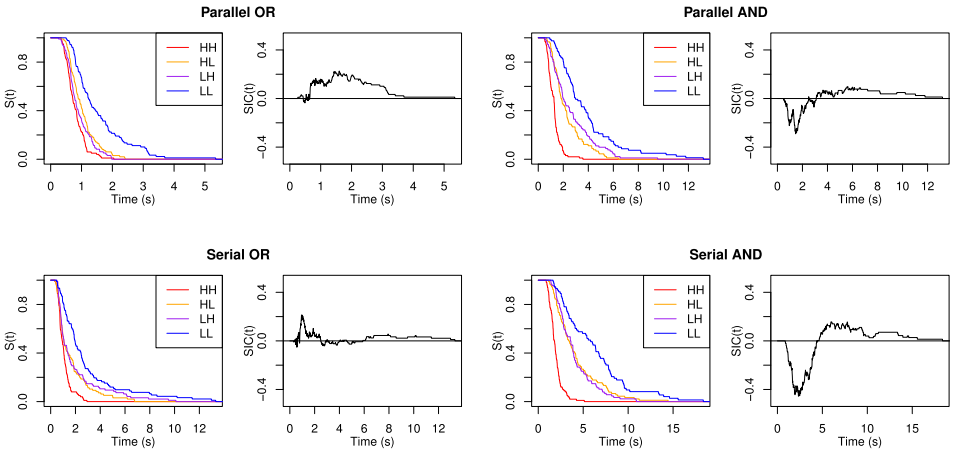

Fig. 6. Simulated survivor and SIC functions using the salience levels suggested by Psi. Models were constructed using each combination of parallel/serial architecture and self-terminating (OR)/exhaustive (AND) stopping rules.

We simulated 100 trials for each condition in the DFP as before and analyzed the data using the sft package (Houpt et al., 2014) in R (R Core Team, 2017).

For each generating model below, the survivor functions of all ten "subjects" were ordered appropriately according to a series of pair-wise Kolmogorov–Smirnov tests. That is, we can reject the null hypotheses that $S_{LL}(t) < \{S_{LH}(t), S_{HL}(t), S_{HH}(t)\}$ and $S_{HH}(t) > \{S_{HL}(t), S_{LH}(t), S_{LL}(t)\}$. Therefore, we can interpret the SIC and MIC of each subject.

For the parallel-OR models, the statistics for the positive extent of the SIC (D^+) ranged from 0.109 to 0.575. All were above the critical value for the SIC null-hypothesis test at $\alpha = .33$, and all but two were significant at $\alpha = .05$. The statistics for the negative part of the SIC (D^-) ranged from 0.0 to 0.109. Only one was significant at $\alpha = .33$, and none were significant at $\alpha = .05$. The MIC was significantly positive for all cases at $\alpha = .33$, nine of which were significant at $\alpha = .05$.

For the parallel-AND models, the statistics for the negative extent of the SIC ranged from 0.145 and 0.551. All were above the critical value at $\alpha = .33$, and eight were significant at $\alpha = .05$. The statistics for the positive extent of the SIC ranged from 0.029 to 0.081. None were significant at $\alpha = .33$. The MIC was significantly negative for all 10 cases at $\alpha = .05$.

For the serial-OR models, the statistics for the positive extent of the SIC ranged from 0.024 to 0.132. Two were significant at $\alpha = .33$, and none were significant at $\alpha = .05$. The statistics for the negative extent ranged from 0.034 to 0.124. One was significant at $\alpha = .33$ but not at $\alpha = .05$. The MIC was significantly different from zero for eight at $\alpha = .33$, two of which were also significant at $\alpha = .05$.[10]

For the serial-AND models, the statistics for the positive extent of the SIC ranged from 0.027 to 0.183. Four were significant at $\alpha = .33$, two of which were also significant at $\alpha = .05$. The statistics for the negative extent ranged from 0.151 to 0.495. All but two were significant at $\alpha = .05$, but all were significant at $\alpha = .33$. The MIC was significantly different from zero for six at $\alpha = .33$, five of which were also significant at $\alpha = .05$.

5.2.2. LNRM

Using the same simulation approach as above, with the DDM drift rate given by an affine transformation of the stimulus intensity, we simulated each architecture/stopping-rule combination when salience levels for the DFP were estimated with the LNRM. The basic process was similar to the parameter convergence simulations above. We first generated data from the DDM at a preselected set of stimulus intensity levels, then fit the LNRM to estimate the mapping between stimulus intensity and LNRM mean separation. Once an estimate of the mapping was obtained, we then inverted the mapping to estimate the appropriate salience levels to achieve a prespecified difference in mean separation for the high salience and low salience trials. In these simulations, we used a separation of $z_{\text{correct}} - z_{\text{incorrect}} = 8$ for high salience trials and 1.3 for low salience trials.

For the four standard models (parallel-OR, parallel-AND, serial-OR, serial-AND), we estimated the high and low salience levels with 100 trials at each of 10 salience levels. This yielded a high salience level of 1.71 and a low salience level of 0.580. Thus, for the DFP simulation, each subprocess was simulated with a drift rate $\xi_{high} = 1.71z$ for high salience trials and $\xi_{low} = 0.580z$ for low salience trials. The other parameters were fixed at the following: decision bounds $(a - b)/2 = 3$, $z = 2$, base time $t_{er} = 0.1$, and standard deviation of the drift $s = 0.2$. We used 100 trials per each combination of the salience levels to estimate the survivor functions and SICs. The simulations produced the characteristic survivor and SIC functions for each standard model as expected (compare Fig. 7 to the examples in Fig. 1.). These results are nearly identical to those produced by our simulations using the Psi method.

Next, we simulated data approximately following a typical SFT study. We generated data from ten distinct sets of DDM parameters, representing ten unique subjects, for each combination of architecture and stopping rule. Parameters were drawn from truncated normal distributions with means set based on the example simulations above, with the exception of a higher mean threshold (3.3) and mean drift rates set to have a drift rate to threshold ratio equal to that of the example models. Standard deviations were 1/8 the mean for all parameters, and distributions were truncated at zero. For each subject, 100 trials were simulated from each combination of salience levels. Data were analyzed using the `sft` package (Houpt et al., 2014) in R (R Core Team, 2017).

The first simulation was based on a parallel-OR system. Of the ten simulated subjects, none had statistically significantly misordered survivor functions, and only two did not reach statistically significantly ordered distributions for all pair-wise Kolmogorov–Smirnov tests. These two subjects' survivor functions were still ordered appropriately upon visual inspection so we interpret them with the others. The statistics for the positive extent of the

[10] The null hypothesis test used for the SIC (Houpt & Townsend, 2010) is stronger than the Adjusted Rank Transform (ART) test used for the MIC, Leys and Schumann (2010) and Reinach (1965). Whenever the two measures disagree, we defer to the SIC results.

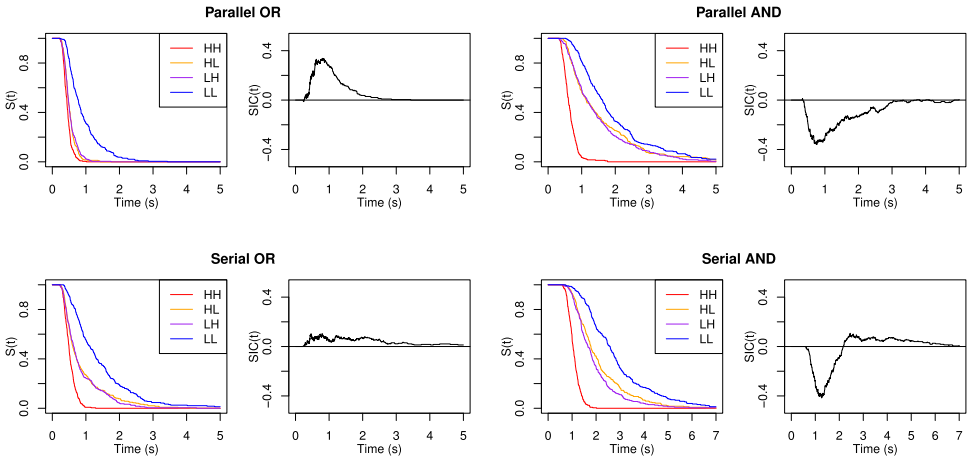

Fig. 7. Example survivor functions and SICs from a simulated DFP for each combination of stopping rule and architecture. Salience levels for the DFP were determined based on fitting a lognormal race model to data generated with a drift–diffusion model.

SIC ranged from 0.240 to 0.660. All were above the critical value for the SIC null-hypothesis test at $\alpha = .33$, and only one was not significant at $\alpha = .05$. The statistics for the negative part of the SIC ranged from 0.002 to 0.122. None were significant at $\alpha = .33$. The MIC was significantly positive for all 20 cases at $\alpha = .05$.

For the parallel-AND models, none of the survivor functions were significantly misordered, and only two were not significantly ordered. They were still visually ordered appropriately. The statistics for the negative extent of the SIC ranged from 0.340 to 0.609. All were significant at $\alpha = .33$. The statistics for the positive extent ranged from 0.011 to 0.141. None were significant at $\alpha = .33$. The MIC was significantly negative for seven at $\alpha = 0.33$, three of which were also significant at $\alpha = .05$.

For the serial-OR models, all but one subject passed the survivor function ordering tests. That one was visually ordered appropriately and not statistically significantly misordered. The statistics for the negative part of the SIC ranged from 0.030 to 0.211. Three were significant at $\alpha = .33$, but none were significant at $\alpha = .05$. The statistics for the positive extent ranged from 0.011 to 0.284. Two were significant at $\alpha = .05$. The MIC was significantly different from zero for nine at $\alpha = .33$, six of which were also significant at .05.

For the serial-AND models, all distributions passed the survivor function ordering tests. The statistics for the negative part of the SIC ranged from 0.208 to 0.659. One was not significant at $\alpha = .05$, but all were significant at $\alpha = .33$. The statistics for the positive extent ranged from 0.093 to 0.266. Seven were significant at $\alpha = .33$, two of which were also significant at $\alpha = .05$. The MIC was significantly different from zero for five at $\alpha = .33$, two of which were also significant at $\alpha = .05$.

5.3. Discussion of simulations

We used a drift–diffusion model to simulate human subjects in a perceptual decision making task. We used our model of the human to evaluate the convergence of parameters in each of our psychophysical models. We then simulated a full adaptive DFP study by estimating high and low salience values from either the accuracy-only or joint RT–accuracy methods and then using those values to estimate the respective SICs and MICs. Our efforts revealed that both methods led to adequate survivor function orderings, indicative of selective influence, and we successfully recovered the appropriate generating architecture and stopping rule for each virtual subject.

The Psi method and the LNRM method achieved nearly identical outcomes in our simulations. The number of trials needed to sufficiently fit the underlying psychophysical models was the primary difference between them. The Psi algorithm strategically chooses the next stimulus intensity to test based on the responses made so far, picking the intensities that will be most informative for estimating the model's parameters. In our joint RT–accuracy approach, we had to first collect data using the method of constant stimuli before fitting a LNRM. In this regard, the latter approach is much less efficient: despite using two sources of information, it does not collect the data in an optimal manner.

Based on our simulations, 100 to 150 trials for each stimulus dimension should be sufficient for the Psi method and 50 to 100 trials per level of intensity (500 to 1000 total) per dimension when using the LNRM approach should be sufficient. As stated above, these estimates are influenced by the specifics of our simulation, so the numbers should be considered approximate. Assuming 4 s per trial on average, Psi would need 15 to 20 min to complete both psychophysical blocks, whereas the LNRM approach would need 1 to 2 h. Hence, the accuracy-based approach is generally more practical, although one could drastically improve the efficiency of either method by specifying an informative prior distribution over the model parameters. It may also be possible to get away with using fewer than the above recommended number of trials. For example, we recommend using 150 trials for Psi because that is how many trials it takes for the slope (beta) parameter to converge, but after 40 trials both model parameters have overcome their initialization bias. At this point the psychophysical model fit may not be the best that it could be, but it may be good enough to set usable salience levels for the DFP. In the following section, we conduct a study with human subjects using far fewer trials (25 trials for Psi and 5 repetitions of 10 levels (50 trials total) for LNRM) and demonstrate that both methods remain fairly successful.

There is one substantial situation in which the experimenter should favor the joint RT–accuracy approach over the accuracy-only approach. The accuracy-only approach relies on the assumption that accuracy and response latency are negatively correlated with respect to stimulus salience — as salience decreases, accuracy decreases and RT increases. If this relationship was disrupted (e.g., by changing strategies) or relatively weak (e.g. so that accuracy would need to be sacrificed to an unacceptable degree to obtain sufficient RT differences), then the accuracy-only approach will fail. One should be especially wary of under motivated subjects and make sure that the task instructions are clear and consistent in their emphasis on speed vs. accuracy.

6. Demonstration with human subjects

6.1. Method

Having demonstrated the adaptive SFT experiment using the simulations above, we now set out to showcase the Psi and LNRM techniques with human subjects. Unless otherwise noted, the methods underlying the two conditions are exactly the same. The task we chose to use for the human study is adapted from our previous work with visual search (Glavan, Haggit, & Houpt, 2019). Subjects have to determine the presence of a target defined by two features, here color and orientation. On each trial, distractors are presented that differ from the target along one or both feature dimensions. We manipulate salience by adjusting the degree of dissimilarity between the target and distractors. The reader may notice that we collect fewer trials from the human subjects than we used in our simulations. Whereas the goal with the simulations was to propose recommendations for ideal conditions, in the current section we attempt to demonstrate the effectiveness of our techniques under more realistic time constraints.

6.1.1. Participants

We recruited eight undergraduate students to participate in the study, half of whom were assigned to the Psi condition and the other half to the LNRM condition. Each subject completed two one-hour sessions administered over consecutive days and was awarded class credit as compensation for their time. The study was conducted at Wright State University and approved by its Institutional Review Board. All participants gave written informed consent before beginning the study and indicated that they had normal or corrected to normal color vision and hearing, unencumbered use of both hands, and no history of epilepsy or brain trauma.

6.1.2. Materials

We conducted the study in a dark room (i.e. lights off, door shut, no windows, etc.) to control ambient light levels. We presented the task using PsychoPy (Peirce, 2009) on a 20″ Sony Trinitron monitor positioned 90 cm from the edge of a table at which the participants sat. The display spanned 40.5 cm (25.361 degrees of visual angle) across and 30.5 cm (19.234 degrees of visual angle) in height with a resolution of 1280 × 1024 pixels. Participants responded using the computer's optical mouse.

Each stimulus consisted of a circle with a line through its center, similar to the international prohibition sign that readers may recognize from "no parking" or "no smoking" signs. The diameter of each stimulus was 0.700 degrees of visual angle, and all line orientations are reported in degrees of counterclockwise rotation from horizontal.

The target was always red with 45 degrees of tilt. The choice of color was somewhat arbitrary, but it gave some connection to our prior work (Glavan et al., 2019). We chose to use an oblique angle to avoid the facilitation observers exhibit at completely vertical or horizontal (i.e. cardinal) orientations (Appelle, 1972).

We parameterized the color of our stimuli using the CIELAB color space in order to better correspond to human perceptual properties. This allowed us to control luminance by fixing the L parameter at 50. We also fixed the a parameter, which controls the green–red component, at 110. We manipulated the b parameter, which controls the blue–yellow component, to yield more red-like stimuli at higher values and more magenta-like stimuli at lower values.

Because PsychoPy does not currently support CIELAB, we translated code from Ruzon (2009) into Python to convert between CIELAB and RGB color spaces. This code is available at https://github.com/jhoupt/adaptiveSFT, along with all the other code described in this article.

The color used for targets was ($L = 50$, $a = 110$, $b = 110$), and the color used for the background of the search field was neutral gray ($L = 50$, $a = 0$, $b = 0$). In the Psi condition, we only estimated low salience values and fixed the high salience intensities at ($L = 50, a = 110, b = -50$) for color and 90 degrees for orientation. In the LNRM condition, we estimated both low and high salience intensities from the model.

6.1.3. Procedure

On the first day of the study, the experimenter guided the subject into a darkened room, whose only source of light came from the computer used in the experiment. Once seated at the computer, the experimenter explained the informed consent protocol and demographics survey. While the subject read and completed these forms on the computer, the experimenter stood on the other side of a divider to give the subject some privacy but remained in the room to avoid introducing extraneous light from outside the room. After the subject indicated that they had completed the informed consent process, the experimenter loaded the experiment. At this point, the luminary characteristics of the display matched those of the rest of the experiment, and the experimenter explained the task to the subject. After the subject completed a few practice trials, the experimenter quietly left the room after confirming that the subject did not have any questions.

Following the departure of the experimenter, the subject completed one psychophysical block where the distractors shared the target's color and one psychophysical block in which the distractors shared the target's orientation, requiring them to make orientation-only or color-only judgments, respectively. Within these blocks, 50% of the trials were catch trials, i.e., contained no targets. The third, final block used the salience levels estimated from the psychophysical blocks to create distractors that differed from the target in color and/or orientation according to the DFP. Each block was preceded by written instructions, example images of the target and distractors, and a set of practice trials. Ten high salience practice trials preceded the psychophysical blocks, and 16 practice trials (one for each of the possible trial conditions) preceded the DFP block. The practice trials provided feedback as to whether the subject's response on each trial was correct and included a brief explanation if it was incorrect.

We assumed that the time needed to fully explain the instructions, combined with the informed consent process, was sufficient for the subject to adapt to the lighting of the room, but to further ensure that adaptation would not confound our results, we always administered the orientation psychophysics block before the color psychophysics block because orientation judgments should be less sensitive to cone cell adaptation.

On each trial, the words "Get ready..." were presented for 1 s, followed by a field of randomly placed stimuli. The target was present on exactly half of all trials. If the target was absent, a distractor replaced it so that there were always 24 objects displayed. The distractors on a given trial had identical color and orientation; however, the color and orientation of the distractors were factorially varied across trials to obtain the salience manipulations needed to calculate a SIC. This resulted in 8 distractor types:[11]

- High salience color difference with high salience orientation difference (HH),
- High salience color with low salience orientation (HL),
- High salience color with target orientation (HA),
- Low salience color with high salience orientation (LH),

[11] Note that we did not include AA trials as this would have been a search array entirely composed of targets.

- Low salience color with low salience orientation (LL),
- Low salience color with target orientation (LA),
- Target color with high salience orientation (AH),
- Target color with low salience orientation (AL).

Each distractor type was equally likely.

We instructed subjects to determine whether the target was present "as quickly and accurately as possible". They responded positively by clicking the left mouse button and negatively by clicking the right mouse button. If they indicated that the target was absent, then the screen cleared and advanced to the next trial after a 1 s delay. If the subject responded that the target was present, then the stimuli would all change into black outlines of triangles at their respective positions. The mouse cursor, which was not normally visible, would appear in the center of the display, and the subject would have to left click on the triangle that corresponded to where they found the target. We only used this additional response procedure when scoring accuracy; all search times reported here reflect the latency of the initial present/absent response. The trial timed out if no response was made within 20 s, and these scratched trials were discarded from analysis.

In the Psi condition, the psychophysical blocks each terminated after the subject completed 50 trials (25 target-present and 25 target-absent) or after 5 min had elapsed, whichever came first. Whereas Psi adaptively selects the next trial intensity to use, the LNRM must be fit to choice-RT data post-hoc. Thus in the LNRM condition, we had subjects first complete a method of constant stimuli for each psychophysical block and then fit a model to obtain high and low salience intensities. Through pilot testing, we learned that using the same stopping criterion as Psi for the method of constant stimuli did not test enough different intensities, so we doubled the duration of the psychophysical blocks in the LNRM condition, terminating each after the subject completed 5 repetitions of 10 intensities for both present and absent targets (100 trials) or after 10 min had elapsed, whichever came first. Because of these disparities, subjects in the Psi condition completed more DFP trials (720 per session) than the subjects in the LNRM condition (576 DFP trials per session).

In both conditions, only the target-present trials were used to estimate salience levels. For the LRNM condition, this is not a problem because the method of constant stimuli specifies the intensities to use for the distractors on both target-present and target-absent trials; however, in the Psi condition, the adaptive algorithm only specifies intensities for the target-present trials. We were concerned that subjects may be able to use the magnitude of intensity change between trials as a cue for target presence, so on each target-absent trial we used the intensity from the last target-present trial plus some normally distributed random noise independently sampled from $\mathcal{N}(\mu = 0,\ \sigma = 5)$.

On the second day of the study, subjects followed the same procedure as they had for the first session with one exception. In place of the previously completed informed consent procedure, subjects sat quietly at the computer in the darkened room for five minutes to allow their eyes to adjust. After the experimenter reviewed the instructions with the subject and verified that they had no new questions, the subject completed the psychophysical and DFP blocks as before.

6.2. Results

Subjects completed most of the psychophysical blocks within the time allotted (Table 1). We present the low salience intensities estimated using Psi in Table 2 and high and low salience intensities estimated using the LNRM in Table 3. These are the unique intensities that were used to create the stimuli for each subject in the DFP.

Table 1
Number of target-present trials completed in each psychophysical block.

Psi		Type of judgment	
Subject	Session	Color	Orientation
1	1	25	25
	2	25	25
2	1	23	11
	2	25	16
3	1	25	18
	2	25	25
4	1	22	20
	2	25	23
LNRM		Type of judgment	
Subject	Session	Color	Orientation
5	1	32	23
	2	50	41
6	1	34	26
	2	50	42
7	1	47	31
	2	50	50
8	1	50	50
	2	50	37

Note. The maximum number of trials that could be completed was 25 and 50 for the Psi and LNRM conditions, respectively.

Table 2
Psi condition stimulus intensities.

Color			
Subject	Session	Low salience	High salience
1	1	−25.927	−50
	2	−21.656	−50
2	1	−13.007	−50
	2	−9.510	−50
3	1	−25.062	−50
	2	−21.176	−50
4	1	−7.928	−50
	2	−2.822	−50
Orientation			
Subject	Session	Low salience	High salience
1	1	81.633	90
	2	81.398	90
2	1	57.229	90
	2	59.063	90
3	1	70.672	90
	2	71.567	90
4	1	65.669	90
	2	62.166	90

Note. High salience was fixed for all subjects.

Subjects were very accurate. The worst individual accuracy we observed in one of the redundant signals conditions (i.e., HH, HL, LH, LL), which are the critical conditions for a SIC, was 89%. A Bayesian ANOVA revealed strong evidence for an effect of session on accuracy ($BF = 6.299 \times 10^4$), so we breakdown accuracy by condition and session in Fig. 8. Mean correct RTs reflected our salience manipulations and were longer for target-absent trials. We found some evidence against an effect of session on mean RT ($BF = 7.271$) but plot them by condition and session in Fig. 9 for consistency.

In order to interpret the SIC, we must first affirm the assumption that $S_{HH}(t) < \{S_{HL}(t),\ S_{LH}(t)\} < S_{LL}(t)$ holds. One way to do this is with a series of Kolmogorov–Smirnov tests. While we did not find strict evidence for the above ordering (i.e. we could not reject the null hypothesis that $S_{HH}(t) > \{S_{HL}(t),\ S_{LH}(t)\}$ or $S_{LL}(t) < \{S_{LH}(t),\ S_{HL}(t)\}$) for any subject, we also did not find any significant violations of the ordering (i.e. we could not reject the null hypothesis that $S_{HH}(t) < \{S_{HL}(t),\ S_{LH}(t)\}$ and $S_{LL}(t) > \{S_{LH}(t),\ S_{HL}(t)\}$). In this ambiguous situation, we assume that so long as the survivor functions are appropriately ordered upon

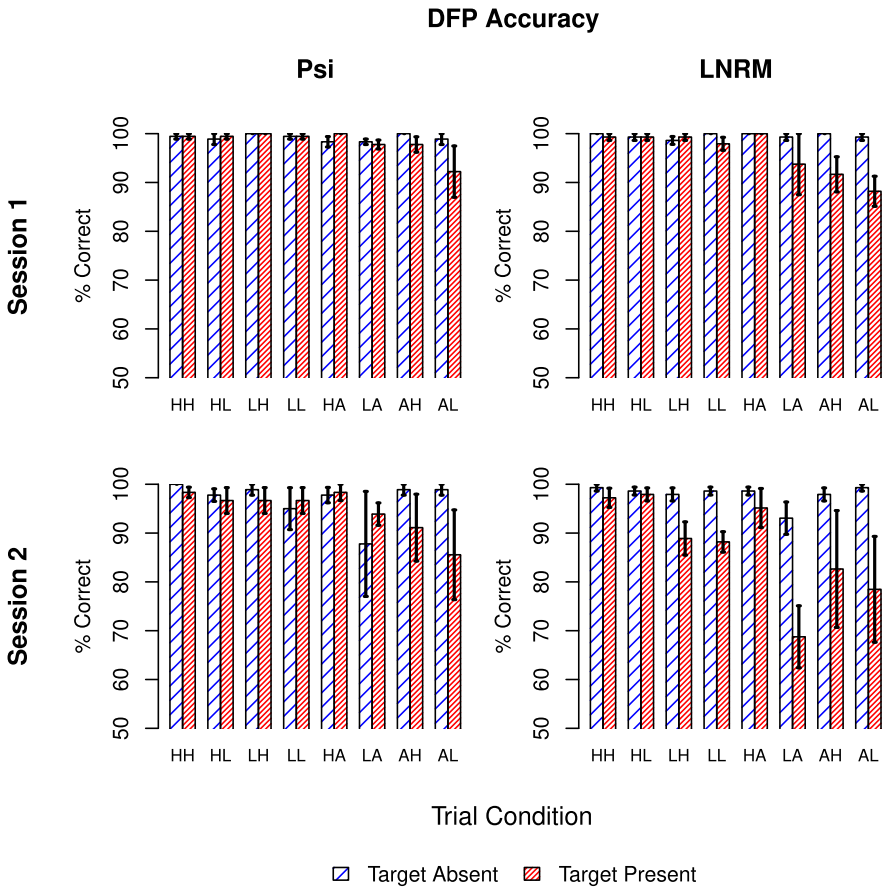

Fig. 8. Accuracy for the DFP portion of the study. Error bars indicate standard error of the mean across subjects. H, L, and A indicate trial conditions where the difference between the target and a particular distractor dimension was high, low, or absent (i.e. identical to the target), respectively. Color dissimilarity always precedes orientation dissimilarity such that HL indicates the high color salience and low orientation salience condition.

Table 3
LNRM condition stimulus intensities.

Color			
Subject	Session	Low salience	High salience
5	1	−25.880	−42.551
	2	−12.906	−24.927
6	1	−7.271	−16.585
	2	−1.736	−10.918
7	1	−37.097	−52.274
	2	−30.642	−45.683
8	1	−9.254	−21.741
	2	1.933	−9.414
Orientation			
Subject	Session	Low salience	High salience
5	1	70.219	80.717
	2	66.350	74.889
6	1	63.206	76.178
	2	63.841	71.819
7	1	79.230	88.490
	2	84.708	93.614
8	1	72.074	81.446
	2	62.374	71.538

visual inspection then any failure to reject the Kolmogorov–Smirnov null hypothesis is due to the limited number of trials rather than a real violation of selective influence. To this end, we inspected each subject's survivor functions and determined that we cannot interpret SICs for Subjects 1 and 8 in the target-present condition and either SIC for Subject 7. We plot the SICs and corresponding survivor functions for each subject separately for target present and absent conditions. The results for the Psi condition may be found in Figs. 10 and 11, and the results for the LNRM condition may be found in Figs. 12 and 13.

We conducted statistical tests of the SIC and MIC (Table 4). Following the recommendation of Houpt and Burns (2017), we use an α of .33. In the target-present condition, Subjects 2, 4, and 6 had significantly positive SIC deviations from zero ($D^+ \in [0.286, 0.369]$, $p \in [.003, .026]$), non-significant negative SIC deviations from zero ($D^- \in [0.011, 0.156]$, $p \in [.337, .995]$), and significantly positive MICs (ART $\in [0.223, 0.657]$, $p < .001$). Subject 3's SIC did not significantly deviate above zero ($D^+ = 0.103$, $p = .625$) but significantly deviated below zero ($D^- = 0.195$, $p = .183$). Their MIC did not significantly deviate from zero (ART $= 0.010$, $p = .809$). Subject 5's SIC and MIC did not significantly deviate from zero ($D^+ = 0.093$, $p = .743$; $D^- = 0.123$, $p = .594$; ART $= -0.059$, $p = .927$).

In the target-absent condition, Subjects 2, 4, and 5 had significantly positive SIC deviations from zero ($D^+ \in [0.198, 0.598]$, $p \in [.000, .246]$), non-significant negative SIC deviations from zero ($D^- \in [0.022, 0.105]$, $p \in [.673, .978]$), and significantly positive MICs (ART $\in [0.303, 1.250]$, $p \in [.000, .068]$). Subjects 1 and 3 had non-significant positive SIC deviations from zero ($D^+ \in [0.037, 0.122]$, $p \in [.519, .940]$) and significant negative SIC deviations from zero ($D^- \in [0.176, 0.212]$, $p \in [.136, .253]$). Subject 3 had an MIC that was significantly different from zero (ART $= -0.056$, $p = .204$) and Subject 1 did not (ART $= 0. - 0.020$, $p = .624$). Subjects 6 and 8 did not have SICs or MICs that significantly deviated from zero ($D^+ \in [0.094, 0.111]$, $p \in [.641, .732]$; $D^- \in [0.102, 0.153]$, $p \in [.432, .694]$; ART $\in [-0.006, 0.089]$, $p \in [.473, .825]$).

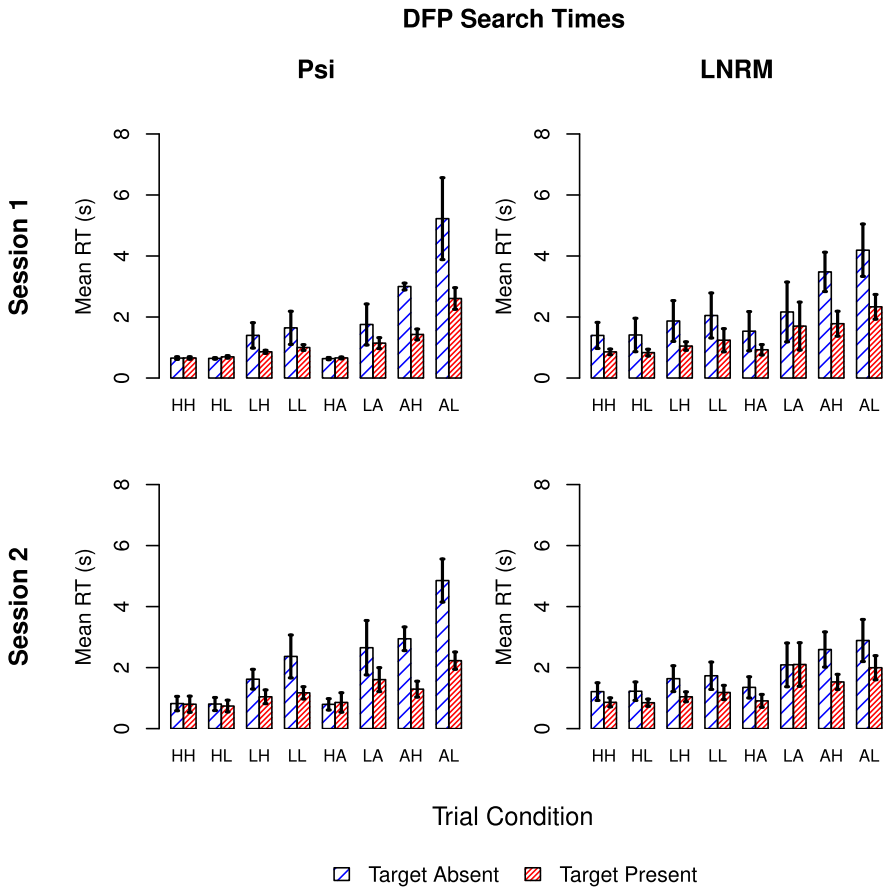

Fig. 9. Mean search times for correct responses in the DFP portion of the study. Error bars indicate standard error of the mean across subjects. H, L, and A indicate trial conditions where the difference between the target and a particular distractor dimension was high, low, or absent (i.e. identical to the target), respectively. Color dissimilarity always precedes orientation dissimilarity such that HL indicates the high color salience and low orientation salience condition.

Table 4
SIC and MIC results.

Psi								
Subject	Target	D^+	p	D^-	p	ART	p	Predicted model
1	Present	–	–	–	–	–	–	–
	Absent	0.122	.519	0.176	.253*	−0.020	.624	Parallel-AND
2	Present	0.286	.026*	0.011	.995	0.223	<.001*	Parallel-OR
	Absent	0.598	<.001*	0.031	.958	1.250	<.001*	Parallel-OR
3	Present	0.103	.625	0.195	.183*	0.010	.809	Parallel-AND
	Absent	0.037	.940	0.212	.136*	−0.056	.204*	Parallel-AND
4	Present	0.356	.003*	0.156	.337	0.292	<.001*	Parallel-OR
	Absent	0.522	<.001*	0.022	.978	0.798	<.001*	Parallel-OR
LNRM								
Subject	Target	D^+	p	D^-	p	ART	p	Predicted model
5	Present	0.093	.743	0.123	.594	−0.059	.927	Serial-OR
	Absent	0.198	.246*	0.105	.673	0.303	.068*	Parallel-OR
6	Present	0.369	.008*	0.057	.893	0.657	<.001*	Parallel-OR
	Absent	0.111	.641	0.153	.432	0.089	.473	Serial-OR
7	Present	–	–	–	–	–	–	–
	Absent	–	–	–	–	–	–	–
8	Present	–	–	–	–	–	–	–
	Absent	0.094	.732	0.102	.694	−0.006	.825	Serial-OR

Note. D^+ and D^- are Houpt–Townsend statistics for the positive and negative SIC deviations, respectively. ART is the Adjusted Rank Transform test statistic for the MIC. Significant p-values are indicated by asterisks ($\alpha = .33$; Houpt & Burns, 2017). We have replaced with dashes the values for subjects who violated selective influence to discourage interpretation of these results.

6.3. Discussion of human results

We conducted a visual search task where the color and orientation of distractors were adaptively chosen for each session and subject based on the subject's performance in a preliminary psychophysical portion of the experiment. One condition used the Psi psychophysical method (Kontsevich & Tyler, 1999) to recommend low salience intensities, and the other condition used the LNRM (Rouder et al., 2015) to estimate low and high salience intensities based on joint choice-RT distributions.

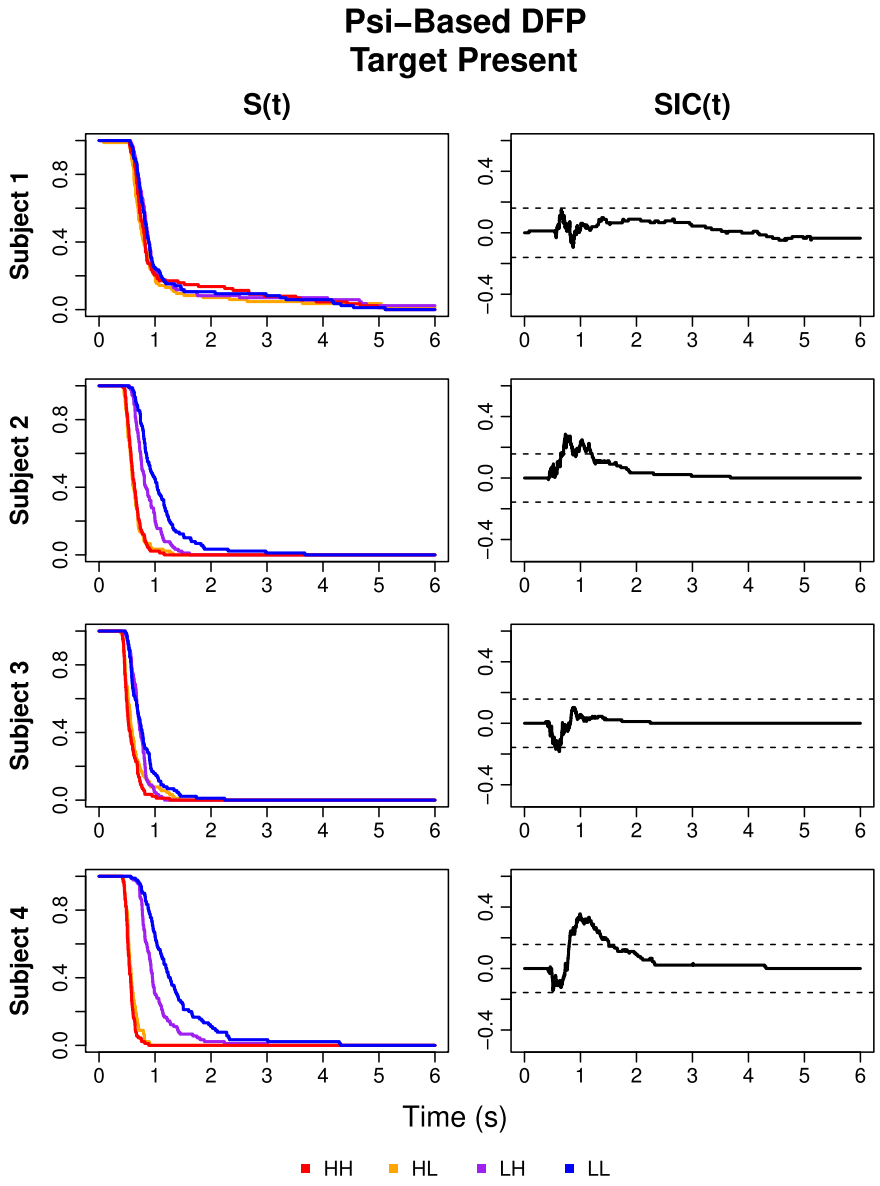

Fig. 10. Survivor and SIC functions on target present trials for each subject in the Psi condition. Dashed lines indicate the critical D^+ and D^- values the SIC must surpass to reject the null-hypothesis that SIC = 0 with $\alpha = .33$. Note that we cannot interpret the target-present SIC for Subject 1.

The range of intensities recommended for each group of subjects was consistent across the two psychophysical methods. As expected, thresholds varied greatly between subjects but relatively little within subjects (i.e. across sessions). The color salience levels proposed for subjects in the LNRM condition differed across sessions somewhat more than their counterparts in the Psi condition, which may reflect a larger change in the time to discriminate colors compared to the accuracy of such discriminations. The LNRM would take into account this dimension of learning when estimating thresholds whereas Psi would not.

We consider two sets of trial conditions when evaluating accuracy. Participants were very accurate on trials where distractors differed from the target in both color and orientation (HH, HL, LH, LL; Fig. 8), which is important because these are the trials used in the SIC calculation. Accuracy was somewhat lower for the trials where distractors differed from the target in only one dimension, particularly for the target-present trials, which suggests that subjects were more likely to miss the target than to misidentify it in these more difficult conditions. Subjects were less accurate in the second session, and although one might expect accuracy to improve with additional sessions, thresholds did tend to decrease between sessions. Because these intensities were estimated from performance at the beginning of a session, subjects likely improved since the first psychophysical blocks, resulting in higher accuracy than expected for the first session. By the start of the second session, subjects were well practiced such that the thresholds estimated were lower but accuracy did not subsequently improve as before.

Mean RTs followed the general pattern we expected, increasing as the perceptual difference between distractors and the target decreased. Participants were faster when the target was present and when they could exploit color dissimilarity.

The SIC and MIC results largely support parallel processing of color and orientation during visual search, which is consistent with what we have previously found for color and shape (Glavan et al., 2019). In the target-present condition, Subjects 2, 4, and 6 demonstrated results consistent with self-terminating parallel processing. Subject 3's SIC results suggested that they used exhaustive parallel processing, although if this was the case then we would expect their MIC to be significantly negative.

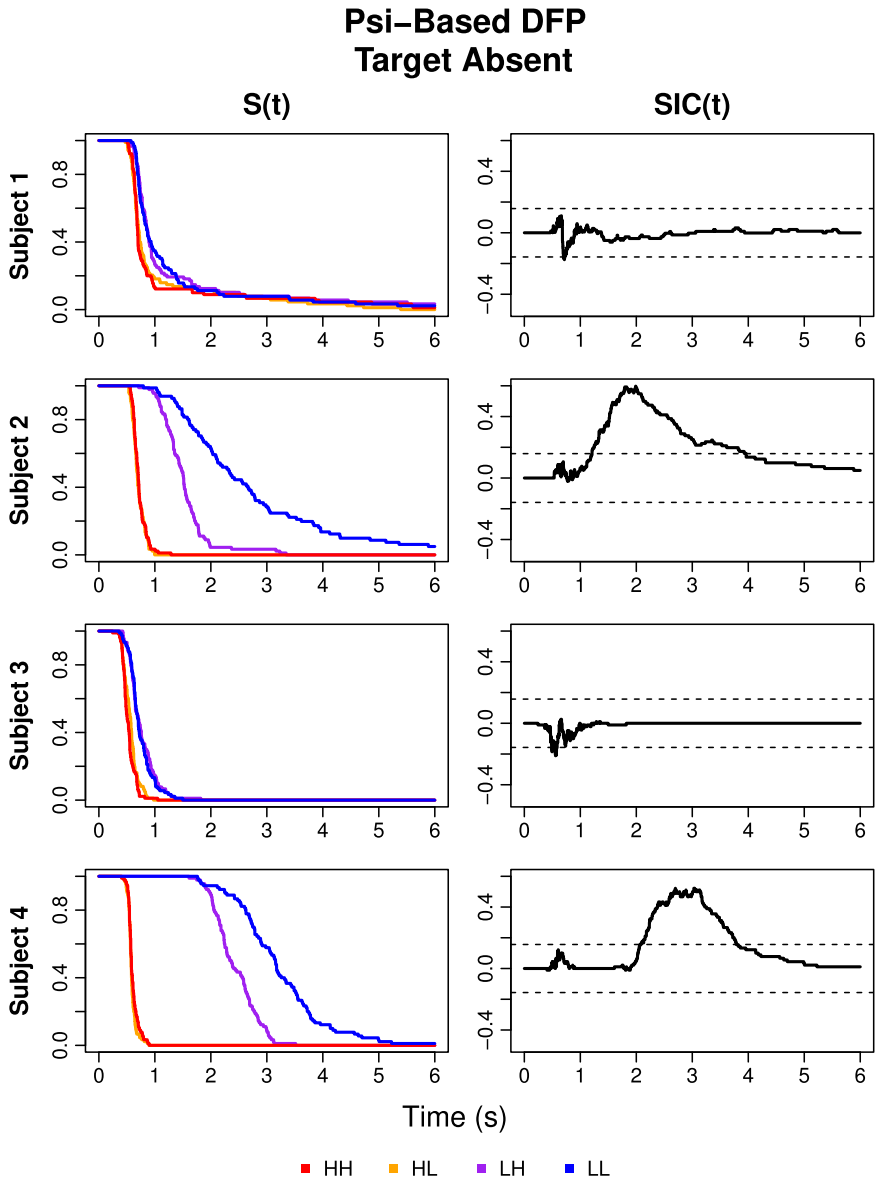

Fig. 11. Survivor and SIC functions on target absent trials for each subject in the Psi condition. Dashed lines indicate the critical D^+ and D^- values the SIC must surpass to reject the null-hypothesis that SIC $= 0$ with $\alpha = .33$.

It may instead be the case that they used an exhaustive serial strategy, and we simply failed to detect the positive SIC deviation. Similarly, Subject 5's results are consistent with self-terminating serial processing, but this may also be a failure to reject caused by insufficient power.

In the target-absent condition, Subject 3 demonstrated results consistent with exhaustive parallel processing, while Subjects 2, 4, and 5 appeared to use self-terminating parallel processing. A self-terminating stopping rule is bemusing considering that subjects must fully process all color and orientation information in order to respond accurately when the target was absent. Workload capacity analysis (Townsend & Nozawa, 1995) may resolve this discrepancy by revealing facilitation between processing channels (Eidels et al., 2011; Glavan et al., 2019). Subject 1's SIC results suggested that they used exhaustive parallel processing, but, like Subject 3 above, their MIC was not significantly negative as we would expect, leading us to suspect that maybe some subjects used an exhaustive serial strategy. Subjects 6 and 8's SICs and MICs did not differ from zero, which is consistent with self-terminating serial processing, but, again because we expect exhaustive processing to be necessary to be accurate when the target is absent, these non-significant results would likely change with additional trials.

This study was based on relatively few participants because our main focus was on demonstrating the adaptive SFT methods. Hence, we caution readers against putting too much credence in these conclusions. A larger study should be conducted to more comprehensively address the research question introduced here. Not only did we restrict the overall experiment time, but because the Psi method is more efficient than the LNRM approach, we had to collect more psychophysical trials in the LNRM condition than the Psi condition. Nevertheless, the number of psychophysical trials we actually collected for the LNRM condition (5 repetitions of 10 levels) compared to our recommended minimum number of trials (50 repetitions of 10 levels) was proportionally much lower than the number used for the Psi condition (25 trials compared to the recommended minimum of 100). Therefore, the quality of the stimulus levels proposed by the LNRM may have been more impacted by the time constraint than those proposed by Psi. Additionally, the extra psychophysical trials in the LNRM

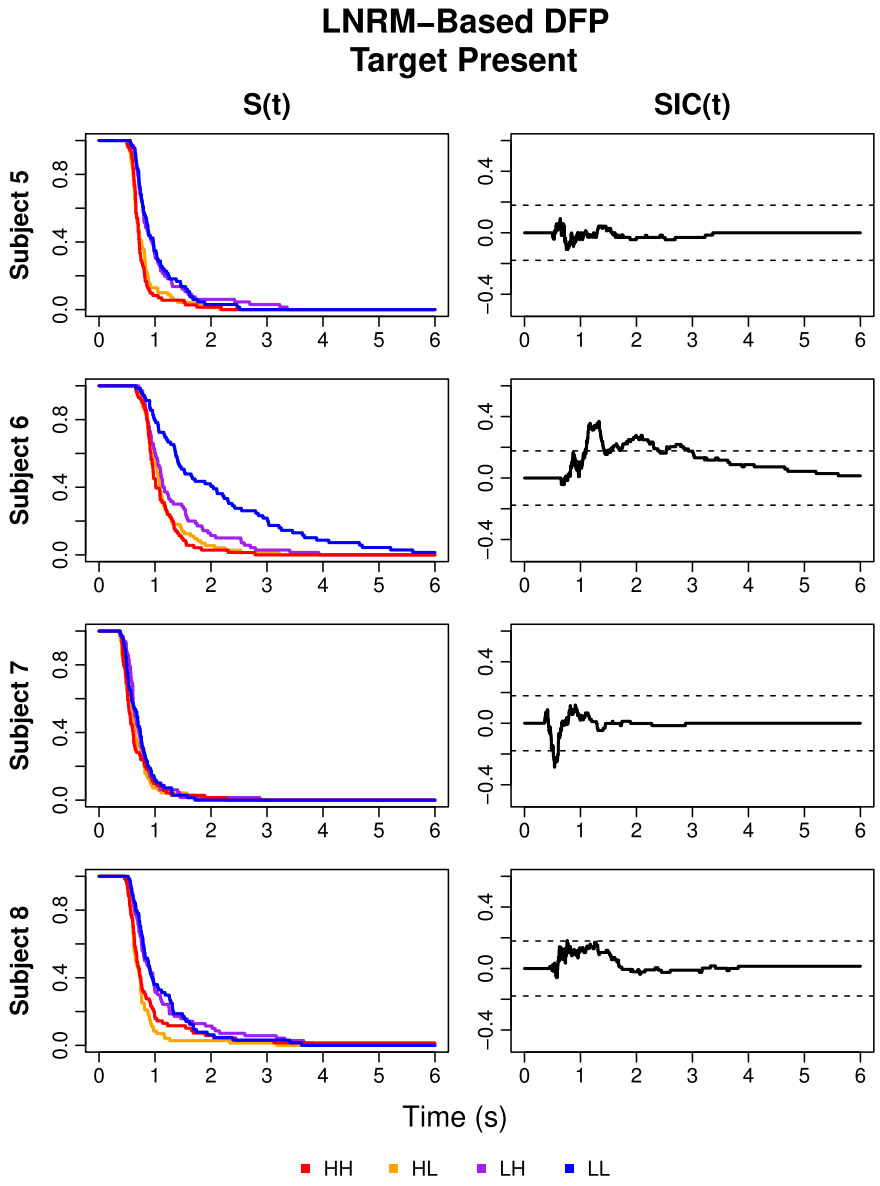

Fig. 12. Survivor and SIC functions on target present trials for each subject in the LNRM condition. Dashed lines indicate the critical D^+ and D^- values the SIC must surpass to reject the null-hypothesis that SIC = 0 with $\alpha = .33$. Note that we cannot interpret the target-present SIC for Subjects 7 and 8.

condition meant less time available to collect DFP data. Subjects in the Psi condition completed 90 trials per salience condition over two one-hour sessions, compared to only 72 trials for the LNRM condition. Therefore, the statistical tests for the SIC and MIC had less statistical power for the LNRM condition than the Psi condition. When one also considers that a serial-OR model is effectively the null hypothesis of the SIC statistical tests, it becomes clear that differences in the human results between adaptive method conditions may be because the LNRM method is less robust to the practical constraints we imposed than due to true differences in architecture between the groups.

The variability between adaptive methods that we found in the human study highlights a limitation of the current approach. Although choosing efficient stimulus properties may improve the power of the DFP experiment, it is not the sole contributing factor. The number of samples used to estimate each SIC distribution also has a major impact. Unfortunately, when the overall experiment time is limited, including an adaptive routine like the ones we propose will reduce the number of DFP trials that can be collected within a given session. Furthermore, the tradeoff between effect size and number of samples may not be easy to estimate a priori; a preliminary power analysis may be necessary to determine the appropriate number of DFP trials. From there, the choice between accuracy-based and joint RT–accuracy methods can be made depending on remaining constraints.

7. General discussion

Understanding the fundamental characteristics of how a cognitive system combines multiple sources of information is critical for understanding that cognitive system. SFT provides a framework for formalizing and examining those characteristics. SFT methods are both powerful and general because they are based on formal mathematical derivations that do not rely on parametric or distributional assumptions. Unfortunately, SFT is often out of reach for experimentalist hoping to apply these methods in their domains. One major hindrance to applying SFT is determining the stimulus characteristics that will lead to interpretable data.

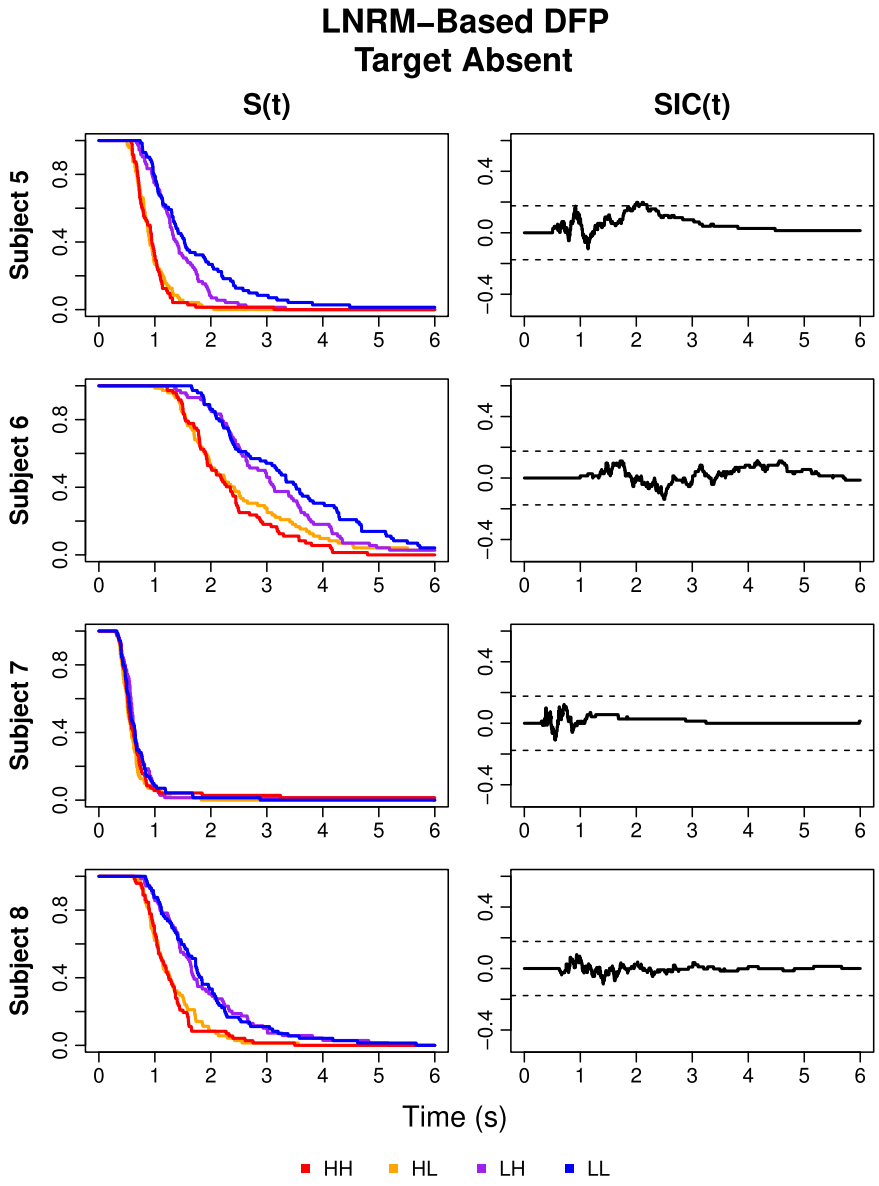

Fig. 13. Survivor and SIC functions on target absent trials for each subject in the LNRM condition. Dashed lines indicate the critical D^+ and D^- values the SIC must surpass to reject the null-hypothesis that SIC = 0 with $\alpha = .33$. Note that we cannot interpret the target-absent SIC for Subject 7.

In this paper, we demonstrated two approaches to finding stimulus salience levels that are likely to lead to clean and interpretable estimates of the SIC across individuals. In keeping with the fundamental strengths of SFT, the methods we proposed are general, making only minimal assumptions about the stimulus dimensions to which they are applied. They are also applicable at the individual subject level, avoiding the pitfalls inherent in focusing exclusively on group-level data (Estes, 1956).

The first method we proposed for determining the appropriate stimulus salience levels was Kontsevich and Tyler's (1999) Psi method, which focuses solely on a participant's accuracy. We proposed that the appropriate setting for the low salience stimulus level is the level at which a subject should achieve 90% accuracy, under the assumption that accuracy is negatively correlated with difficulty, and hence response times. For the high salience stimulus level, we suggested using the maximum salience, when such a value exists, or a level that is otherwise phenomenologically "easy", leading to maximal accuracy. Although we did not explore alternatives for setting the high salience level with Psi in this paper, there are other options. For example, one could choose a stimulus level that corresponded with an even higher level of accuracy on the psychometric function (e.g. 95%) or choose a constant multiple of the low salience stimulus intensity.

From a series of simulations we found that reasonably stable individual stimulus thresholds could be obtained within approximately 100–150 trials using the Psi method. Without good priors, we recommend a minimum of 100 trials per stimulus dimension. Next, we simulated a fully adaptive DFP study that used the Psi method to estimate individualized stimulus salience levels. To do so, we randomly selected sets of parameters for a drift–diffusion model that resulted in human-like choice and RT performance, with each set of parameters representing a unique individual. We then determined the low salience level for each "individual" on both of two dimensions, putatively color and orientation. We used those stimulus intensities and the maximum physical stimulus intensities to generate responses from correspondingly parameterized DDMs constructed from specific combinations of architecture and stopping rule. The resulting SIC functions matched the patterns predicted by the chosen architecture and stopping rule quite well, indicating promise for the Psi method in the experimental design pipeline.

Finally, we applied the Psi method for determining stimulus salience levels to human observers. Following the same procedure used for the simulations, we first ran participants through sequences of trials following the Psi method to estimate the color for which a participant should achieve 90% accuracy on color target detection and the orientation of a line for which that participant should achieve 90% accuracy on orientation target detection. Because each subject completed these trials on each day, we were able to set low salience target values that were specific to each participant on each day of the experiment. Using those low salience target levels along with a predetermined high salience target level, we created individualized stimuli for the DFP by factorially combining the stimulus levels across color and orientation. While the approach did not work for all participants, many participants' data indicated clear ordering of survivor functions according to stimulus salience levels, and clearly identifiable SIC shapes.

As an alternative to the accuracy-only method, we also proposed a method based on the log-normal race model (Rouder et al., 2015) for adapting jointly to RT and accuracy. Like the accuracy-based approach, we first examined the LNRM through simulations where data were generated from a DDM. The first simulation focused on the precision of parameter estimates as a function of the number of trials. Next, we examined the LNRM in the context of a DFP paradigm by using the stimulus intensity levels indicated by the LNRM for the high and low salience levels on two stimulus dimensions. To examine the robustness of the approach to individual variation, we randomly perturbed the parameters to generate a variety of choice–RT profiles. Once we completed simulation testing, we then applied the approach with human participants.

For the joint RT–accuracy approach, we used the method of constant stimuli and fit the LNRM post-hoc rather than developing an online adaptive procedure. The method of constant stimuli can vary in its extent based on the number of stimulus salience levels tested and the number of trials at each stimulus level. As with most models, more data yielded more precise parameter estimates, but we did not explore the trade-offs in precision between using more salience levels or using more trials per level. We used ten salience levels to cover a wide range of potential performance patterns. Using data generated from the DDM, we estimated posterior highest density intervals as a function of the number of trials at each of those ten levels. When relatively vague priors were used, we found that the posterior intervals reached asymptote at around 100 trials per level. This would suggest that in the absence of prior information, no more than 1000 trials (100 by ten levels) are necessary. Based on the simulation results, we recommend a minimum of 50 trials per level (500 total trials) per dimension. While this is a large amount of trials, we see it as less of a cost than having to throw out participants' data due to ineffective salience manipulations or insufficient accuracy. In some cases, this may not result in a practical improvement in data collection time and cost. We are currently developing adaptive methods in line with Psi and other approaches (Kim, Pitt, Lu, & Myung, 2017; Kim, Pitt, Lu, Steyvers, & Myung, 2014) for joint RT–accuracy modeling, which we hope will further reduce the recommended number of trials.

Following the parameter convergence simulations, we fit the LNRM to responses generated from a DDM that was run through a method of constant stimuli to determine salience levels for a DFP study. Like the values chosen based on the Psi method, the high and low salience levels extracted from the LNRM yielded clear survivor orderings and SIC shapes that matched the generating architecture and stopping rule. Furthermore, when we generated data from a collection of parameter sets for the DDM, we found that the LNRM-based approach was robust to variation across individuals in the generated performance profiles.

When applied with human subjects, the LNRM approach yielded similar outcomes to the Psi approach. For some participants, there was not a clear separation of survivor functions across salience levels. For those that did show a clear separation, the resulting SICs were reasonably identifiable. Although the sample size is small, these results seem to indicate an advantage for the Psi approach. This is likely due to the inefficiency of the method of constant stimuli used to inform the LNRM; in addition to collecting fewer trials per salience level, this method takes longer, which means fewer DFP trials can be collected in the same amount of experiment time. Another possible explanation may be non-stationarity in the participants' performance, either from learning or strategy shifts. Nevertheless, it was clear that different subjects needed drastically different salience levels to achieve similar performance, and hence, choosing a single set of salience levels for all participants would have yielded even more problematic results.

In the human study, we emphasized practicality, using far fewer trials than recommended by our simulation results. This likely explains why we did not produce selective influence across all subjects. Furthermore, we know from our simulations that the joint RT–accuracy approach requires many more trials than the accuracy-only approach because of its reliance on the method of constant stimuli. This likely explains why we found more violations of selective influence for the LNRM condition than the Psi condition. In other words, these results are to be expected given our simulation findings. That being said, even in these sub-optimal conditions the techniques we developed did not catastrophically fail, and compliance rates were fairly similar to those we have found with the traditional "pilot and test" method, which suggests they are more robust than one might expect given the parameter convergence simulations. In the future, it may be worth investing a day or two in collecting nothing but psychophysical data so as to establish a strong group-level prior for the stimuli/task being used. Such an informative prior could drastically reduce the number of trials needed to adapt the stimuli to individuals on days on which the full DFP is administered. We look forward to pioneering such hierarchical methods within the SFT framework.

8. Conclusion

The process of finding the right salience levels is one of the more complex and resource intensive aspects of the SFT methodology. The traditional approach, which relies on extensive pilot testing for determining a single set of salience levels at the group level, can lead to results wherein a large number of participants' data is unusable because either the group-set salience levels were too difficult, and accuracy was too low, or the levels did not yield sufficiently different RTs across salience conditions. In this paper, we investigated two approaches to determining individualized salience levels. The first approach was based on the Psi method, a well established method in the psychophysics literature for estimating the mapping between stimulus salience and accuracy. The second approach was a novel approach relying on the LNRM for estimating the mapping between stimulus salience and both RT and accuracy. We hope that these approaches will contribute to making the powerful SFT methodology more accessible to a wider range of experimental psychologists and more applicable to a wider range of domains.

References

Appelle, S. (1972). Perception and discrimination as a function of stimulus orientation: the "oblique effect" in man and animals. *Psychological Bulletin, 78*(4), 266.

Blaha, L. M., Houpt, J. W., McIntire, J. P., Havig, P. R., & Morris, M. B. (manuscript in preparation). Characterizing stereoscopic disparity information processing with systems factorial technology.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., et al. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1).

Donkin, C., Brown, S., & Heathcote, A. (2011). Drawing conclusions from choice response time models: A tutorial using the linear ballistic accumulator. *Journal of Mathematical Psychology*, *55*(2), 140–151.

Eidels, A., Houpt, J. W., Altieri, N., Pei, L., & Townsend, J. T. (2011). Nice guys finish fast and bad guys finish last: Facilitatory vs. inhibitory interaction in parallel systems. *Journal of Mathematical Psychology*, *55*(2), 176–190.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*(2), 134.

Fifić, M., & Little, D. R. (2017). Stretching mental processes: An overview of and guide for SFT applications. In D. R. Little, N. Altieri, M. Fifić, & C.-T. Yang (Eds.), *Systems factorial technology: A theory driven methodology for the identification of perceptual and cognitive mechanisms* (pp. 27–51). Elsevier.

Fox, E. L., & Houpt, J. W. (2016). The perceptual processing of fused multi-spectral imagery. *Cognitive Research: Principles and Implications*, *1*(1), 31.

Glavan, J. J., Haggit, J. M., & Houpt, J. W. Temporal organization of color and shape processing during visual search. Attention, Perception, & Psychophysics, Special issue in honor of the contributions of Anne Treisman, in press.

Houpt, J. W., Blaha, L. M., McIntire, J. P., Havig, P. R., & Townsend, J. T. (2014). Systems factorial technology with r. *Behavior Research Methods*, *46*(2), 307–330.

Houpt, J. W., & Burns, D. M. (2017). Statistical analyses for systems factorial technology. In D. R. Little, N. Altieri, M. Fifić, & C.-T. Yang (Eds.), *Systems factorial technology* (pp. 55–67). San Diego: Elsevier, http://dx.doi.org/10.1016/B978-0-12-804315-8.00005-7, URL http://www.sciencedirect.com/science/article/pii/B9780128043158000057.

Houpt, J. W., & Fifić, M. (2017). A hierarchical bayesian approach to distinguishing serial and parallel processing. *Journal of Mathematical Psychology*, *79*, 13–22.

Houpt, J. W., MacEachern, S. N., Peruggia, M., & Townsend, J. T. (2016). Semiparametric bayesian approaches to systems factorial technology. *Journal of Mathematical Psychology*, *75*, 68–85.

Houpt, J. W., & Townsend, J. T. (2010). The statistical properties of the survivor interaction contrast. *Journal of Mathematical Psychology*, *54*(5), 446–453.

Houpt, J. W., & Townsend, J. T. (2011). An extension of SIC predictions to the Wiener coactive model. *Journal of Mathematical Psychology*, *55*(3), 267–270.

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

Kim, W., Pitt, M. A., Lu, Z.-L., & Myung, J. I. (2017). Planning beyond the next trial in adaptive experiments: A dynamic programming approach. *Cognitive Science*, *41*(8), 2234–2252.

Kim, W., Pitt, M. A., Lu, Z.-L., Steyvers, M., & Myung, J. I. (2014). A hierarchical adaptive approach to optimal experimental design. *Neural Computation*, *26*(11), 2465–2492.

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

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

Leys, C., & Schumann, S. (2010). A nonparametric method to analyze interactions: The adjusted rank transform test. *Journal of Experimental Social Psychology*, *46*(4), 684–688.

Molenaar, D., Tuerlinkcx, F., & van der Maas, H. (2015). Fitting diffusion item response theory models for responses and response times using the r package diffirt. *Journal of Statistical Software*, *66*, 1–34.

Mulder, M. J., Wagenmakers, E.-J., Ratcliff, R., Boekel, W., & Forstmann, B. U. (2012). Bias in the brain: a diffusion model analysis of prior probability and potential payoff. *Journal of Neuroscience*, *32*(7), 2335–2343.

Murphy, P. R., Vandekerckhove, J., & Nieuwenhuis, S. (2014). Pupil-linked arousal determines variability in perceptual decision making. *PLoS Computational Biology*, *10*(9), e1003854.

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, *2*, 10.

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, URL https://www.R-project.org/.

Rammsayer, T. H. (1992). An experimental comparison of the weighted up-down method and the transformed up-down method. *Bulletin of the Psychonomic Society*, *30*(5), 425–427.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59.

Reinach, S. (1965). A nonparametric analysis for a multi-way classification with one element per cell. *South African Journal of Agricultural Science*, *8*(4), 941–960.

Rouder, J. N., Province, J. M., Morey, R. D., Gomez, P., & Heathcote, A. (2015). The lognormal race: A cognitive-process model of choice and latency with desirable psychometric properties. *Psychometrika*, *80*(2), 491–513.

Ruzon, M. Lab2rgb, https://www.mathworks.com/matlabcentral/fileexchange/24010-lab2rgb/ . Version 1.0.0.

Stone, M. (1960). Models for choice-reaction time. *Psychometrika*, *25*(3), 251–260.

Taylor, M., & Creelman, C. (1967). Pest: Efficient estimates on probability functions. *The Journal of the Acoustical Society of America*, *41*(4A), 782–787.

Townsend, J. T., & Nozawa, G. (1995). Spatio-temporal properties of elementary perception: An investigation of parallel, serial, and coactive theories. *Journal of Mathematical Psychology*, *39*(4), 321–359.

Watson, A. B. (2017). Quest+: A general multidimensional bayesian adaptive psychometric method. *Journal of Vision*, *17*(3), 10.

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